SCHOOL LAUNCH · MATHEMATICS · GRADES 10–12 FIRST EDITION

STATISTICS

Learning from data in a world of variation and chance

Statistics is the science of learning from data when the things we measure vary: no two eggs weigh the same, no two summers bring the same rainfall, and no batter gets a hit every time. This one-year course begins with graphs and numerical summaries of data, moves to the normal distribution, correlation and regression, and then to how good data are collected through careful sampling and well-designed experiments. The second half builds the tools of chance (probability, random variables, the binomial and geometric distributions, and sampling distributions) and uses them for statistical inference: confidence intervals and significance tests for proportions and means, comparisons of two groups, and a short unit on chi-square tests.

The course teaches the reasoning behind every method, not only the steps. You will learn why each method works and what its answer means, and you will practice reasoning carefully and honestly about uncertain things. Every answer is interpreted in a sentence in the context of the problem, because a number without its meaning is only half an answer. Daily practice sets build skill steadily, and short history notes tell how men such as Halley, de Moivre, Gauss, Fisher and Gosset discovered these ideas while studying real problems in astronomy, farming and industry.

Who it is for: Students in grades 10–12 (ages 15–18) who have completed Algebra I (Geometry and Algebra II are helpful but not required); the course is good preparation for college statistics and for further study in the sciences, agriculture, business, engineering and medicine.

Contents

1. Course guide: how to use, scope and sequence, pacing, project, materials, grading, note to parents

2. Unit 1. Exploring Data: Variables and Graphs

- 1.1 What Is Statistics?
- 1.2 Displaying Categorical Data
- 1.3 Two-Way Tables
- 1.4 Displaying Quantitative Data
- 1.5 Describing a Distribution

3. Unit 2. Describing Data with Numbers

- 2.1 Measures of Center
- 2.2 Quartiles, the Five-Number Summary and IQR
- 2.3 Boxplots and Outliers
- 2.4 Standard Deviation
- 2.5 Comparing Distributions
- 2.6 Transforming Data

4. Unit 3. Position and the Normal Distribution

- 3.1 Percentiles and z-Scores
- 3.2 Density Curves
- 3.3 Normal Distributions and the 68–95–99.7 Rule
- 3.4 Finding Normal Proportions
- 3.5 Finding Values from Normal Proportions
- 3.6 Assessing Normality

5. Unit 4. Scatterplots, Correlation and Regression

- 4.1 Scatterplots and Association
- 4.2 The Correlation Coefficient
- 4.3 The Least-Squares Regression Line
- 4.4 Residuals and r^2
- 4.5 Cautions about Regression and Causation

6. Unit 5. Collecting Data: Samples and Experiments

- 5.1 Populations, Samples and Bias
- 5.2 Random Sampling Methods
- 5.3 Problems with Surveys
- 5.4 Observational Studies and Experiments
- 5.5 Designing Experiments
- 5.6 Blocking, Matched Pairs and the Scope of Conclusions

7. Unit 6. Probability

- 6.1 Randomness and Simulation
- 6.2 Probability Rules
- 6.3 Conditional Probability
- 6.4 Independence and the Multiplication Rule
- 6.5 Tree Diagrams
- 6.6 Counting: Permutations and Combinations

8. Unit 7. Random Variables

- 7.1 Discrete Random Variables
- 7.2 Expected Value and Standard Deviation
- 7.3 Transforming and Combining Random Variables
- 7.4 Normal Random Variables and Their Sums

9. Unit 8. Binomial and Geometric Distributions

- 8.1 Binomial Settings and the Binomial Formula
- 8.2 Binomial Mean, Standard Deviation and Cumulative Probabilities
- 8.3 The Normal Approximation to the Binomial
- 8.4 Geometric Distributions

L0. Unit 9. Sampling Distributions

9.1 Parameters, Statistics and Sampling Variability
9.2 The Sampling Distribution of a Sample Proportion
9.3 The Sampling Distribution of a Sample Mean
9.4 The Central Limit Theorem

L1. Unit 10. Confidence Intervals

10.1 The Idea of a Confidence Interval
10.2 Estimating a Population Proportion
10.3 The t Distributions
10.4 Estimating a Population Mean
10.5 Margin of Error and Sample Size

L2. Unit 11. Significance Tests

11.1 The Logic of Significance Tests
11.2 Testing a Claim about a Proportion
11.3 Testing a Claim about a Mean
11.4 Type I and Type II Errors and Power
11.5 Tests and Intervals Together

L3. Unit 12. Comparing Two Groups

12.1 Comparing Two Proportions
12.2 Comparing Two Means
12.3 Paired Data
12.4 Choosing the Right Procedure

L4. Unit 13. Chi-Square Tests

13.1 The Chi-Square Goodness-of-Fit Test
13.2 Chi-Square Tests for Two-Way Tables
13.3 Thinking Wisely about Statistics

L5. Midterm and final exams

L6. Statistical tables

Course guide

How to use this course

- The course fills one school year: 32 weeks, 5 days a week, about 45 minutes a day, for 13 units, 63 lessons, a midterm, a final exam and a short statistics project.
- Each day, begin with the lesson's practice habit, then read the core ideas with pencil and calculator in hand, work through the two worked examples yourself, do the practice set, check your answers, and correct every mistake.
- Most lessons take two days: on day 1, study the teaching and examples and do the first half of the practice set; on day 2, review the core ideas and finish the second half. Lessons 1.1 and 13.3 take one day.
- Interpret every answer in a complete sentence in context, with units: say what the number means for the eggs, the fields or the free throws in the problem.
- Each unit ends with a review day (use the mastery checklist and redo missed problems) and a unit test.
- The midterm exam (Units 1–6) comes after Unit 6, and the cumulative final exam (Units 1–13) ends the year.
- The statistics project has two days of planning after Unit 5, data collection in spare time during the following weeks, and three days of analysis and a short written report after Unit 13.
- From Unit 2 on, use a scientific calculator with one- and two-variable statistics; the normal, t and chi-square tables are at the back of the course. A graphing calculator or a free online statistics calculator is optional. Small differences in the last decimal place between table and technology answers are fine.
- Students who need more time may stretch the course over more weeks; steady understanding matters more than speed.

Scope and Sequence

UNIT 1 · WEEKS 1–3

Exploring Data: Variables and Graphs

Students learn what statistics is, classify variables, and display and describe categorical and quantitative data with tables and graphs.

- 1.1 What Is Statistics?
- 1.2 Displaying Categorical Data
- 1.3 Two-Way Tables
- 1.4 Displaying Quantitative Data
- 1.5 Describing a Distribution

UNIT 2 · WEEKS 3–5

Describing Data with Numbers

Students measure center and spread with the mean, median, quartiles and standard deviation, draw boxplots, identify outliers and compare distributions.

- 2.1 Measures of Center
- 2.2 Quartiles, the Five-Number Summary and IQR
- 2.3 Boxplots and Outliers
- 2.4 Standard Deviation
- 2.5 Comparing Distributions
- 2.6 Transforming Data

UNIT 3 · WEEKS 6–8

Position and the Normal Distribution

Students describe position with percentiles and z-scores and use density curves and normal distributions to find proportions and values.

- 3.1 Percentiles and z-Scores
- 3.2 Density Curves
- 3.3 Normal Distributions and the 68–95–99.7 Rule
- 3.4 Finding Normal Proportions
- 3.5 Finding Values from Normal Proportions
- 3.6 Assessing Normality

UNIT 4 · WEEKS 8–11

Scatterplots, Correlation and Regression

Students describe relationships between two quantitative variables with scatterplots, the correlation r and the least-squares regression line, and learn the cautions about causation.

- 4.1 Scatterplots and Association
- 4.2 The Correlation Coefficient
- 4.3 The Least-Squares Regression Line
- 4.4 Residuals and r^2
- 4.5 Cautions about Regression and Causation

UNIT 5 · WEEKS 11–13

Collecting Data: Samples and Experiments

Students learn how random sampling and randomized experiments produce trustworthy data and what conclusions each allows.

- 5.1 Populations, Samples and Bias
- 5.2 Random Sampling Methods
- 5.3 Problems with Surveys
- 5.4 Observational Studies and Experiments
- 5.5 Designing Experiments
- 5.6 Blocking, Matched Pairs and the Scope of Conclusions

UNIT 6 · WEEKS 14–17

Probability

Students study chance through simulation, the rules of probability, conditional probability, independence, tree diagrams and counting.

- 6.1 Randomness and Simulation
- 6.2 Probability Rules
- 6.3 Conditional Probability
- 6.4 Independence and the Multiplication Rule
- 6.5 Tree Diagrams
- 6.6 Counting: Permutations and Combinations

UNIT 7 · WEEKS 17–19

Random Variables

Students describe random outcomes with probability distributions, expected values and standard deviations, and combine random variables.

- 7.1 Discrete Random Variables
- 7.2 Expected Value and Standard Deviation
- 7.3 Transforming and Combining Random Variables
- 7.4 Normal Random Variables and Their Sums

UNIT 8 · WEEKS 19–21

Binomial and Geometric Distributions

Students count successes in repeated trials with the binomial distribution, approximate it with a normal curve, and model waiting times with the geometric distribution.

- 8.1 Binomial Settings and the Binomial Formula
- 8.2 Binomial Mean, Standard Deviation and Cumulative Probabilities
- 8.3 The Normal Approximation to the Binomial
- 8.4 Geometric Distributions

UNIT 9 · WEEKS 21–23

Sampling Distributions

Students learn how sample proportions and sample means vary from sample to sample, and the central limit theorem.

- 9.1 Parameters, Statistics and Sampling Variability
- 9.2 The Sampling Distribution of a Sample Proportion
- 9.3 The Sampling Distribution of a Sample Mean
- 9.4 The Central Limit Theorem

UNIT 10 · WEEKS 23–25

Confidence Intervals

Students estimate population proportions and means with confidence intervals, learn the t distributions, and choose sample sizes.

- 10.1 The Idea of a Confidence Interval
- 10.2 Estimating a Population Proportion
- 10.3 The t Distributions
- 10.4 Estimating a Population Mean
- 10.5 Margin of Error and Sample Size

UNIT 11 · WEEKS 26–28

Significance Tests

Students test claims about a proportion or a mean, interpret P-values, and weigh Type I and Type II errors and power.

- 11.1 The Logic of Significance Tests
- 11.2 Testing a Claim about a Proportion
- 11.3 Testing a Claim about a Mean
- 11.4 Type I and Type II Errors and Power
- 11.5 Tests and Intervals Together

UNIT 12 · WEEKS 28–30

Comparing Two Groups

Students compare two proportions or two means with intervals and tests, analyze paired data, and choose the right procedure.

- 12.1 Comparing Two Proportions
- 12.2 Comparing Two Means
- 12.3 Paired Data
- 12.4 Choosing the Right Procedure

UNIT 13 · WEEKS 30–31

Chi-Square Tests

Students test categorical distributions and two-way tables with chi-square tests and review how to think wisely about statistical claims.

- 13.1 The Chi-Square Goodness-of-Fit Test
- 13.2 Chi-Square Tests for Two-Way Tables
- 13.3 Thinking Wisely about Statistics

Weekly pacing calendar

Five school days a week for 32 weeks (160 days).

Week	Plan
1	• Day 1: Lesson 1.1 What Is Statistics? • Day 2: Lesson 1.2 Displaying Categorical Data (day 1 of 2) • Day 3: Lesson 1.2 Displaying Categorical Data (day 2 of 2) • Day 4: Lesson 1.3 Two-Way Tables (day 1 of 2) • Day 5: Lesson 1.3 Two-Way Tables (day 2 of 2) *Families on a 4-day week can combine the two days of a two-day lesson when the student is ready, or stretch the course to about 40 weeks.*
2	• Day 1: Lesson 1.4 Displaying Quantitative Data (day 1 of 2) • Day 2: Lesson 1.4 Displaying Quantitative Data (day 2 of 2) • Day 3: Lesson 1.5 Describing a Distribution (day 1 of 2) • Day 4: Lesson 1.5 Describing a Distribution (day 2 of 2) • Day 5: Unit 1 review (mastery checklist and missed problems)
3	• Day 1: Unit 1 test • Day 2: Lesson 2.1 Measures of Center (day 1 of 2) • Day 3: Lesson 2.1 Measures of Center (day 2 of 2) • Day 4: Lesson 2.2 Quartiles, the Five-Number Summary and IQR (day 1 of 2) • Day 5: Lesson 2.2 Quartiles, the Five-Number Summary and IQR (day 2 of 2)
4	• Day 1: Lesson 2.3 Boxplots and Outliers (day 1 of 2) • Day 2: Lesson 2.3 Boxplots and Outliers (day 2 of 2) • Day 3: Lesson 2.4 Standard Deviation (day 1 of 2) • Day 4: Lesson 2.4 Standard Deviation (day 2 of 2) • Day 5: Lesson 2.5 Comparing Distributions (day 1 of 2)
5	• Day 1: Lesson 2.5 Comparing Distributions (day 2 of 2) • Day 2: Lesson 2.6 Transforming Data (day 1 of 2) • Day 3: Lesson 2.6 Transforming Data (day 2 of 2) • Day 4: Unit 2 review (mastery checklist and missed problems) • Day 5: Unit 2 test
6	• Day 1: Lesson 3.1 Percentiles and z-Scores (day 1 of 2) • Day 2: Lesson 3.1 Percentiles and z-Scores (day 2 of 2) • Day 3: Lesson 3.2 Density Curves (day 1 of 2) • Day 4: Lesson 3.2 Density Curves (day 2 of 2) • Day 5: Lesson 3.3 Normal Distributions and the 68–95–99.7 Rule (day 1 of 2)

Week	Plan
7	• Day 1: Lesson 3.3 Normal Distributions and the 68–95–99.7 Rule (day 2 of 2) • Day 2: Lesson 3.4 Finding Normal Proportions (day 1 of 2) • Day 3: Lesson 3.4 Finding Normal Proportions (day 2 of 2) • Day 4: Lesson 3.5 Finding Values from Normal Proportions (day 1 of 2) • Day 5: Lesson 3.5 Finding Values from Normal Proportions (day 2 of 2)
8	• Day 1: Lesson 3.6 Assessing Normality (day 1 of 2) • Day 2: Lesson 3.6 Assessing Normality (day 2 of 2) • Day 3: Unit 3 review (mastery checklist and missed problems) • Day 4: Unit 3 test • Day 5: Lesson 4.1 Scatterplots and Association (day 1 of 2)
9	• Day 1: Lesson 4.1 Scatterplots and Association (day 2 of 2) • Day 2: Lesson 4.2 The Correlation Coefficient (day 1 of 2) • Day 3: Lesson 4.2 The Correlation Coefficient (day 2 of 2) • Day 4: Lesson 4.3 The Least-Squares Regression Line (day 1 of 2) • Day 5: Lesson 4.3 The Least-Squares Regression Line (day 2 of 2)
10	• Day 1: Lesson 4.4 Residuals and r^2 (day 1 of 2) • Day 2: Lesson 4.4 Residuals and r^2 (day 2 of 2) • Day 3: Lesson 4.5 Cautions about Regression and Causation (day 1 of 2) • Day 4: Lesson 4.5 Cautions about Regression and Causation (day 2 of 2) • Day 5: Unit 4 review (mastery checklist and missed problems)
11	• Day 1: Unit 4 test • Day 2: Lesson 5.1 Populations, Samples and Bias (day 1 of 2) • Day 3: Lesson 5.1 Populations, Samples and Bias (day 2 of 2) • Day 4: Lesson 5.2 Random Sampling Methods (day 1 of 2) • Day 5: Lesson 5.2 Random Sampling Methods (day 2 of 2)
12	• Day 1: Lesson 5.3 Problems with Surveys (day 1 of 2) • Day 2: Lesson 5.3 Problems with Surveys (day 2 of 2) • Day 3: Lesson 5.4 Observational Studies and Experiments (day 1 of 2) • Day 4: Lesson 5.4 Observational Studies and Experiments (day 2 of 2) • Day 5: Lesson 5.5 Designing Experiments (day 1 of 2)
13	• Day 1: Lesson 5.5 Designing Experiments (day 2 of 2) • Day 2: Lesson 5.6 Blocking, Matched Pairs and the Scope of Conclusions (day 1 of 2) • Day 3: Lesson 5.6 Blocking, Matched Pairs and the Scope of Conclusions (day 2 of 2) • Day 4: Unit 5 review (mastery checklist and missed problems) • Day 5: Unit 5 test

Week	Plan
14	• Day 1: Statistics project, part 1: choose a question and plan the study • Day 2: Statistics project, part 1: begin collecting data • Day 3: Lesson 6.1 Randomness and Simulation (day 1 of 2) • Day 4: Lesson 6.1 Randomness and Simulation (day 2 of 2) • Day 5: Lesson 6.2 Probability Rules (day 1 of 2) *Project week: choose a question you care about and a way to collect honest data. Keep collecting in spare time over the next weeks.*
15	• Day 1: Lesson 6.2 Probability Rules (day 2 of 2) • Day 2: Lesson 6.3 Conditional Probability (day 1 of 2) • Day 3: Lesson 6.3 Conditional Probability (day 2 of 2) • Day 4: Lesson 6.4 Independence and the Multiplication Rule (day 1 of 2) • Day 5: Lesson 6.4 Independence and the Multiplication Rule (day 2 of 2)
16	• Day 1: Lesson 6.5 Tree Diagrams (day 1 of 2) • Day 2: Lesson 6.5 Tree Diagrams (day 2 of 2) • Day 3: Lesson 6.6 Counting: Permutations and Combinations (day 1 of 2) • Day 4: Lesson 6.6 Counting: Permutations and Combinations (day 2 of 2) • Day 5: Unit 6 review (mastery checklist and missed problems)
17	• Day 1: Unit 6 test • Day 2: Midterm review (Units 1–6) • Day 3: Midterm exam (Units 1–6) • Day 4: Lesson 7.1 Discrete Random Variables (day 1 of 2) • Day 5: Lesson 7.1 Discrete Random Variables (day 2 of 2) *Midterm week. Use any spare time to correct missed problems from Units 1–6.*
18	• Day 1: Lesson 7.2 Expected Value and Standard Deviation (day 1 of 2) • Day 2: Lesson 7.2 Expected Value and Standard Deviation (day 2 of 2) • Day 3: Lesson 7.3 Transforming and Combining Random Variables (day 1 of 2) • Day 4: Lesson 7.3 Transforming and Combining Random Variables (day 2 of 2) • Day 5: Lesson 7.4 Normal Random Variables and Their Sums (day 1 of 2)
19	• Day 1: Lesson 7.4 Normal Random Variables and Their Sums (day 2 of 2) • Day 2: Unit 7 review (mastery checklist and missed problems) • Day 3: Unit 7 test • Day 4: Lesson 8.1 Binomial Settings and the Binomial Formula (day 1 of 2) • Day 5: Lesson 8.1 Binomial Settings and the Binomial Formula (day 2 of 2)

Week	Plan
20	• Day 1: Lesson 8.2 Binomial Mean, Standard Deviation and Cumulative Probabilities (day 1 of 2) • Day 2: Lesson 8.2 Binomial Mean, Standard Deviation and Cumulative Probabilities (day 2 of 2) • Day 3: Lesson 8.3 The Normal Approximation to the Binomial (day 1 of 2) • Day 4: Lesson 8.3 The Normal Approximation to the Binomial (day 2 of 2) • Day 5: Lesson 8.4 Geometric Distributions (day 1 of 2)
21	• Day 1: Lesson 8.4 Geometric Distributions (day 2 of 2) • Day 2: Unit 8 review (mastery checklist and missed problems) • Day 3: Unit 8 test • Day 4: Lesson 9.1 Parameters, Statistics and Sampling Variability (day 1 of 2) • Day 5: Lesson 9.1 Parameters, Statistics and Sampling Variability (day 2 of 2)
22	• Day 1: Lesson 9.2 The Sampling Distribution of a Sample Proportion (day 1 of 2) • Day 2: Lesson 9.2 The Sampling Distribution of a Sample Proportion (day 2 of 2) • Day 3: Lesson 9.3 The Sampling Distribution of a Sample Mean (day 1 of 2) • Day 4: Lesson 9.3 The Sampling Distribution of a Sample Mean (day 2 of 2) • Day 5: Lesson 9.4 The Central Limit Theorem (day 1 of 2)
23	• Day 1: Lesson 9.4 The Central Limit Theorem (day 2 of 2) • Day 2: Unit 9 review (mastery checklist and missed problems) • Day 3: Unit 9 test • Day 4: Lesson 10.1 The Idea of a Confidence Interval (day 1 of 2) • Day 5: Lesson 10.1 The Idea of a Confidence Interval (day 2 of 2)
24	• Day 1: Lesson 10.2 Estimating a Population Proportion (day 1 of 2) • Day 2: Lesson 10.2 Estimating a Population Proportion (day 2 of 2) • Day 3: Lesson 10.3 The t Distributions (day 1 of 2) • Day 4: Lesson 10.3 The t Distributions (day 2 of 2) • Day 5: Lesson 10.4 Estimating a Population Mean (day 1 of 2)
25	• Day 1: Lesson 10.4 Estimating a Population Mean (day 2 of 2) • Day 2: Lesson 10.5 Margin of Error and Sample Size (day 1 of 2) • Day 3: Lesson 10.5 Margin of Error and Sample Size (day 2 of 2) • Day 4: Unit 10 review (mastery checklist and missed problems) • Day 5: Unit 10 test
26	• Day 1: Lesson 11.1 The Logic of Significance Tests (day 1 of 2) • Day 2: Lesson 11.1 The Logic of Significance Tests (day 2 of 2) • Day 3: Lesson 11.2 Testing a Claim about a Proportion (day 1 of 2) • Day 4: Lesson 11.2 Testing a Claim about a Proportion (day 2 of 2) • Day 5: Lesson 11.3 Testing a Claim about a Mean (day 1 of 2)

Week	Plan
27	• Day 1: Lesson 11.3 Testing a Claim about a Mean (day 2 of 2) • Day 2: Lesson 11.4 Type I and Type II Errors and Power (day 1 of 2) • Day 3: Lesson 11.4 Type I and Type II Errors and Power (day 2 of 2) • Day 4: Lesson 11.5 Tests and Intervals Together (day 1 of 2) • Day 5: Lesson 11.5 Tests and Intervals Together (day 2 of 2)
28	• Day 1: Unit 11 review (mastery checklist and missed problems) • Day 2: Unit 11 test • Day 3: Lesson 12.1 Comparing Two Proportions (day 1 of 2) • Day 4: Lesson 12.1 Comparing Two Proportions (day 2 of 2) • Day 5: Lesson 12.2 Comparing Two Means (day 1 of 2)
29	• Day 1: Lesson 12.2 Comparing Two Means (day 2 of 2) • Day 2: Lesson 12.3 Paired Data (day 1 of 2) • Day 3: Lesson 12.3 Paired Data (day 2 of 2) • Day 4: Lesson 12.4 Choosing the Right Procedure (day 1 of 2) • Day 5: Lesson 12.4 Choosing the Right Procedure (day 2 of 2)
30	• Day 1: Unit 12 review (mastery checklist and missed problems) • Day 2: Unit 12 test • Day 3: Lesson 13.1 The Chi-Square Goodness-of-Fit Test (day 1 of 2) • Day 4: Lesson 13.1 The Chi-Square Goodness-of-Fit Test (day 2 of 2) • Day 5: Lesson 13.2 Chi-Square Tests for Two-Way Tables (day 1 of 2)
31	• Day 1: Lesson 13.2 Chi-Square Tests for Two-Way Tables (day 2 of 2) • Day 2: Lesson 13.3 Thinking Wisely about Statistics • Day 3: Unit 13 review (mastery checklist and missed problems) • Day 4: Unit 13 test • Day 5: Statistics project, part 2: organize the data and make graphs
32	• Day 1: Statistics project, part 2: analyze (summaries, and an interval or test where it fits) • Day 2: Statistics project, part 2: write and present the report • Day 3: Final review (Units 1–6) • Day 4: Final review (Units 7–13) • Day 5: Final exam (Units 1–13) *Final exam week.*

The Statistics Project

In this project you carry out one complete statistical study of your own, from a question to a careful conclusion. Choose something you can measure honestly over a few weeks: a garden, a flock, a sports team, the weather or a family business.

Steps

1. Ask a clear question about one variable ("What is the mean weight of our hens' eggs?") or a comparison ("Do seeds of brand A germinate more often than seeds of brand B?").
2. Decide whether you will take a sample or run an experiment, and plan how chance will choose your sample or assign your treatments (random digits, drawing names, a random number generator).
3. Collect your data honestly over a few weeks, recording every value with its units and noting anything unusual, without leaving out results you do not like.
4. Make graphs (dotplots, histograms, boxplots, bar graphs or scatterplots) and compute summaries (counts, percents, means, medians, standard deviations) that fit your data.
5. When your data allow it, use a confidence interval or significance test from Units 10–13, and check its conditions first.
6. State your conclusions in context, and state their limits: what your data can and cannot show, and to which population or setting they apply.

The written report

- Write 2–4 pages, neatly organized, with your graphs included.
- The question and why it interested you.
- How you collected the data: the individuals, the variables, the sampling or randomization, and any problems that came up.
- Graphs and summaries, with a short description of what they show.
- Your analysis (an interval or test if you used one, with its conditions) and a conclusion in context.
- What you would do differently if you did the study again.

Project ideas

- Compare the egg production or egg weights of two breeds of hens.
- Compare the germination rates of two brands of seed.
- Study the relationship between weekly rainfall and a garden's yield.
- Compare free-throw percentages with two different shooting routines.
- Compare reading speed on printed pages and on a screen.
- Compare how high bread dough rises at two different temperatures.
- Compare daily high temperatures with the forecast made the day before.

- Count meteors seen in one hour on clear nights and describe the distribution.
- Track the weekly weight gain of calves, lambs or goat kids.
- Study attendance or donations at a parish or community event over several weeks.

Materials

- **Pencil and eraser.** Work in pencil so you can correct mistakes neatly.
- **A notebook.** Keep all practice, corrections and tests in one place.
- **Scientific calculator with one- and two-variable statistics.** For example the TI-30XS MultiView; needed from Unit 2 on for means, standard deviations, regression and r.
- **The tables at the back of this course.** Standard normal, t and chi-square tables; every problem can be done with them and a scientific calculator.
- **Graph paper and a ruler.** For dotplots, histograms, boxplots and scatterplots.
- **A coin, a die and a deck of cards.** For probability experiments and simulations in Unit 6.
- **Index cards.** For practice habits, key facts and formulas to review.
- **Optional: a graphing calculator (TI-83/84) or a free online statistics calculator, and a spreadsheet.** Not required, but they make Units 3 and 8–13 quicker and help with the project.

Grading suggestion

Grades combine daily practice, unit tests, the project and the two exams, with the most weight on the unit tests.

Daily practice sets	15%
Unit tests	45%
Statistics project	10%
Midterm exam	15%
Final exam	15%

Scale: A 90–100, B 80–89, C 70–79, D 60–69, below 60 redo

- Grade daily practice for completion and for corrections: a set counts as done when every problem is attempted and every mistake has been corrected.
- Written interpretations count as much as numbers: an answer needs its meaning in context and its units for full credit.
- A student who scores below 70 on a unit test may study the missed material and take a retake with fresh problems; record the higher score, up to 80.

- Give partial credit for correct work and correct reasoning even when the final number is wrong.
- Accept answers that differ only in the last decimal place because of tables or rounding.

A NOTE TO PARENTS

You do not need to know statistics to teach this course well. Every lesson explains its ideas step by step, every practice answer is provided, and every unit test and exam has a full answer key. Your role is to keep the work steady and to ask good questions.

Help without doing the work for your student. Ask: "What is the question asking?" "What are the individuals and the variable?" "What does that number mean in a sentence?" "How were the data collected?" "Could that be chance?" Make daily practice and the correction of every mistake a habit, and value consistency over speed: a little careful work each day builds lasting understanding.

Statistics matters far beyond the classroom. It helps your student read news reports and advertising claims wisely, and it is used every day in the sciences, in farming and business decisions, and in medicine. Above all, it trains the habit of honest reasoning about uncertain things: weighing evidence fairly, saying exactly what the data show, and admitting what they do not.

UNIT 1

Exploring Data: Variables and Graphs

Statistics is the science of learning from data when the values vary, and variation is everywhere: in rainfall, crop yields, egg weights and batting averages. This first unit teaches the basic vocabulary of data (individuals, variables, populations and samples) and the first and most important tool for understanding data: a well-made graph.

Students learn to display categorical data with frequency tables, bar graphs and pie charts, and to study two categorical variables together with two-way tables and conditional percents. They then display quantitative data with dotplots, stemplots and histograms, and learn to describe a distribution in words: its shape, center, spread and outliers, always in context. Along the way they learn to spot misleading graphs and to ask "compared to what?"

Everything later in the course builds on this unit. Unit 2 adds numerical summaries (mean, median, quartiles, standard deviation) to the pictures drawn here, Unit 4 extends graphs to relationships between two quantitative variables, and the habit of plotting the data first and describing it in context will be used in every unit through the final chi-square tests.

PREREQUISITES

- Work with fractions, decimals and percents, including converting among them.
- Find a percent of a number and the percent one number is of another.
- Read and draw a number line and simple graphs on a coordinate grid.
- Find the mean, median and range of a small set of numbers (from Pre-Algebra).
- Square a number and use area to compare the sizes of similar figures.

MASTERY CHECKLIST

☐ I can identify the individuals and variables in a data set and classify each variable as categorical or quantitative, discrete or continuous.

☐ I can tell a population from a sample and a parameter from a statistic, and I can describe the four steps of a statistical study.

☐ I can make frequency and relative frequency tables, bar graphs and pie charts, and decide when a pie chart makes sense.

☐ I can explain how a graph misleads when its axis does not start at 0 or when a picture grows in two directions.

☐ I can find marginal and conditional percents from a two-way table, using the right denominator.

☐ I can decide whether two categorical variables are associated by comparing conditional percents, and explain why percents, not counts, are compared.
☐ I can make and read dotplots, stemplots (including split stems and back-to-back stemplots) and histograms with equal class widths.
☐ I can explain how a histogram differs from a bar graph and choose a sensible class width.
☐ I can describe the shape, center, spread, outliers, clusters and gaps of a distribution in context, with units.
☐ I can write a paragraph comparing two distributions using comparison words.

LESSON 1.1 · 45 MINUTES

What Is Statistics?

OBJECTIVES

- Explain what statistics is and why variation makes it necessary.
- Identify the individuals and variables in a data set and classify each variable as categorical or quantitative, and each quantitative variable as discrete or continuous.
- Distinguish a population from a sample and a parameter from a statistic.
- Describe the four steps of a statistical study for a simple question.

VIDEO

WATCH

Data and Statistics (4 of 4: Four Types of Data)

Eddie Woo · 5:32

When to watch: Before Example 1: categorical and quantitative variables, with examples of each.

https://www.youtube.com/watch?v=YuBxz45vY-U

BACKUP

Introduction to Sampling & Populations (1 of 4: Why samples?)

Eddie Woo · 8:33

When to watch: Before Example 2: why we study a sample to learn about a population.

https://www.youtube.com/watch?v=Ia_L0Nnjn6g

KEY TERMS

Statistics	The science of learning from data: collecting, organizing, describing and drawing conclusions from data when the values vary.
Individuals and variables	The individuals are the objects a data set describes (people, animals, fields, days, boards). A variable is any characteristic of an individual that can take different values from one individual to the next.
Categorical variable	A variable that places each individual into a group or category, such as breed, color or voice part. Numbers used only as labels (jersey numbers, ZIP codes) are categorical.

Quantitative variable	A variable whose values are numbers that measure or count something, so that adding or averaging them makes sense. It is discrete if it counts (0, 1, 2, ...) and continuous if it measures and could take any value in a range (weight, time, rainfall).
Population and sample	The population is the entire group we want to learn about. A sample is the part of the population we actually collect data from.
Parameter and statistic	A parameter is a number that describes the population (usually unknown). A statistic is a number computed from a sample; we use statistics to estimate parameters.

CORE IDEAS

If every hen laid the same number of eggs every week and every July brought the same rainfall, there would be nothing to study: one measurement would tell you everything. But the real world **varies**. Two rows of the same corn, planted on the same day, give different yields; a good batter gets a hit one day and not the next. **Statistics** is the science of learning from data in the presence of this variation. Its central question is always: how much of what we see is a real pattern, and how much is ordinary variation?

Every data set describes some **individuals** and records one or more **variables** about them. The table below shows one week of records for six hens in the Miller family's flock. The individuals are the hens; each column after the name is a variable. Notice the variation even here: six hens under the same care laid anywhere from 4 to 7 eggs.

The Miller family's laying hens (one week)

Hen	Breed	Egg color	Age (years)	Eggs laid last week	Weight (pounds)
Daisy	Rhode Island Red	Brown	2	6	6.5
Pearl	Leghorn	White	1	7	4.2
Hazel	Plymouth Rock	Brown	3	5	7.1
Clover	Leghorn	White	2	6	4.5
Ginger	Rhode Island Red	Brown	4	4	6.8
Maple	Plymouth Rock	Brown	1	6	6.9

Variables come in two kinds, and the kind decides which graphs and calculations make sense. A **categorical** variable sorts individuals into groups: breed and egg color are categorical. A **quantitative** variable is a number that counts or measures something: age, eggs laid and weight are quantitative. A good test: does it make sense to average the values? The average weight of the hens means something;

the "average breed" does not. Numbers can be categorical too: a jersey number is only a name tag, so averaging jersey numbers is meaningless.

Quantitative variables are **discrete** when they count (eggs laid can be 4 or 5, never 4.37) and **continuous** when they measure and could take any value in a range (a hen's weight could be 6.5 pounds, 6.52 pounds, and so on, as precisely as the scale allows). Rainfall, time, length and temperature are continuous; numbers of people, games or meteors are discrete.

We usually want to know about a whole **population**, such as every calf born in the county this year or every board cut at a sawmill today. Measuring all of them is often too costly or impossible, so we study a **sample**. A number that describes the population, such as the mean weight of all the calves, is a **parameter**; it is fixed but usually unknown. A number computed from the sample is a **statistic**; it is known, but it would change if we took a different sample. Much of this course is about how far a statistic can be trusted as an estimate of a parameter, and the answer comes from understanding variation.

A statistical study follows **four steps**. (1) **Ask** a clear question about a population. (2) **Collect** data that can answer it, choosing individuals and variables with care. (3) **Analyze** the data with graphs and numbers. (4) **Conclude** by answering the original question in its context, in plain words, with honest attention to how much the answer could vary. Unit 1 starts on step 3 with graphs; Unit 5 returns to step 2, and the later units build the tools for step 4.

From history. The word "statistics" grew from the same root as "state": it first meant collecting facts about a state, such as its people and its harvests. In 1693 the English astronomer Edmond Halley (for whom the comet is named) used the birth and death records of the city of Breslau to build one of the first life tables, showing how many people of each age could be expected to live another year. It was an early example of finding a dependable pattern in data that vary from person to person.

Practice habit. Before you analyze any data, ask three questions: Who are the individuals? What are the variables? What are the units of each quantitative variable?

WORKED EXAMPLES

Example 1. Use the Miller family's hen records. (a) What are the individuals? (b) Which variables are categorical? (c) Is "eggs laid last week" discrete or continuous? (d) Is "weight" discrete or continuous?

The Miller family's laying hens (one week)

Hen	Breed	Egg color	Age (years)	Eggs laid last week	Weight (pounds)
Daisy	Rhode Island Red	Brown	2	6	6.5
Pearl	Leghorn	White	1	7	4.2
Hazel	Plymouth Rock	Brown	3	5	7.1
Clover	Leghorn	White	2	6	4.5
Ginger	Rhode Island Red	Brown	4	4	6.8
Maple	Plymouth Rock	Brown	1	6	6.9

1. Individuals are the things the rows describe. Each row is one hen, so the individuals are the six hens.
2. The hen's name only labels the row; the variables are breed, egg color, age, eggs laid and weight.
3. Breed and egg color sort the hens into groups, and averaging them makes no sense, so they are categorical.
4. Age, eggs laid and weight are numbers that count or measure, so they are quantitative.
5. Eggs laid is a count (4, 5, 6, 7, never 5.3), so it is discrete.
6. Weight is a measurement that could be any value in a range (6.5, 6.52, ...), so it is continuous.

Answer: (a) The six hens. (b) Breed and egg color. (c) Discrete. (d) Continuous.

Example 2. A county extension agent wants to know the mean weaning weight of all 1,200 beef calves born in the county this spring. The agent weighs 50 calves chosen at random and finds that their mean weight is 512 pounds. (a) Identify the population and the sample. (b) What is the value of the statistic? (c) What parameter does it estimate, and why can't the agent report 512 pounds as the exact mean weight of all the calves?

1. The population is the whole group the question is about: all 1,200 beef calves born in the county this spring.
2. The sample is the part actually measured: the 50 calves that were weighed.
3. A statistic is a number computed from the sample: the sample mean, 512 pounds.
4. The parameter is the matching number for the population: the mean weaning weight of all 1,200 calves.
5. Calves vary in weight. A different random 50 calves would give a somewhat different sample mean, so 512 pounds is an estimate, not the exact population mean.
6. Conclusion in context: the agent's best estimate of the mean weaning weight of all the county's spring calves is about 512 pounds, and later units will show how to say how close that estimate is likely to be.

Answer: (a) Population: all 1,200 spring calves in the county; sample: the 50 calves weighed. (b) 512 pounds (the sample mean). (c) It estimates the mean weight of all 1,200 calves. Because calves vary, another sample of 50 would give a different mean, so 512 pounds is only an estimate of the population mean.

PRACTICE

1. EASY A beekeeper records the number of frames filled with honey in each hive. Is this variable categorical or quantitative?
2. EASY A gardener records the color of each tulip that blooms (red, yellow, white or pink). Is this variable categorical or quantitative?
3. EASY Is the number of people attending Sunday service each week a discrete or a continuous variable?
4. EASY Is the time it takes a runner to finish a mile a discrete or a continuous variable?
5. MEDIUM A soccer coach records each player's jersey number. Is jersey number categorical or quantitative? Explain in one sentence.

6. MEDIUM The table shows county fair records for some 4-H members. Not counting the member's name, which only labels each row, how many individuals and how many variables does the table describe?

County fair records for six 4-H members

Member	Animal raised	Age (years)	Animal's weight (pounds)	Ribbon
Samuel	Lamb	14	118	Blue
Ruth	Goat	12	86	Red
Henry	Steer	16	1240	Blue
Grace	Lamb	13	124	White
Anna	Pig	15	268	Red
Thomas	Goat	11	79	Blue

7. MEDIUM A choir director records three things about each member: voice part (soprano, alto, tenor or bass), the number of whole years the member has sung in the choir, and the member's height in inches. (a) Is voice part categorical or quantitative? (b) Is years in the choir discrete or continuous? (c) Is height discrete or continuous?

8. MEDIUM A sawmill cuts 2,000 boards on Tuesday. To check whether the saw is set correctly, the foreman measures the lengths of 40 of those boards chosen at random. (a) What is the population? (b) How many individuals are in the sample?

9. MEDIUM The church secretary counts every family in the parish directory and finds 142 families. Then the secretary phones 25 families chosen at random and learns that 9 of them have teenagers. Is 142 a parameter or a statistic? Is 9 out of 25 a parameter or a statistic? Explain.

10. HARD Two rows of the same corn variety, planted on the same day in the same field, yield 152 and 161 bushels per acre. Explain why this happens, and why it matters to a farmer who wants to compare two different seed varieties.

11. HARD The table shows July rainfall at the Carter farm for eight years. (a) In how many of the years was July rainfall more than 4.0 inches? (b) What is the range (largest minus smallest), in inches? (c) Explain why the farmer cannot plan on "the" July rainfall as a single number.

July rainfall at the Carter farm

Year	July rainfall (inches)
2018	3.2
2019	4.8
2020	2.1
2021	5.6
2022	3.9
2023	4.4
2024	1.7
2025	3.5

12. HARD On five clear nights in August, an astronomy club counted 12, 18, 9, 15 and 21 meteors in one hour. (a) Is "meteors counted in one hour" discrete or continuous? (b) Find the mean number of meteors per hour for these five nights. (c) The club wants to know the mean hourly count for all clear August nights. Is your answer to (b) a parameter or a statistic? Explain.

13. HARD Ruth wonders whether her tomato plants produce more fruit when watered in the morning than when watered in the evening. Describe briefly what she could do in each of the four steps of a statistical study.

14. CHALLENGE Henry wants to know whether Rhode Island Red hens in his county lay heavier eggs than Leghorn hens. Name the population(s) he cares about, a sensible sample, the variables he must record for each egg (and the type of each), and one parameter and one statistic in his study.

COMMON MISTAKES

- Calling a number-coded variable quantitative. Jersey numbers, ZIP codes and phone numbers are labels; ask whether an average would mean anything.
- Mixing up individuals and variables. The individuals are what the rows describe (hens, boards, days); the variables are the columns recorded about them.
- Calling a count continuous. If the values come from counting (eggs, people, meteors), the variable is discrete.
- Treating a sample statistic as if it were the exact population parameter. A statistic is an estimate; a different sample would give a different value.

ANSWERS TO PRACTICE

1. Quantitative: it is a count, and averaging it makes sense.
2. Categorical: it sorts each tulip into a color group.
3. Discrete: it is a count of people (whole numbers only).
4. Continuous: time is measured and could take any value in a range (5 minutes 41.27 seconds, for example).
5. Categorical: the number is only a label that names the player; the average of the jersey numbers would mean nothing.
6. 6 individuals (the six members) and 4 variables (animal raised, age, animal's weight, ribbon).
7. (a) Categorical. (b) Discrete (a count of whole years). (c) Continuous (a measurement).
8. (a) All 2,000 boards cut on Tuesday. (b) 40 boards.
9. 142 is a parameter: it describes the whole population (every family in the directory). 9 out of 25 is a statistic: it comes from a sample of 25 families.
10. Yields vary naturally because of small differences in soil, water, sunlight and pests, even with the same seed. So if one variety beats another by a few bushels in a single comparison, the difference might be ordinary variation rather than a better seed; the farmer needs several plots of each variety and statistical reasoning to tell a real difference from chance.
11. (a) 3 years (4.8, 5.6 and 4.4 inches). (b) $5.6 - 1.7 = 3.9$ inches. (c) July rainfall varies a great deal from year to year (from 1.7 to 5.6 inches), so the farmer should plan for a range of possible rainfall, not a single fixed amount.
12. (a) Discrete (a count). (b) $\dfrac{12 + 18 + 9 + 15 + 21}{5} = \dfrac{75}{5} = 15$ meteors per hour. (c) A statistic: it comes from a sample of five nights and estimates the mean for all clear August nights, which is the parameter.
13. Ask: Do tomato plants watered in the morning produce more pounds of fruit than plants watered in the evening? Collect: water some plants in the morning and some in the evening all summer (otherwise treated alike) and weigh the fruit from each plant. Analyze: make graphs of the weights for each group and compare their typical values and variation. Conclude: answer the question in context, for example "the morning-watered plants produced about 2 pounds more fruit per plant," and say whether the difference is large compared with the plant-to-plant variation.

14. Populations: all eggs laid by Rhode Island Red hens in the county and all eggs laid by Leghorn hens in the county (the individuals are eggs, because egg weight is recorded for each egg). Sample: eggs collected from a number of hens of each breed on several farms. Variables: breed (categorical) and egg weight in grams (quantitative, continuous). Parameter: the mean weight of all eggs laid by the county's Rhode Island Reds. Statistic: the mean weight of the Rhode Island Red eggs in his sample.

LESSON 1.2 · 2 SESSIONS OF 45 MINUTES

Displaying Categorical Data

OBJECTIVES

- Construct frequency and relative frequency tables for a categorical variable.
- Make and read bar graphs and pie charts, and decide when a pie chart makes sense.
- Explain why relative frequencies, not counts, are used to compare groups of different sizes.
- Identify and explain misleading graphs, including a vertical axis that does not start at 0 and pictographs that grow in two directions.

VIDEO

WATCH

Constructing Column/Bar Graphs

Eddie Woo · 12:50

When to watch: Before Example 1: building a bar graph from a frequency table.

https://www.youtube.com/watch?v=RRtjWDkDMk8

BACKUP

Constructing Sector & Divided Bar Graphs

Eddie Woo · 7:18

When to watch: After Example 1: pie charts (sector graphs) and divided bar graphs for the same kind of data.

https://www.youtube.com/watch?v=ap9a4-MJx1I

KEY TERMS

Distribution	A description of what values a variable takes and how often it takes each one.
Frequency table	A table that lists each category and the count (frequency) of individuals in it.
Relative frequency	The fraction, decimal or percent of all individuals that fall in a category: the count in the category divided by the total.
Bar graph	A graph with one separate bar for each category; the height of each bar shows the count or percent in that category.
Pie chart	A circle divided into slices, one per category, whose sizes show each category's share of one whole.

Pictograph A graph that uses pictures (jars, baskets, people) instead of bars to show amounts.

CORE IDEAS

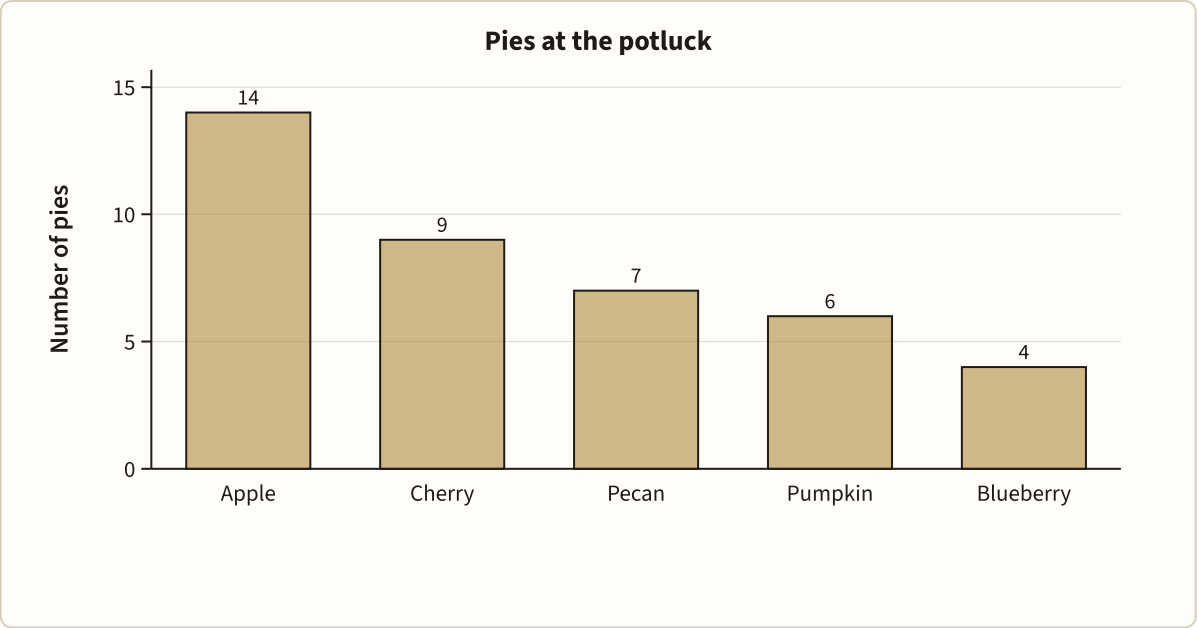
For a categorical variable the data are names of groups, so the first step is to count. The **distribution** of a categorical variable tells which categories occur and how often. At the Hollow Creek church potluck, 40 pies were brought, and the table counts them by kind. The count in each category is its **frequency**; dividing by the total gives the **relative frequency**. For apple, $\frac{14}{40} = 0.35$, or 35%.

Pies at the Hollow Creek church potluck

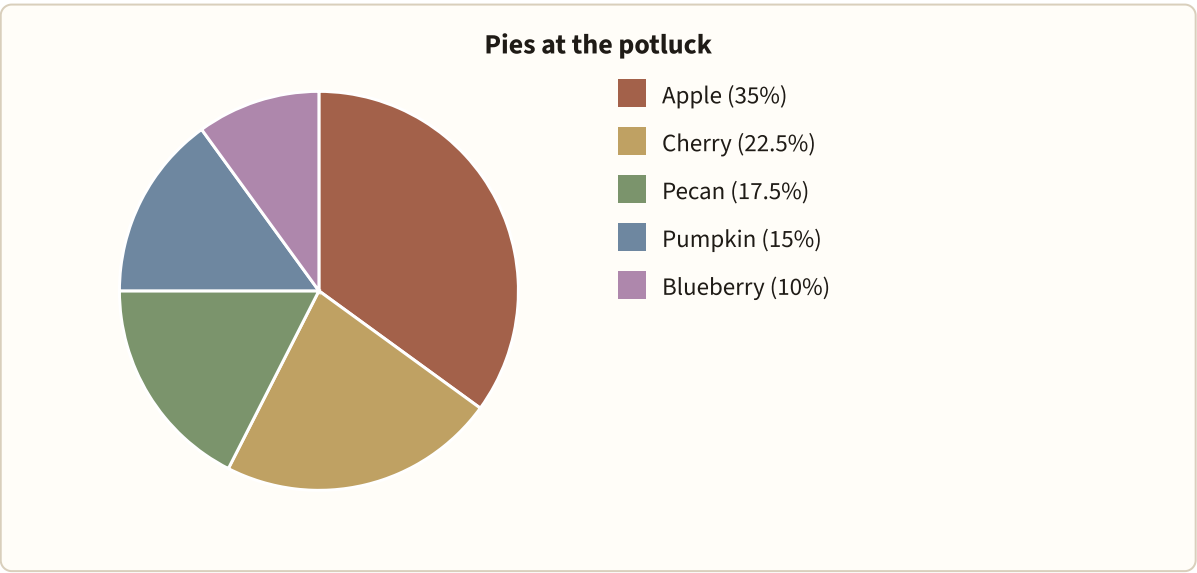
Kind of pie	Count	Relative frequency	Percent
Apple	14	0.35	35%
Cherry	9	0.225	22.5%
Pecan	7	0.175	17.5%
Pumpkin	6	0.15	15%
Blueberry	4	0.1	10%
Total	40	1	100%

Why bother with relative frequencies when we already have counts? Because a count only means something compared with the total. Fourteen apple pies out of 40 is a large share; fourteen out of 400 would be a small one. Relative frequencies put groups of different sizes on the same scale, so they are what we compare. A quick check: the relative frequencies always add to 1 (or 100%), apart from small rounding errors.

A **bar graph** draws one bar per category, with the height showing the count (or percent). The bars are separated by gaps because the categories are separate groups, and they can be listed in any order; putting the tallest first often makes the graph easiest to read. Because our eyes compare bar heights, the vertical axis must start at 0, or the heights will not be in the right proportion.



A **pie chart** shows how one whole is divided into parts. Each slice's angle is its relative frequency times 360 degrees: apple gets $0.35 \times 360 = 126$ degrees. A pie chart makes sense only when the categories are parts of ONE whole: every individual falls in exactly one category, and the percents add to 100%. If people can choose more than one answer ("Which animals do you raise?"), the percents add to more than 100% and a pie chart is impossible; use a bar graph instead. Bar graphs also make it easier to compare categories of nearly equal size.



Graphs can mislead even when the numbers are true. (1) **A vertical axis that does not start at 0.** If bake sale earnings of \$520 and \$560 are drawn on an axis starting at \$500, the bars are 20 and 60 units tall, so the second looks three times as large though it is only about 8% larger. (2) **Pictographs that grow in two directions.** If a picture of a honey jar is drawn twice as tall to show twice the honey, and it is also drawn twice as wide to keep its shape, its area is $2 \times 2 = 4$ times as large. Our eyes judge area, so the growth looks four-fold. Honest graphs start bar axes at 0 and change only one dimension of a picture, or use plain bars.

KEY FORMULAS

$$\text{relative frequency} = \frac{\text{count in the category}}{\text{total count}}$$

$$\text{angle of a pie slice} = (\text{relative frequency}) \times 360^\circ$$

If a picture is scaled by a factor k in both height and width, its area is multiplied by k^2.

From history. The Scottish engineer William Playfair is usually credited with inventing the bar graph, which he published in 1786, and the pie chart, which he published in 1801. Before him, numbers were nearly always shown in tables of figures only.

Practice habit. Ask "compared to what?" Check that the relative frequencies add to 100% (allowing for rounding), and look at where the vertical axis starts before you believe what a graph seems to show.

WORKED EXAMPLES

Example 1. The table shows the 240 books checked out of a parish library in March. (a) What percent of the checkouts were fiction? (b) In a pie chart of these data, what angle (in degrees) would the fiction slice have?

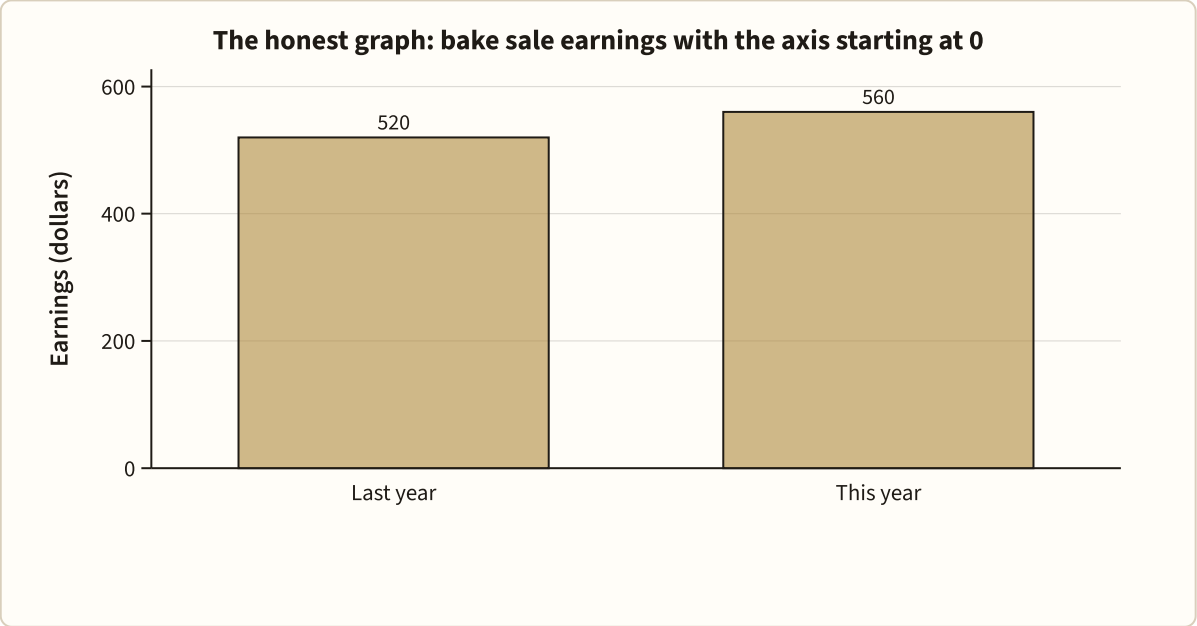
Parish library checkouts in March

Kind of book	Checkouts
Fiction	84
Biography	36
History	30
Science	24
Children's	66
Total	240

1. Relative frequency of fiction: $\frac{84}{240} = 0.35$.
2. As a percent: $0.35 = 35\%$.
3. Check that a pie chart makes sense: each checkout is exactly one kind of book, so the categories are parts of one whole.
4. Slice angle: $0.35 \times 360^\circ = 126^\circ$.
5. Conclusion: 35% of the library's March checkouts were fiction, so fiction fills a 126-degree slice, a little more than a third of the pie.

Answer: (a) $\frac{84}{240} = 0.35$, so 35%. (b) $0.35 \times 360^\circ = 126^\circ$.

Example 2. A church bake sale earned \$520 last year and \$560 this year. A poster shows the two amounts as bars on a vertical axis that starts at \$500. (a) How many times as tall as last year's bar does this year's bar look? (b) By what percent did earnings actually increase? Round to one decimal place. (c) Explain why the poster is misleading.

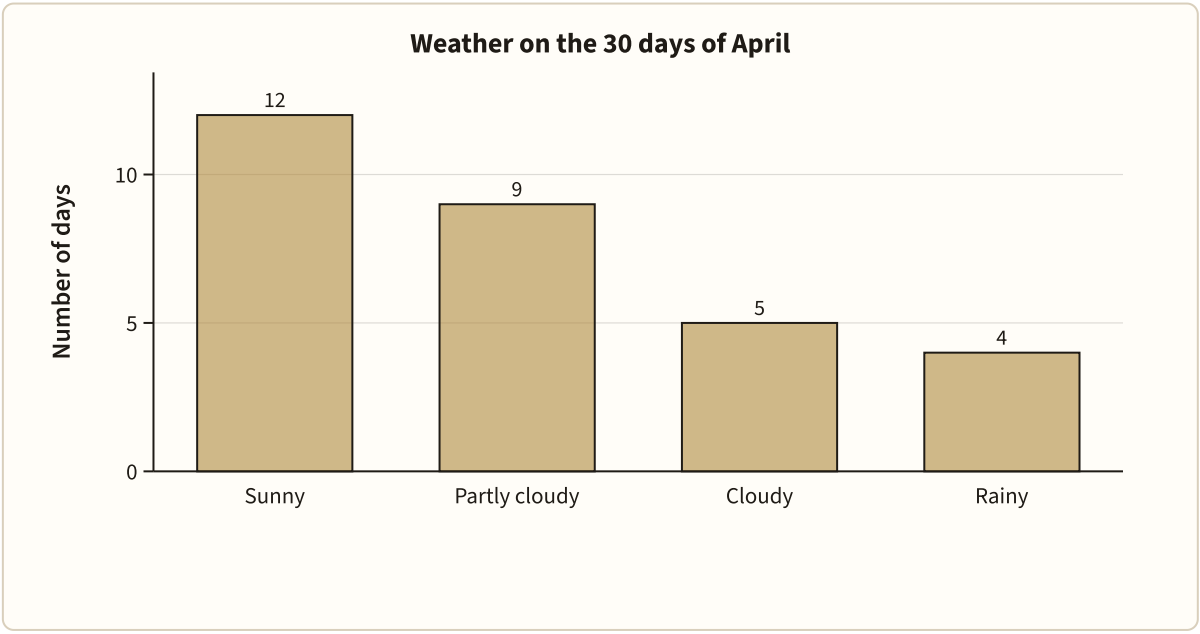


1. On an axis starting at \$500, last year's bar is $520 - 500 = 20$ units tall.
2. This year's bar is $560 - 500 = 60$ units tall.
3. Ratio of heights: $\frac{60}{20} = 3$, so this year's bar looks 3 times as tall.
4. Actual increase: $560 - 520 = 40$ dollars, and $\frac{40}{520} = 0.07692\ldots$, or about 7.7%.
5. The chart shown, drawn with the axis starting at 0, gives the honest picture: the bars are nearly the same height.
6. Conclusion: the poster makes a 7.7% rise look like a tripling, because the eye compares bar heights and the bottom \$500 of each bar was cut off.

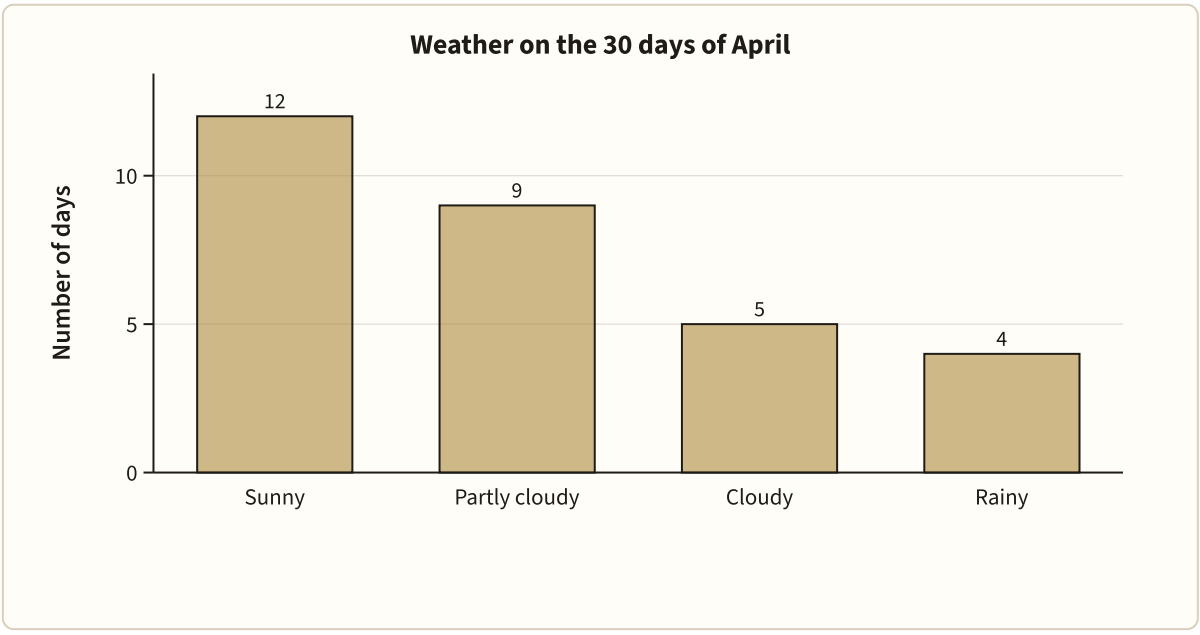
Answer: (a) $\frac{560 - 500}{520 - 500} = \frac{60}{20} = 3$ times as tall. (b) $\frac{40}{520} \approx 0.077$, an increase of about 7.7%. (c) Cutting off the bottom of the axis exaggerates a small increase; the bars should start at 0.

PRACTICE

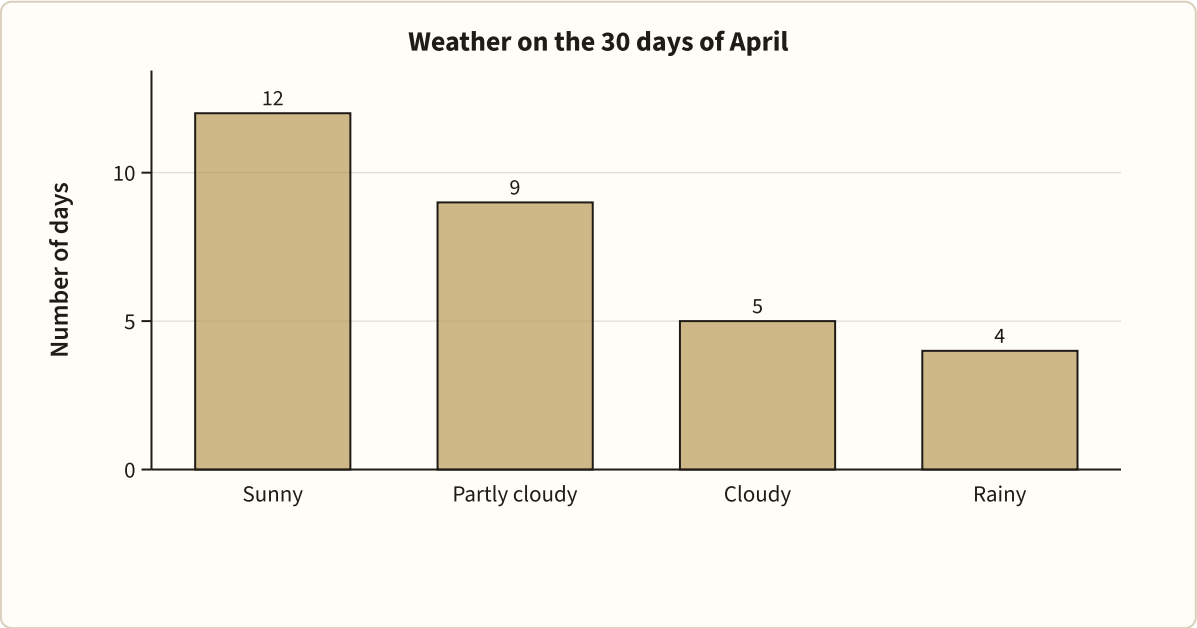
1. EASY The bar graph shows the weather on the 30 days of April at a farm. On how many days was it sunny or partly cloudy?



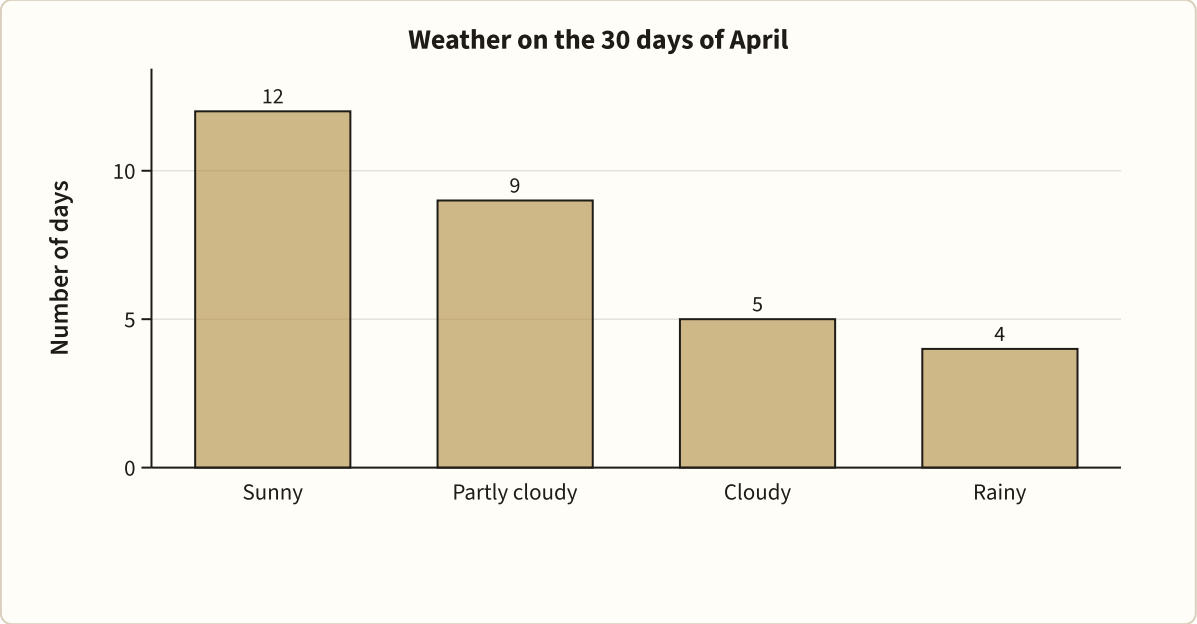
2. EASY Using the April weather graph, what percent of the 30 days were sunny?



3. **EASY** Using the April weather graph, what is the relative frequency of rainy days as a proportion? Round to three decimal places.



4. **EASY** Would a pie chart be a sensible way to display the April weather data? Answer yes or no and give a reason.



5. **MEDIUM** Grace's garden has 80 plants, and 24 of them are tomato plants. In a pie chart of the plants by kind, what angle (in degrees) would the tomato slice have?

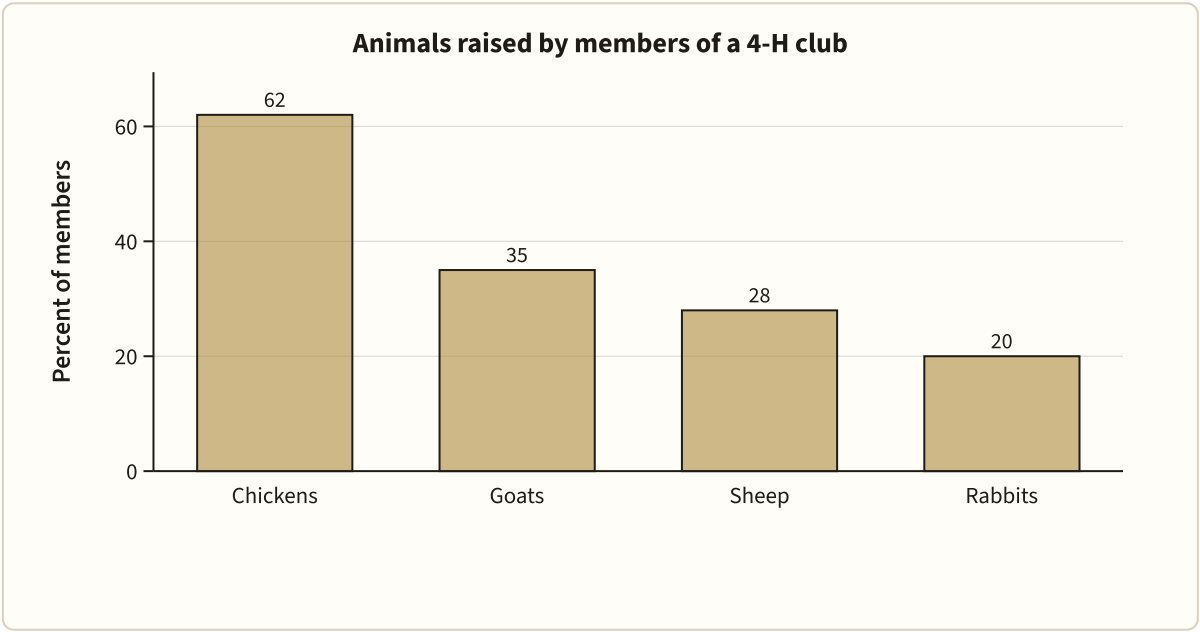
6. **MEDIUM** A choir of 20 members has 7 sopranos, 6 altos, 3 tenors and 4 basses. (a) What percent of the members are altos? (b) What percent sing tenor or bass?

7. MEDIUM Ridgeview Orchard graded 250 apples, as shown in the table. (a) What percent of the apples were graded Utility? (b) How many apples were graded Fancy?

Grades of 250 apples at Ridgeview Orchard

Grade	Percent of apples
Extra Fancy	18%
Fancy	46%
Utility	?
Cull	8%

8. MEDIUM The bar graph shows the percent of members of a 4-H club who raise each kind of animal. Could these percents be displayed in a pie chart? Explain.

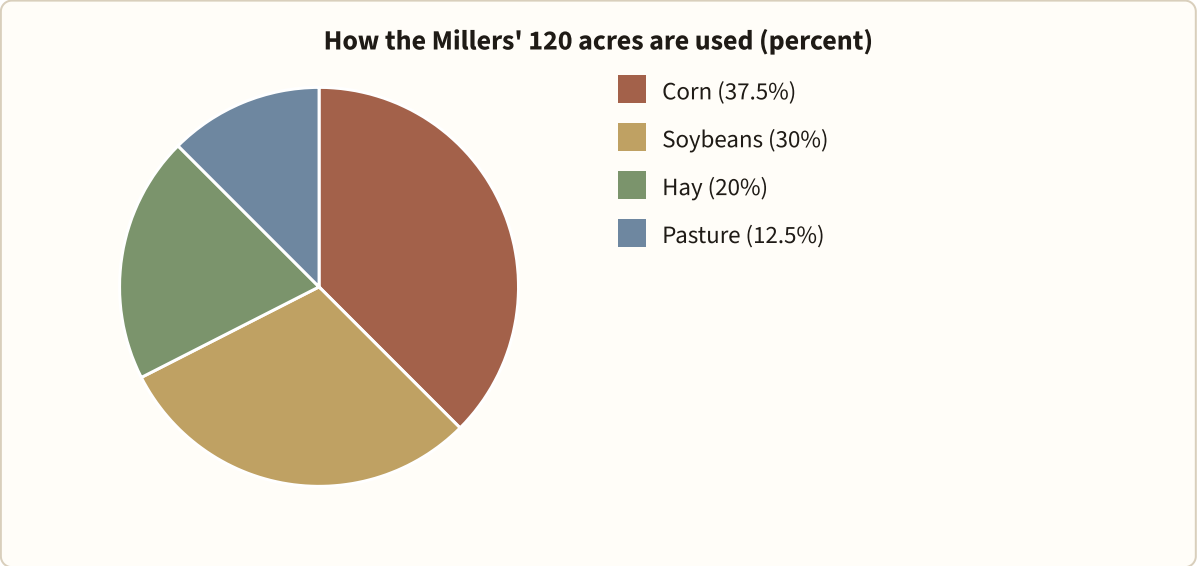


9. MEDIUM Samuel draws a bar graph of the potluck pies with the bars in alphabetical order (Apple, Blueberry, Cherry, Pecan, Pumpkin). Ruth draws the same data with the bars from tallest to shortest. Does the order change the information? Which graph is easier to read, and why?

10. HARD A beekeeper's honey harvest doubled from one year to the next. A poster shows the first year as a picture of a jar, and the second year as the same jar drawn twice as tall and twice as wide. (a) The second jar's area is how many times the first jar's area? (b) Describe an honest way to draw the poster.

11. HARD Field A yielded 48 bushels of wheat per acre and Field B yielded 52. A bar graph shows these yields on a vertical axis that starts at 40 bushels. (a) How many times as tall as Field A's bar does Field B's bar look? (b) How many times as large is Field B's yield really? Round to two decimal places.

12. HARD The pie chart shows how the Miller family uses its 120 acres: corn 37.5%, soybeans 30%, hay 20% and pasture 12.5%. (a) How many acres are in hay? (b) What angle (in degrees) is the corn slice?



13. HARD At harvest, 60 of the 400 apples inspected from Orchard A were bruised, and 45 of the 250 apples inspected from Orchard B were bruised. (a) What percent of Orchard A's apples were bruised? (b) What percent of Orchard B's apples were bruised? (c) Henry says Orchard A has the bigger bruising problem because it had more bruised apples. Is he right? Explain.

14. CHALLENGE A family orchard picked 200 bushels of apples last year and 300 bushels this year. A poster shows last year's harvest as a basket 2 inches tall and this year's as the same basket 3 inches tall, with the width enlarged in the same proportion. (a) By what factor did the harvest actually grow? (b) By what factor did the picture's area grow? (c) Explain why the poster misleads, using your answers.

COMMON MISTAKES

- Comparing counts when the groups have different totals. Convert to percents (relative frequencies) first, then compare.
- Making a pie chart when the categories are not parts of one whole, for example when people could choose more than one answer. Check that the percents add to 100%.
- Trusting bar heights without checking the vertical axis. If it does not start at 0, compare the actual numbers, not the heights.
- Forgetting that a picture enlarged in both height and width grows in area by the square of the scale factor.

ANSWERS TO PRACTICE

1. $12 + 9 = 21$ days.

2. $\frac{12}{30} = 0.40$, so 40%.
3. $\frac{4}{30} \approx 0.133$.
4. Yes. Each of the 30 days falls in exactly one weather category, so the categories are parts of one whole (the month), and the percents add to 100%.
5. $\frac{24}{80} = 0.30$ and $0.30 \times 360^\circ = 108^\circ$.
6. (a) $\frac{6}{20} = 0.30$, so 30%. (b) $\frac{3+4}{20} = \frac{7}{20} = 0.35$, so 35%.
7. (a) The percents must add to 100%: $100 - 18 - 46 - 8 = 28$, so 28%. (b) $0.46 \times 250 = 115$ apples.
8. No. The percents add to $62 + 35 + 28 + 20 = 145\%$, because many members raise more than one kind of animal. The categories are not parts of one whole, so a pie chart cannot show them; a bar graph is the right choice.
9. The order does not change the information: each bar still shows the same count, and the categories have no natural order. Ruth's graph is easier to read because the most and least common kinds stand out at once and neighboring bars are easy to compare.
10. (a) $2 \times 2 = 4$ times the area, so the harvest looks four times as large instead of twice as large. (b) Use two bars on an axis starting at 0, or show the harvest with a row of equal-sized jars (for example 2 jars and 4 jars), or stretch only the height of the picture.
11. (a) The bars are $48 - 40 = 8$ and $52 - 40 = 12$ units tall, and $\frac{12}{8} = 1.5$ times as tall. (b) $\frac{52}{48} \approx 1.08$ times as large (about 8% more).
12. (a) $0.20 \times 120 = 24$ acres. (b) $0.375 \times 360^\circ = 135^\circ$.
13. (a) $\frac{60}{400} = 0.15$, so 15%. (b) $\frac{45}{250} = 0.18$, so 18%. (c) No. Orchard A had more bruised apples only because more of its apples were inspected. Comparing percents, which puts both orchards on the same scale, shows Orchard B had the higher rate of bruising (18% versus 15%).
14. (a) $\frac{300}{200} = 1.5$. (b) Both height and width are multiplied by $\frac{3}{2} = 1.5$, so the area is multiplied by $1.5^2 = 2.25$. (c) Readers judge the size of a picture by its area, so the harvest appears to have grown 2.25 times when it really grew 1.5 times.

LESSON 1.3 · 2 SESSIONS OF 45 MINUTES

Two-Way Tables

OBJECTIVES

- Read a two-way table and find its marginal distributions.
- Compute conditional distributions by choosing the correct group as the denominator.
- Compare conditional distributions to decide whether two categorical variables are associated, and explain the result in context.
- Explain why we compare percents, not counts, when the groups have different sizes.

VIDEO

WATCH

Two-Way Tables (2 of 2: Michael Jordan vs. Reggie Miller)

Eddie Woo · 9:05

When to watch: Before Example 1: reading a two-way table and comparing conditional percents (two basketball players).

https://www.youtube.com/watch?v=67U_kBMfFc4

BACKUP

Two-Way Tables (1 of 2: Basic worked example)

Eddie Woo · 12:23

When to watch: Extra practice: a worked two-way table with row and column totals and percents.

https://www.youtube.com/watch?v=zFx7kiun9cc

KEY TERMS

Two-way table	A table of counts that classifies individuals by two categorical variables at once: one variable for the rows and one for the columns, with totals in the margins.
Marginal distribution	The distribution of one variable by itself, found from the row totals or the column totals in the margins of a two-way table, usually as percents of the grand total.
Conditional distribution	The distribution of one variable among only the individuals who share a given value of the other variable, such as "among the sheep raisers only".

Association	Two variables are associated if knowing the value of one tells you something about the value of the other; in a two-way table, the conditional distributions differ from group to group.
Segmented bar graph	A graph with one bar for each group, each bar standing for 100% of that group and divided into segments that show the conditional percents of the other variable.

CORE IDEAS

Lesson 1.2 described one categorical variable. Often we want to know how two categorical variables are related. A **two-way table** counts individuals by both variables at once. The table below shows 200 members of a county 4-H program classified by the animal they raised and whether they showed it at the county fair. The numbers inside are counts; the totals in the right column and bottom row are in the **margins**.

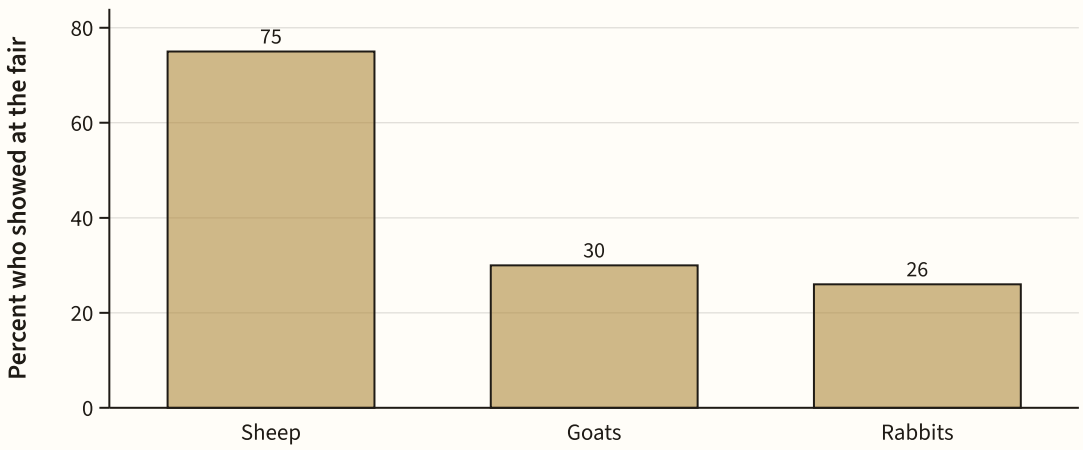
200 county 4-H members

Animal raised	Showed at fair	Did not show	Total
Sheep	30	10	40
Goats	18	42	60
Rabbits	26	74	100
Total	74	126	200

A **marginal distribution** looks at one variable alone, using the totals in the margin. Of all 200 members, 74 showed at the fair: $\frac{74}{200} = 37\%$. By animal, $\frac{40}{200} = 20\%$ raised sheep, 30% raised goats and 50% raised rabbits. Marginal distributions describe each variable, but they cannot show how the two variables are related.

A **conditional distribution** looks at only one group. "Of the 4-H members who raised sheep, what percent showed at the fair?" restricts us to the sheep row: $\frac{30}{40} = 75\%$. For goats, $\frac{18}{60} = 30\%$; for rabbits, $\frac{26}{100} = 26\%$. The rule is simple: the group named after "of the" (or "among the") supplies the denominator. Changing the condition changes the question: "Of the members who showed, what percent raised sheep?" uses the "Showed" column, $\frac{30}{74} \approx 40.5\%$, a different number answering a different question.

Why percents and not counts? Thirty sheep raisers and 26 rabbit raisers showed at the fair, so the counts look almost the same. But there were only 40 sheep raisers and 100 rabbit raisers. Counts mix two things together: how common the outcome is and how big the group is. Dividing by the group size removes the group size, leaving the rate for each group, so 75% against 26% is a fair comparison.



Comparing the conditional distributions shows whether the variables are **associated**. Here the percent who showed is very different for the three animals (75%, 30%, 26%), so animal raised and showing at the fair are associated: knowing which animal a member raised tells you something about whether the member showed. If every group had the same conditional percent, knowing the animal would tell you nothing, and there would be no association. An association does not by itself prove that one variable causes the other; perhaps sheep raisers tend to be older members with more experience.

A **segmented bar graph** pictures conditional distributions. Each group (sheep, goats, rabbits) gets one bar standing for 100% of that group, divided into a "showed" segment and a "did not show" segment. Because every bar has the same total height, the segments can be compared directly even though the groups have different sizes. If the segments line up at the same height in every bar, there is no association; if they differ, there is.

KEY FORMULAS

$$\text{marginal percent} = \frac{\text{row (or column) total}}{\text{grand total}} \times 100\%$$

$$\text{conditional percent} = \frac{\text{count in the cell}}{\text{total of the group you are restricted to}} \times 100\%$$

Practice habit. Before dividing, say the question in the form "Of the ___, what percent ___?" The group in the first blank gives the denominator.

WORKED EXAMPLES

Example 1. Use the table of 200 county 4-H members. (a) What percent of all the members showed at the fair? (b) Of the members who raised goats, what percent showed at the fair? (c) Of the members who showed at the fair, what percent raised sheep? Round to one decimal place where needed.

200 county 4-H members

Animal raised	Showed at fair	Did not show	Total
Sheep	30	10	40
Goats	18	42	60
Rabbits	26	74	100
Total	74	126	200

1. (a) The group is all members, so the denominator is the grand total 200. Showed: 74. $\frac{74}{200} = 0.37$, or 37%.
2. (b) "Of the members who raised goats": the denominator is the goat row total, 60. Goat raisers who showed: 18.
3. $\frac{18}{60} = 0.30$, or 30%.
4. (c) "Of the members who showed": the denominator is the "Showed" column total, 74. Of these, 30 raised sheep.
5. $\frac{30}{74} = 0.4054\ldots$, or about 40.5%.
6. Conclusion: 37% of all members showed; only 30% of goat raisers showed; and about 40.5% of those who showed at the fair were sheep raisers. Parts (b) and (c) use different denominators because they ask about different groups.

Answer: (a) $\frac{74}{200} = 37\%$. (b) $\frac{18}{60} = 30\%$. (c) $\frac{30}{74} \approx 40.5\%$.

Example 2. Ruth stored 200 bean seeds in a cool cellar and 150 in a warm shed over the winter, then planted them all in spring. (a) What percent of the cellar seeds sprouted? (b) What percent of the shed seeds sprouted? (c) Of the seeds that sprouted, what percent had been stored in the cellar? Round to one decimal place. (d) Is there an association between storage place and sprouting?

Bean seeds stored over the winter

Storage	Sprouted	Did not sprout	Total
Cool cellar	168	32	200
Warm shed	99	51	150
Total	267	83	350

1. (a) Restrict to the cellar row: $\frac{168}{200} = 0.84$, or 84%.
2. (b) Restrict to the shed row: $\frac{99}{150} = 0.66$, or 66%.
3. We compare percents, not counts, because the groups have different sizes (200 and 150).
4. (c) Restrict to the "Sprouted" column, total 267: $\frac{168}{267} = 0.6292\ldots$, or about 62.9%.
5. (d) The conditional percents sprouting differ a good deal (84% versus 66%), so the variables are associated.
6. Conclusion: seeds stored in the cool cellar sprouted at a higher rate (84%) than seeds stored in the warm shed (66%), so storage place and sprouting are associated in Ruth's seeds.

Answer: (a) $\frac{168}{200} = 84\%$. (b) $\frac{99}{150} = 66\%$. (c) $\frac{168}{267} \approx 62.9\%$. (d) Yes: 84% of cellar seeds sprouted but only 66% of shed seeds.

PRACTICE

1. EASY The table shows the electives chosen by 90 students at a homeschool co-op. How many grade 11 students chose woodworking?

Electives chosen at a homeschool co-op

Grade	Woodworking	Choir	Astronomy	Total
10	12	13	15	40
11	10	12	8	30
12	8	3	9	20
Total	30	28	32	90

2. EASY Using the co-op electives table, how many students in all chose astronomy?

Electives chosen at a homeschool co-op

Grade	Woodworking	Choir	Astronomy	Total
10	12	13	15	40
11	10	12	8	30
12	8	3	9	20
Total	30	28	32	90

3. EASY Using the co-op electives table, what percent of all 90 students chose astronomy? Round to one decimal place.

Electives chosen at a homeschool co-op

Grade	Woodworking	Choir	Astronomy	Total
10	12	13	15	40
11	10	12	8	30
12	8	3	9	20
Total	30	28	32	90

4. EASY Using the co-op electives table, of the grade 12 students, what percent chose choir?

Electives chosen at a homeschool co-op

Grade	Woodworking	Choir	Astronomy	Total
10	12	13	15	40
11	10	12	8	30
12	8	3	9	20
Total	30	28	32	90

5. MEDIUM Using the co-op electives table, of the students who chose choir, what percent were in grade 10? Round to one decimal place.

Electives chosen at a homeschool co-op

Grade	Woodworking	Choir	Astronomy	Total
10	12	13	15	40
11	10	12	8	30
12	8	3	9	20
Total	30	28	32	90

6. MEDIUM Using the co-op electives table, give the marginal distribution of grade level as percents of all 90 students (grades 10, 11, 12 in order). Round to one decimal place.

Electives chosen at a homeschool co-op

Grade	Woodworking	Choir	Astronomy	Total
10	12	13	15	40
11	10	12	8	30
12	8	3	9	20
Total	30	28	32	90

7. MEDIUM A farmer kept track of the morning forecast and the afternoon weather on 60 summer days. (a) On the days when rain was forecast, what percent of the time did it rain? (b) On the days when no rain was forecast, what percent of the time did it rain? Round to one decimal place.

Morning forecast and afternoon weather on 60 summer days

Forecast	Rained	Did not rain	Total
Rain forecast	16	8	24
No rain forecast	6	30	36
Total	22	38	60

8. MEDIUM Using the forecast table, is there an association between the forecast and whether it rained? Explain using conditional percents.

Morning forecast and afternoon weather on 60 summer days

Forecast	Rained	Did not rain	Total
Rain forecast	16	8	24
No rain forecast	6	30	36
Total	22	38	60

9. MEDIUM Using the forecast table, of the days on which it rained, what percent had a rain forecast? Round to one decimal place.

Morning forecast and afternoon weather on 60 summer days

Forecast	Rained	Did not rain	Total
Rain forecast	16	8	24
No rain forecast	6	30	36
Total	22	38	60

10. MEDIUM The table classifies 100 eggs by the breed of hen that laid them and by size. (a) What percent of the Leghorn eggs were large? (b) What percent of the Rhode Island Red eggs were large? (c) Is there an association between breed and egg size in these data?

Eggs from two breeds of hens

Breed	Large eggs	Medium eggs	Total
Leghorn	45	15	60
Rhode Island Red	30	10	40
Total	75	25	100

11. HARD Of the 120 people at a church picnic, 70 were adults and the rest were children. Forty-five of the adults and 30 of the children played in the softball game. Make a two-way table (adult or child; played or did not play) and use it to answer: (a) What percent of the children played? (b) What percent of the adults played? (c) Of the people who played, what percent were children? Round to one decimal place where needed.

12. HARD In June, 90 of the 300 books lent by Library A were children's books, and 60 of the 150 books lent by Library B were children's books. (a) What percent of Library A's loans were children's books? (b) What percent of Library B's loans were children's books? (c) A volunteer says Library A lends a larger share of children's books because 90 is more than 60. Explain the error.

13. HARD Describe a segmented bar graph for the forecast table: what the bars stand for, how tall the segments would be, and what the graph would show about the association.

Morning forecast and afternoon weather on 60 summer days

Forecast	Rained	Did not rain	Total
Rain forecast	16	8	24
No rain forecast	6	30	36
Total	22	38	60

14. **CHALLENGE** The table shows hits and at bats for two batters. (A batting average is hits divided by at bats.) Henry had the higher average in each half of the season: 0.300 and 0.250 against Samuel's 0.290 and 0.225. (a) Find Henry's batting average for the whole season. (b) Find Samuel's batting average for the whole season. Round to three decimal places. (c) Explain how Samuel can have the higher season average even though Henry was better in each half.

Hits and at bats for two batters

Batter	First half	Second half	Whole season
Henry	12 hits in 40 at bats	25 hits in 100 at bats	37 hits in 140 at bats
Samuel	29 hits in 100 at bats	9 hits in 40 at bats	38 hits in 140 at bats

COMMON MISTAKES

- Dividing by the wrong total. "Of the goat raisers, what percent showed?" divides by the goat row total, not by the grand total or the "Showed" column total.
- Treating "the percent of sheep raisers who showed" and "the percent of those who showed who raised sheep" as the same. They have different denominators and usually different values.
- Comparing counts across groups of different sizes. Always compare conditional percents.
- Concluding that one variable causes the other because they are associated. An association may be explained by other variables.

ANSWERS TO PRACTICE

1. 10 students.
2. $15 + 8 + 9 = 32$ students (the column total).
3. $\frac{32}{90} \approx 0.356$, so about 35.6%.
4. $\frac{3}{20} = 0.15$, so 15%.
5. $\frac{13}{28} \approx 0.464$, so about 46.4%.
6. Grade 10: $\frac{40}{90} \approx 44.4\%$; grade 11: $\frac{30}{90} \approx 33.3\%$; grade 12: $\frac{20}{90} \approx 22.2\%$. (They add to 99.9% because of rounding.)
7. (a) $\frac{16}{24} \approx 66.7\%$. (b) $\frac{6}{36} \approx 16.7\%$.
8. Yes. It rained on about 66.7% of the days when rain was forecast but on only about 16.7% of the days when no rain was forecast. Because these conditional percents are very different, knowing the forecast tells you something about whether it will rain.
9. $\frac{16}{22} \approx 0.727$, so about 72.7%.

10. (a) $\frac{45}{60} = 75\%$. (b) $\frac{30}{40} = 75\%$. (c) No: the conditional percents are the same, so knowing the breed tells you nothing about whether an egg is large. (The counts 45 and 30 differ only because there were more Leghorn eggs.)
11. Table: adults 45 played, 25 did not (70); children 30 played, 20 did not (50); totals 75 played, 45 did not (120). (a) $\frac{30}{50} = 60\%$. (b) $\frac{45}{70} \approx 64.3\%$. (c) $\frac{30}{75} = 40\%$.
12. (a) $\frac{90}{300} = 30\%$. (b) $\frac{60}{150} = 40\%$. (c) The count 90 is larger only because Library A lent twice as many books in all. Comparing percents removes the difference in size and shows that children's books made up a larger share at Library B (40% versus 30%).
13. There would be two bars, one for "rain forecast" days and one for "no rain forecast" days, each standing for 100% of its group. In the first bar the "rained" segment would reach about 66.7% (the rest, 33.3%, "did not rain"); in the second the "rained" segment would reach only about 16.7% (83.3% "did not rain"). The very different segment heights show a strong association between the forecast and the afternoon weather.
14. (a) $\frac{37}{140} \approx 0.264$. (b) $\frac{38}{140} \approx 0.271$. (c) Most of Henry's at bats (100 of 140) came in the second half, when both batters hit worse, while most of Samuel's (100 of 140) came in the stronger first half. Combining the halves weights each batter's average differently, so the season totals reverse the comparison. Looking at the separate groups gives the fairer comparison.

LESSON 1.4 · 2 SESSIONS OF 45 MINUTES

Displaying Quantitative Data

OBJECTIVES

- Construct and read dotplots, stemplots (including split stems and back-to-back stemplots) and histograms.
- Build a frequency table with equal class widths and choose a sensible class width.
- Explain how a histogram differs from a bar graph and why its bars touch.
- Answer questions about a data set from a graph, using counts and percents in context.

VIDEO

WATCH

Histograms

Eddie Woo · 9:37

When to watch: Before Example 2: building and reading a histogram.

https://www.youtube.com/watch?v=WQo_F34vTj4

BACKUP

Stem-and-Leaf Plots

Eddie Woo · 7:48

When to watch: Before Example 1: making a stemplot.

https://www.youtube.com/watch?v=CIJuwHqDgZo

> **Note:** Eddie Woo also shows a cumulative frequency graph (ogive), which this course does not use.

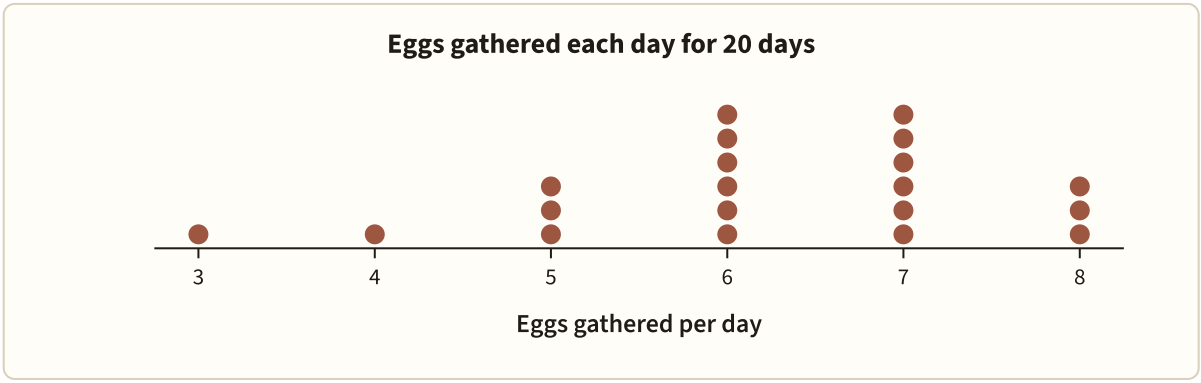
KEY TERMS

Dotplot	A graph with a number line and one dot for each value, stacked when values repeat.
Stemplot	A display that splits each value into a stem (all but the last digit) and a leaf (the last digit), listing the leaves beside their stems so that the actual values can still be read.
Split stems	A stemplot in which each stem appears twice, the first time with leaves 0 to 4 and the second with leaves 5 to 9, used when a plain stemplot has too few stems.

Back-to-back stemplot	Two stemplots that share one column of stems, with one group's leaves to the left and the other's to the right, used to compare two groups.
Histogram	A graph of a quantitative variable that groups values into classes of equal width and draws touching bars whose heights show the count (or percent) of values in each class.
Class	An interval of values in a histogram or frequency table, such as 30 to under 35; in this course a class includes its left edge and excludes its right edge.

CORE IDEAS

A quantitative variable has numbers for values, so its graph needs a number line. Graphs come first because they show the whole **distribution** at once: where the values cluster, how spread out they are, and whether any are unusual. The simplest is the **dotplot**: one dot per value above a number line. Here are the numbers of eggs Ruth gathered each day for 20 days. You can see at a glance that 6 or 7 eggs was typical and that 3 was rare.



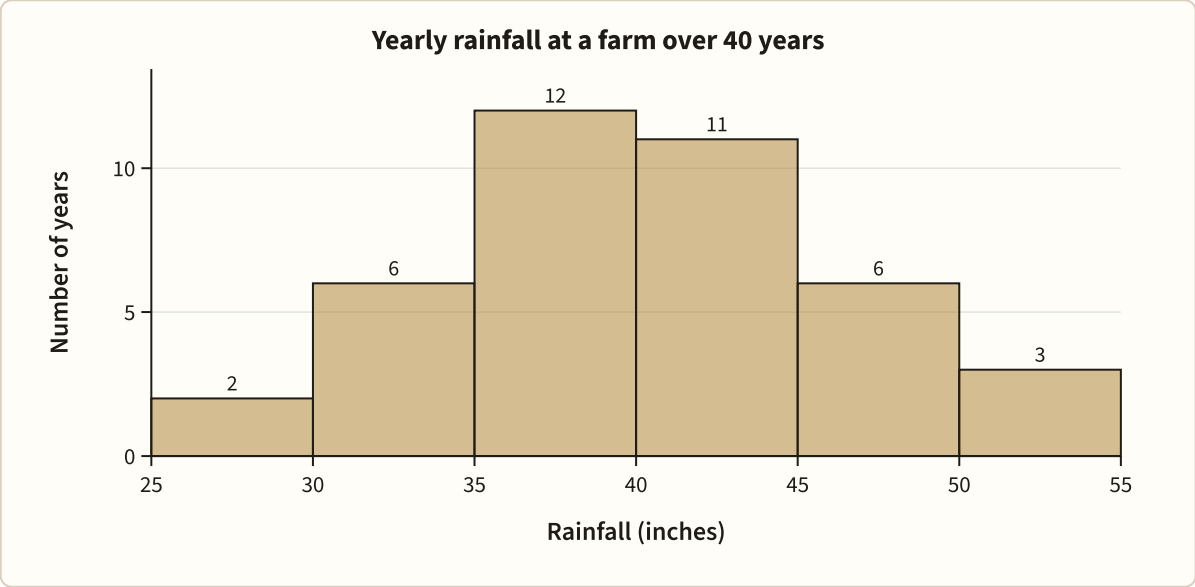
A **stemplot** keeps every actual value while showing the shape. Split each value into a **stem** (all but the last digit) and a **leaf** (the last digit); write the stems in a column, smallest at the top, and each leaf beside its stem in increasing order. Always include a key such as "5 | 2 means 52 pounds". When the data use only two or three stems, everything piles onto a few lines; then **split the stems**, giving leaves 0 to 4 to the first copy of each stem and 5 to 9 to the second. The April high temperatures below span only 51 to 69 degrees, so split stems give four lines instead of two.

Daily high temperatures for 16 days in April (split stems)

5	1 3 4
5	6 7 7 8 9
6	1 2 2 3 4
6	6 8 9

Key: 5 | 1 means 51 degrees Fahrenheit

To compare two groups, a **back-to-back stemplot** puts one group's leaves on the left of a shared stem column and the other's on the right (the left leaves read outward from the stem). For larger data sets, stemplots get crowded, and we use a **histogram**. Divide the range of the data into **classes of equal width**, count the values in each class (a value on a boundary goes into the class that starts there), and draw a bar for each class whose height is the count. Using percents instead of counts gives a relative frequency histogram with exactly the same shape.



Why must the classes have **equal widths**? Our eyes judge a bar by its area. If one class were twice as wide as the others, its bar would look twice as important even with the same count. Why do the bars **touch**? Because the number line has no gaps: the classes 35 to under 40 and 40 to under 45 are neighbors, and every possible value belongs to exactly one class. A bar graph (Lesson 1.2) is different: its bars stand for separate categories that can be put in any order, so they are drawn apart. A histogram's classes have a fixed order along the number line.

Choosing the class width is a judgment call. Too few, wide classes lump everything into one or two bars and hide the shape; too many, narrow classes give a ragged picture with many bars of 0, 1 or 2. For most small data sets, about 5 to 10 classes with round-number edges (widths such as 2, 5, 10 or 20) work well. Try more than one width if you are unsure; a real pattern will show up in each.

One cost of a histogram: individual values are lost. From the rainfall histogram we know that 11 years had between 40 and 45 inches, but not which values they were. Dotplots and stemplots keep every value, which is why they are preferred for small data sets.

KEY FORMULAS

$$\text{class width} \approx \frac{\text{largest value} - \text{smallest value}}{\text{number of classes wanted}}, \text{ rounded up to a convenient number}$$

$$\text{percent in a class} = \frac{\text{count in the class}}{\text{total count}} \times 100\%$$

Practice habit. Always plot the data first. Before any calculation, make a dotplot, stemplot or histogram, and label the axis with the variable and its units.

WORKED EXAMPLES

Example 1. The weaning weights, in pounds, of 12 lambs were 52, 58, 61, 63, 64, 67, 68, 70, 71, 75, 78, 84. (a) Make a stemplot. (b) How many lambs weighed at least 65 pounds?

Weaning weights of 12 lambs

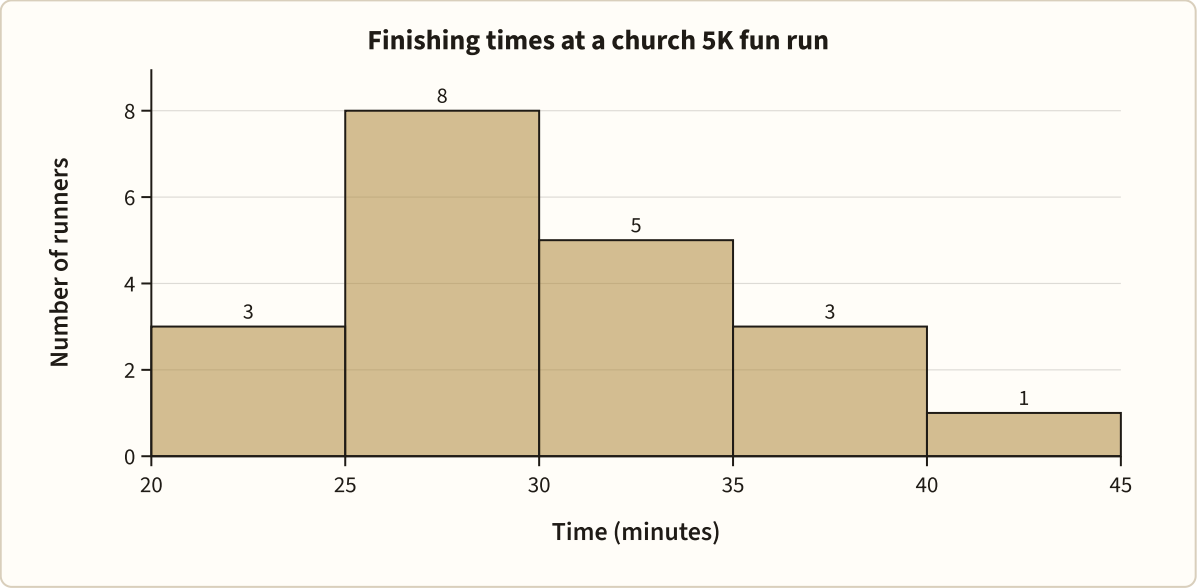
5	2 8
6	1 3 4 7 8
7	0 1 5 8
8	4

Key: 5 | 2 means 52 pounds

1. The values run from 52 to 84, so use the tens digits 5, 6, 7, 8 as stems and the ones digits as leaves.
2. Stem 5: leaves 2, 8 (52 and 58).
3. Stem 6: leaves 1, 3, 4, 7, 8.
4. Stem 7: leaves 0, 1, 5, 8.
5. Stem 8: leaf 4. Add the key "5 | 2 means 52 pounds". The finished stemplot is shown.
6. (b) Count the leaves for 65 or more: on stem 6, the leaves 7 and 8 (2 lambs); all 4 leaves on stem 7; the 1 leaf on stem 8.
7. $2 + 4 + 1 = 7$.
8. Conclusion: 7 of the 12 lambs weighed at least 65 pounds at weaning, and the stemplot shows that most weights fell in the 60s and 70s.

Answer: (a) Stems 5 to 8: 5 | 2 8; 6 | 1 3 4 7 8; 7 | 0 1 5 8; 8 | 4 (key: 5 | 2 means 52 pounds). (b) $2 + 4 + 1 = 7$ lambs.

Example 2. At a church 5K fun run, the finishing times in minutes of 20 runners were 21, 23, 24, 25, 26, 26, 27, 28, 28, 29, 29, 30, 31, 31, 32, 33, 35, 36, 38, 42. Make a histogram using the classes 20 to under 25, 25 to under 30, and so on, and use it to find the percent of runners who finished in under 30 minutes.

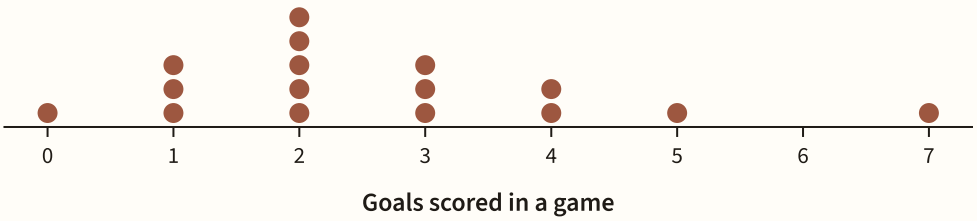


1. The data run from 21 to 42, so classes of width 5 starting at 20 give five classes: 20–25, 25–30, 30–35, 35–40, 40–45, each including its left edge.
2. 20 to under 25: 21, 23, 24, so the count is 3.
3. 25 to under 30: 25, 26, 26, 27, 28, 28, 29, 29, so the count is 8. (The value 25 goes here, in the class that starts at 25.)
4. 30 to under 35: 30, 31, 31, 32, 33, so the count is 5.
5. 35 to under 40: 35, 36, 38, so the count is 3; 40 to under 45: 42, so the count is 1.
6. Check: $3 + 8 + 5 + 3 + 1 = 20$ runners. Draw touching bars with these heights, as shown.
7. Under 30 minutes: $3 + 8 = 11$ runners, and $\dfrac{11}{20} = 0.55$.
8. Conclusion: 55% of the runners finished the 5K in under 30 minutes; the most common times were from 25 to 30 minutes.

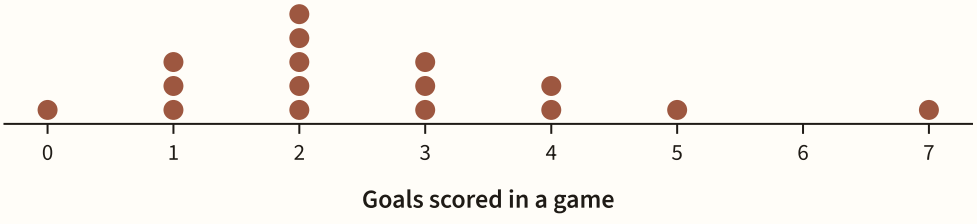
Answer: Counts 3, 8, 5, 3, 1 for the classes 20–25, 25–30, 30–35, 35–40, 40–45 minutes. Under 30 minutes: $\dfrac{3 + 8}{20} = 0.55$, so 55%.

PRACTICE

1. **EASY** The dotplot shows the goals scored by a soccer team in 16 games. In how many games did the team score exactly 2 goals?



2. **EASY** Using the soccer dotplot, in how many games did the team score 3 or more goals?



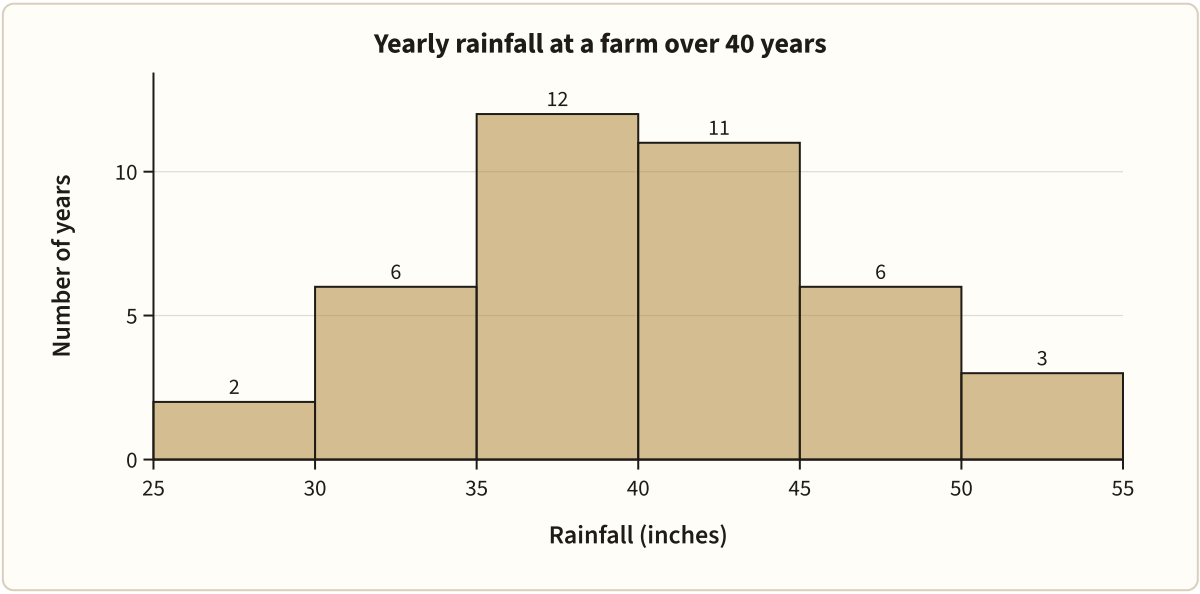
3. **EASY** The back-to-back stemplot shows the weights of 10 pumpkins from each of two patches. What is the weight of the heaviest pumpkin in Patch B?

Pumpkin weights from two patches

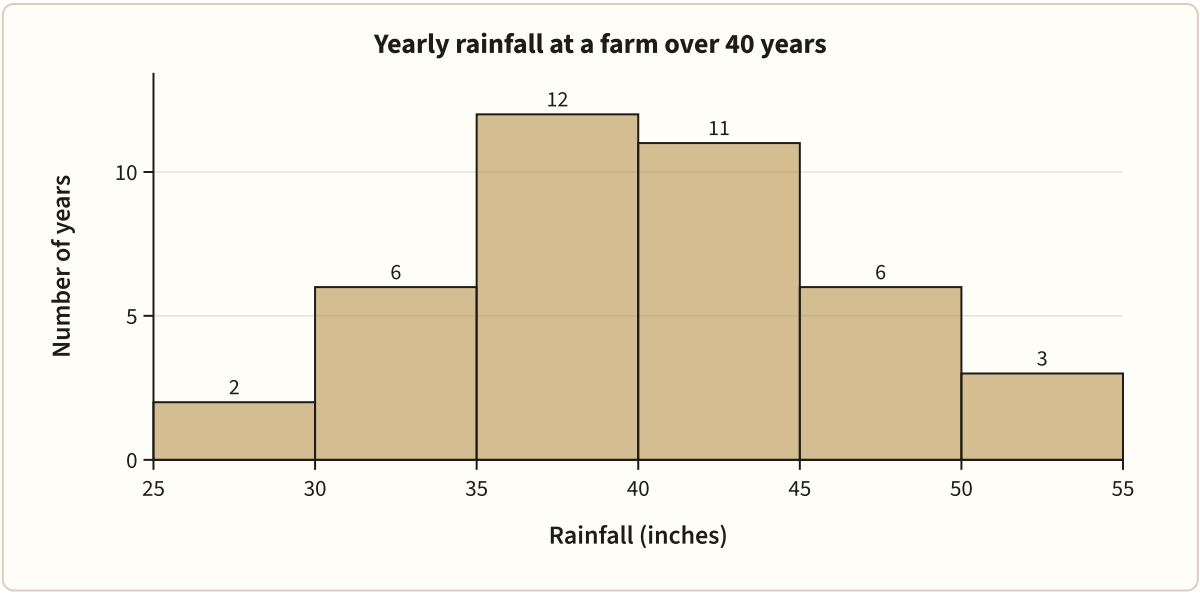
Patch B		Patch A
8	1	2 5 8
9 8 6 3	2	1 2 4 5 7
5 4 2 0	3	1 3
1	4	

Key: 2 | 4 means 24 pounds

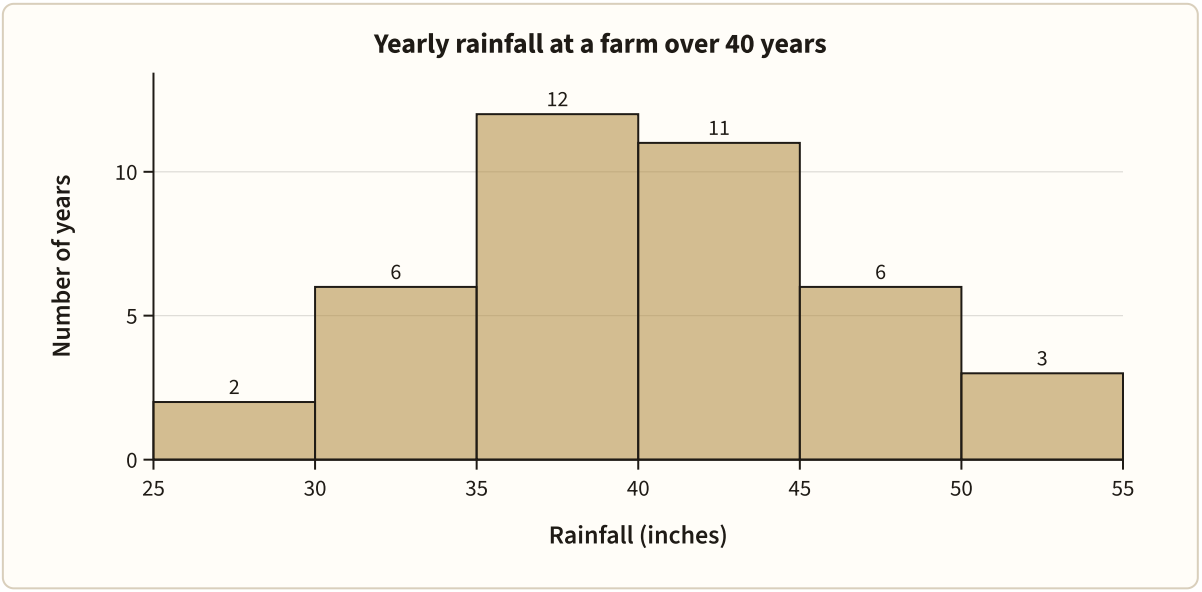
4. EASY The histogram shows the yearly rainfall at a farm over 40 years. In how many years was the rainfall from 40 to under 45 inches?



5. MEDIUM Using the rainfall histogram, what percent of the 40 years had at least 45 inches of rain?



6. MEDIUM Can you find the exact rainfall of the wettest year from the rainfall histogram? Explain.



7. MEDIUM The histogram shows the percent of a farm's eggs in each weight class. What percent of the eggs weighed at least 58 grams?

8. **MEDIUM** Using the egg-weight histogram, if the farm collected 250 eggs, about how many weighed from 62 to under 66 grams?

9. **MEDIUM** Using the pumpkin stemplot, (a) how many Patch A pumpkins weighed less than 20 pounds? (b) Which patch tends to grow heavier pumpkins?

Pumpkin weights from two patches

Patch B					Patch A				
		8	1	2	5	8			
9	8	6	3	2	1	2	4	5	7
5	4	2	0	3	1	3			
		1	4						

Key: 2 | 4 means 24 pounds

10. **MEDIUM** Explain two differences between a histogram and a bar graph, including why the bars of a histogram touch.

11. **HARD** The weights, in pounds, of 15 hay bales were 41, 43, 44, 45, 46, 46, 47, 48, 48, 49, 50, 51, 52, 54, 57. Find the counts for a histogram with the classes 40 to under 44, 44 to under 48, 48 to under 52, 52 to under 56, and 56 to under 60.

12. HARD Using the pumpkin stemplot, what percent of the pumpkins weighed 30 pounds or more (a) in Patch A and (b) in Patch B?

Pumpkin weights from two patches

Patch B		Patch A
8	1	2 5 8
9 8 6 3	2	1 2 4 5 7
5 4 2 0	3	1 3
1	4	

Key: 2 | 4 means 24 pounds

13. HARD Henry says, "In the hay bale histogram from Problem 11, the bale weighing 48 pounds could go in either the 44-to-48 class or the 48-to-52 class." Is he right? Explain the rule and why it matters.

14. CHALLENGE Grace has 40 years of winter snowfall records, ranging from 12 to 87 inches. She first tries a histogram with classes 2 inches wide, then one with classes 40 inches wide. Explain what goes wrong with each choice, and suggest a better class width with its first few classes.

COMMON MISTAKES

- Using classes of unequal width in a histogram. Unequal widths make wide classes look more important; keep every class the same width.
- Counting a boundary value in two classes, or in the wrong one. A class includes its left edge and excludes its right edge.
- Leaving gaps between histogram bars, or drawing a bar graph for quantitative data. Histogram bars touch because the number line is continuous.
- Forgetting the key on a stemplot or writing the leaves out of order. Always give a key with units and list leaves from smallest to largest.

ANSWERS TO PRACTICE

1. 5 games.
2. $3 + 2 + 1 + 1 = 7$ games.
3. 41 pounds.
4. 11 years.
5. $\frac{6 + 3}{40} = \frac{9}{40} = 0.225$, so 22.5%.
6. No. The histogram only shows that 3 years had from 50 to under 55 inches; it groups values into classes, so the individual values are lost.
7. $40 + 24 + 6 = 70\%$.
8. $0.24 \times 250 = 60$ eggs.

9. (a) 3 pumpkins (12, 15 and 18 pounds). (b) Patch B: its leaves sit mostly on the 2 and 3 stems with one 41-pound pumpkin, while Patch A's sit mostly on the 1 and 2 stems.
10. A histogram displays a quantitative variable, with classes in a fixed order along a number line; a bar graph displays a categorical variable, whose categories can be listed in any order. A histogram's bars touch because the classes cover the number line with no gaps between them (one class ends where the next begins); a bar graph's bars are separated because its categories are separate groups.
11. 40–44: 41, 43 (2). 44–48: 44, 45, 46, 46, 47 (5). 48–52: 48, 48, 49, 50, 51 (5). 52–56: 52, 54 (2). 56–60: 57 (1). Check: $2 + 5 + 5 + 2 + 1 = 15$.
12. (a) Patch A: 31 and 33, so $\frac{2}{10} = 20\%$. (b) Patch B: 30, 32, 34, 35, 41, so $\frac{5}{10} = 50\%$.
13. No. Each class includes its left edge and excludes its right edge, so 48 belongs only to the 48-to-under-52 class. A fixed rule makes sure every value is counted exactly once, so the counts add up to the total number of bales.
14. With 2-inch classes there would be about 38 classes for only 40 values, so most bars would be 0, 1 or 2 high and the graph would look ragged, hiding the overall shape. With 40-inch classes there would be only 2 or 3 bars, lumping almost everything together and hiding the shape. A width of 10 inches gives 8 classes (10 to under 20, 20 to under 30, ..., 80 to under 90), enough to show the shape without too much noise.

LESSON 1.5 · 2 SESSIONS OF 45 MINUTES

Describing a Distribution

OBJECTIVES

- Identify the shape of a distribution as symmetric, skewed right or skewed left, and as single-peaked or having more than one peak.
- Describe the center, spread, outliers, clusters and gaps of a distribution from a graph, in context and with units.
- Write a one-paragraph description of a distribution.
- Compare two distributions in a parallel description that uses comparison words.

VIDEO

WATCH

Characterising a Dataset (3 of 3: Which way is data skewed?)

Eddie Woo · 10:00

When to watch: Before Example 1: telling right-skewed from left-skewed data.

https://www.youtube.com/watch?v=c-QoIaidr4Q

BACKUP

Describing Distributions in Statistics

ProfRobBob · 12:55

When to watch: Review: describing shape, center, spread and outliers in context.

https://www.youtube.com/watch?v=uR3MaJdqY3c

KEY TERMS

Symmetric	A distribution is roughly symmetric if the left and right sides of its graph are close to mirror images of each other.
Skewed right and skewed left	A distribution is skewed right if its right side (larger values) stretches out in a long tail, and skewed left if its left side (smaller values) has the long tail. The skew is named for the tail, not for the peak.
Peaks	A distribution is single-peaked if its graph has one clear high point, and has two (or more) peaks if there are separate high points, which often means two different groups are mixed together.

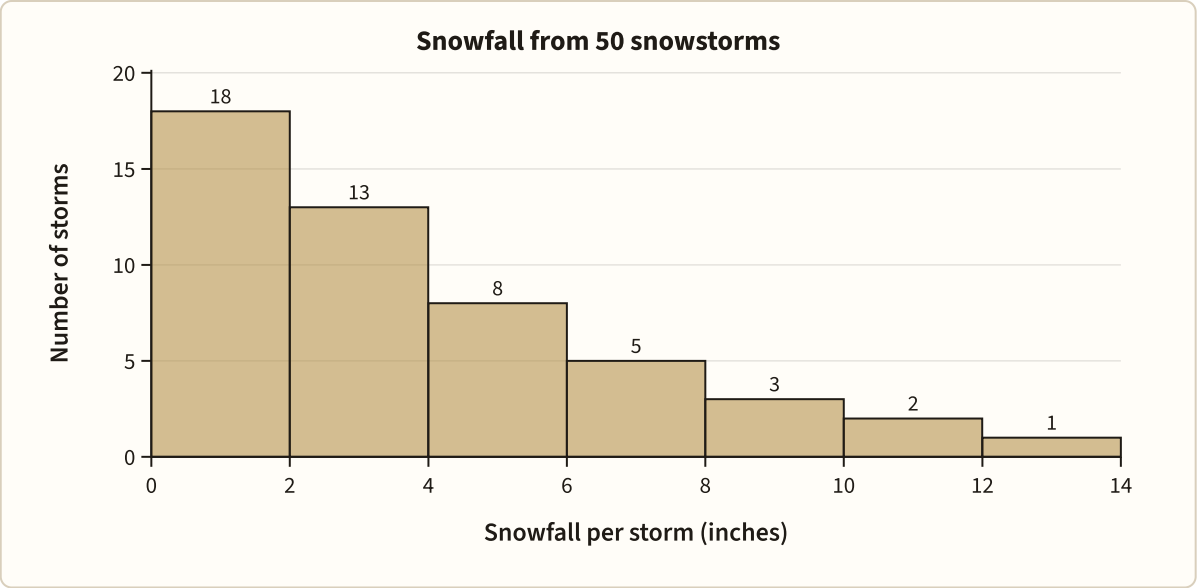
Center and spread	The center is a typical or middle value (for now, the median: the middle value when the data are in order). The spread describes how far the values vary, for now from the smallest to the largest.
Outlier	An individual value that falls clearly outside the overall pattern of the distribution.
Clusters and gaps	A cluster is a group of values close together; a gap is a stretch of the number line with no values, separating clusters.

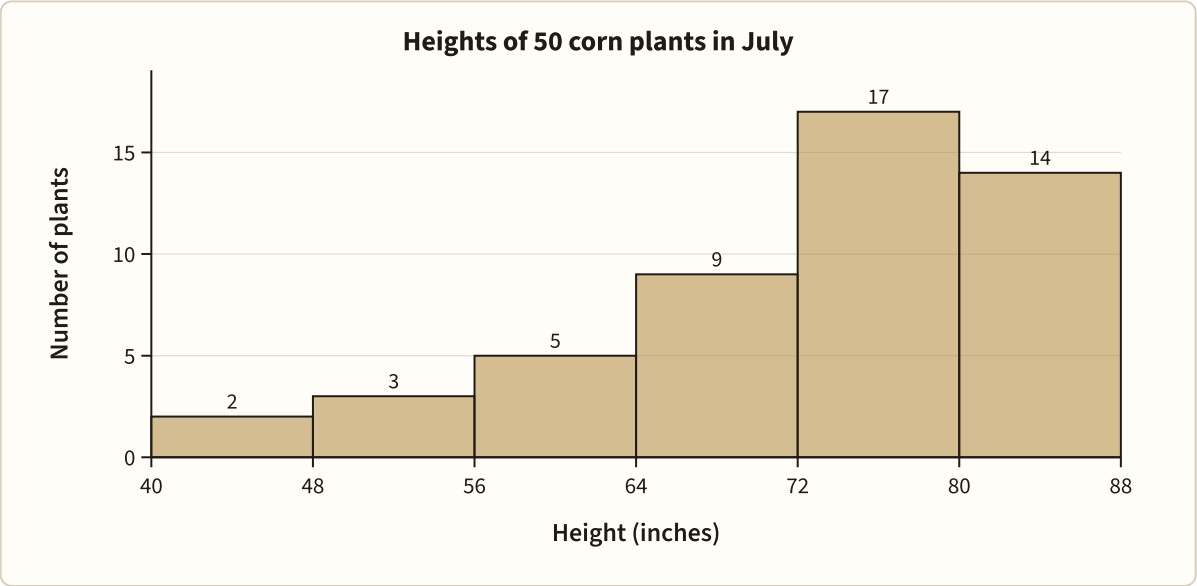
CORE IDEAS

A graph is only half the work; the other half is saying what it shows. To describe the distribution of a quantitative variable, look first for the **overall pattern** and then for **departures** from it. The overall pattern is described by its **shape**, **center** and **spread**; the departures are **outliers**. A helpful memory aid is SOCS: Shape, Outliers, Center, Spread. Every description must be **in context**: name the variable and give its units.

Shape. The 100 egg weights below are roughly **symmetric** and single-peaked: the two sides are near mirror images around about 60 grams. Many measurements of similar things (egg weights, heights of a crop, measurement errors) look like this.

The snowfall from 50 storms is **skewed right**: most storms brought only a few inches, but a few big storms stretch a long tail toward larger values. Right skew is common when values cannot go below 0 but now and then can be large. The heights of 50 corn plants are **skewed left**: most plants grew to full height, but some stunted plants make a tail toward smaller values. Remember that the skew is named for the side with the long, thin **tail**, not for the side with the peak.





Center and spread. For now, describe the center by the **median**, the middle value when the data are in order (or the average of the two middle values when there is an even number), and describe the spread by giving the smallest and largest values or the range. Unit 2 will give more precise measures and explain when each one is the right choice. **Outliers, clusters and gaps.** An outlier is a value that stands clearly apart from the rest; always ask why. It may be a recording mistake, or a real but special case (the library's busiest Saturday was the day of its book sale). Values that fall into separate **clusters** with a **gap** between them usually signal two different kinds of individuals mixed together, and that is worth investigating. (Unit 2 also gives a rule for deciding exactly which values count as outliers.)

Comparing distributions. To compare two groups, write a parallel description: shape, center, spread and outliers for each group, side by side, with **comparison words** such as "greater than", "about the same as" or "more variable than". Saying "Field A's median is 160.5 and Field B's median is 172" lists numbers; saying "Field B's typical yield is about 11.5 bushels per acre higher than Field A's" compares them. Comparing is what answers the question that made us collect the data.

Practice habit. Describe every distribution with SOCS (shape, outliers, center, spread), always in context: name the variable and give the units of every number.

WORKED EXAMPLES

Example 1. The dotplot shows the number of visitors to a parish library on 15 Saturdays. (a) Find the median number of visitors. (b) Describe the distribution.

1. (a) There are 15 values, so the median is the 8th value in order: 18, 20, 21, 21, 22, 23, 23, 23, ...
2. The 8th value is 23, so the median is 23 visitors.
3. (b) Shape: apart from one value, the dots form a single-peaked, roughly symmetric pile from 18 to 28.
4. Outliers: 46 visitors stands far from all the others, an outlier. (It was the Saturday of the library's book sale, a special cause.)
5. Center: the median is 23 visitors.
6. Spread: apart from the outlier, the counts range from 18 to 28 visitors.
7. Conclusion in context: on a typical Saturday the library had about 23 visitors, with counts from 18 to 28; one Saturday, the book sale, was an outlier with 46 visitors.

Answer: (a) 23 visitors. (b) The number of Saturday visitors is roughly symmetric and single-peaked, centered at a median of 23 visitors and ranging from 18 to 28, with one high outlier of 46 visitors (the day of the book sale).

Example 2. A farmer planted two corn varieties on 10 plots each. The back-to-back stemplot shows the yields. (a) Find the median yield of Variety A. (b) Find the median yield of Variety B. (c) Compare the two distributions.

Yields of two corn varieties on 10 plots each

Variety B				Variety A				
			14	8				
		6	15	2	5	8		
	9	6	2	16	0	1	3	5	8
7	4	3	1	17	2				
		5	0	18				

Key: 15 | 2 means 152 bushels per acre

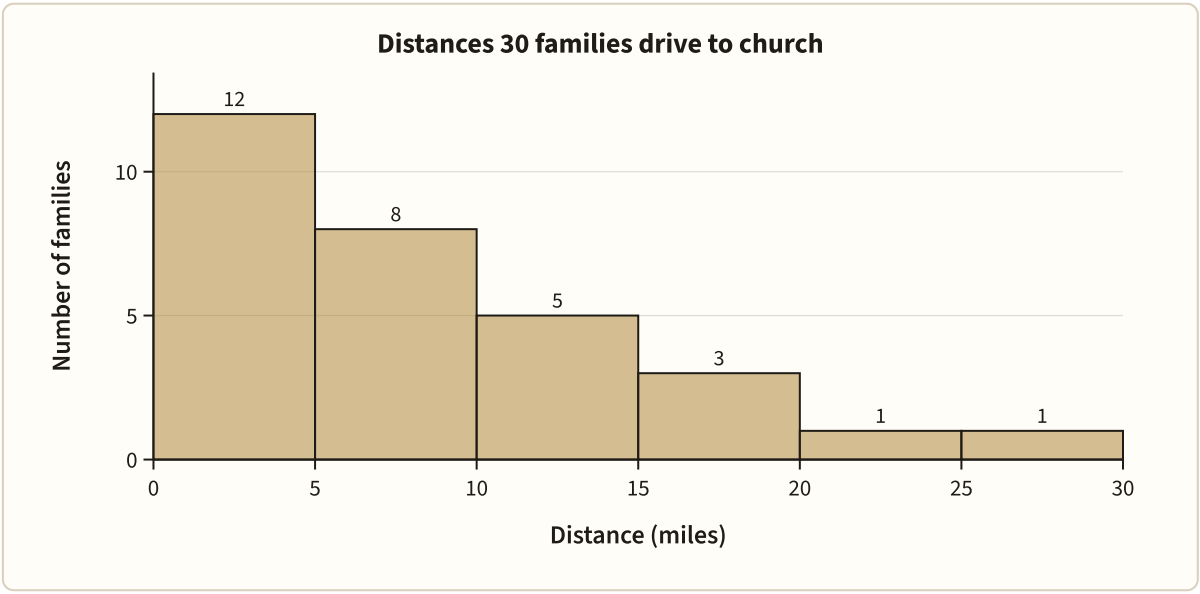
1. (a) Variety A in order: 148, 152, 155, 158, 160, 161, 163, 165, 168, 172. With 10 values, the median is the average of the 5th and 6th: $\frac{160 + 161}{2} = 160.5$ bushels per acre.
2. (b) Variety B in order: 156, 162, 166, 169, 171, 173, 174, 177, 180, 185. Median: $\frac{171 + 173}{2} = 172$ bushels per acre.
3. Shape: both distributions are roughly symmetric and single-peaked.
4. Outliers: neither variety has a value that stands apart from the rest.
5. Center: $172 - 160.5 = 11.5$, so Variety B's typical yield is about 11.5 bushels per acre higher.
6. Spread: Variety A ranges from 148 to 172 (range 24); Variety B from 156 to 185 (range 29), so B is slightly more variable.
7. Conclusion: on these plots, Variety B tended to yield more corn than Variety A, by about 11.5 bushels per acre at the median, with somewhat more variation from plot to plot; both distributions are roughly symmetric with no outliers.

Answer: (a) $\frac{160 + 161}{2} = 160.5$ bushels per acre. (b) $\frac{171 + 173}{2} = 172$ bushels per acre. (c) Both are roughly symmetric with no outliers. Variety B's yields are typically higher (median 172 versus 160.5, about 11.5 bushels per acre more) and slightly more spread out (range 29 versus 24 bushels per acre).

PRACTICE

1. EASY The histogram shows the weights of 40 loaves from a bakery. Is the distribution symmetric, skewed left or skewed right?

2. EASY The histogram shows how far 30 families drive to church. Is the distribution symmetric, skewed left or skewed right?



3. **EASY** The stemplot shows 15 scores on a history test. Is the distribution symmetric, skewed left or skewed right?

Scores on a history test

5	2
6	4
7	1 6
8	0 3 5 7 8
9	0 1 3 4 6 8

Key: 8 | 3 means a score of 83

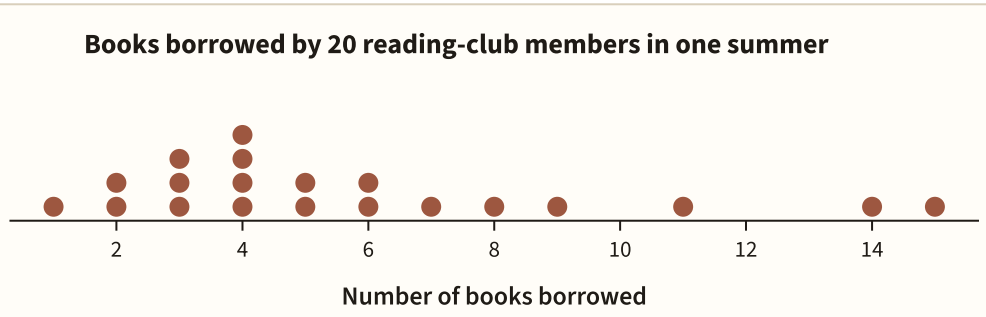
4. **EASY** Using the history test stemplot, find the median score.

Scores on a history test

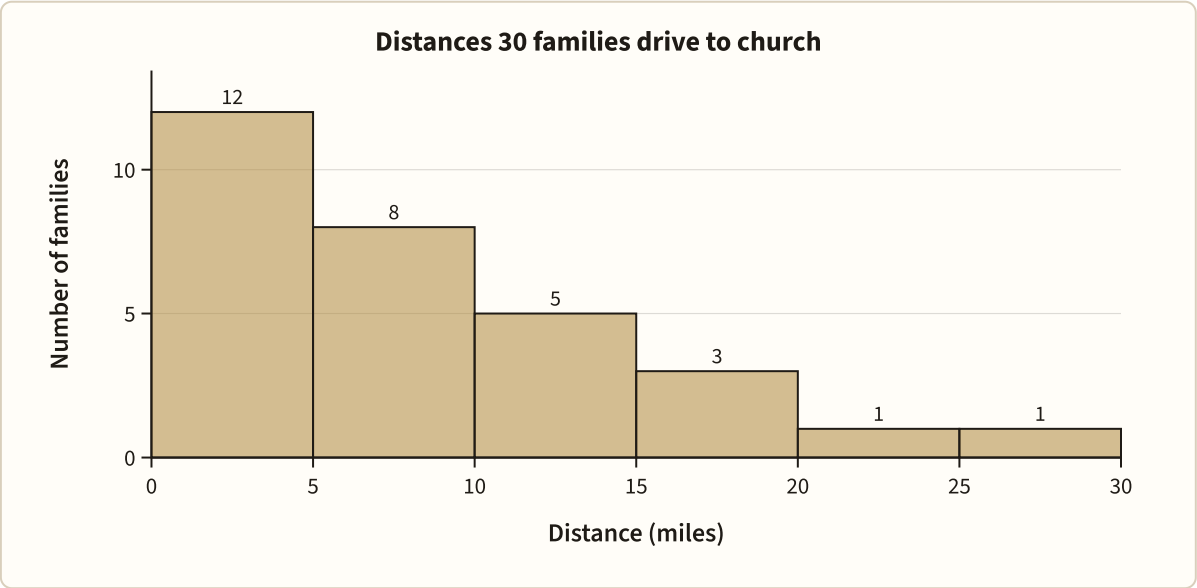
5	2
6	4
7	1 6
8	0 3 5 7 8
9	0 1 3 4 6 8

Key: 8 | 3 means a score of 83

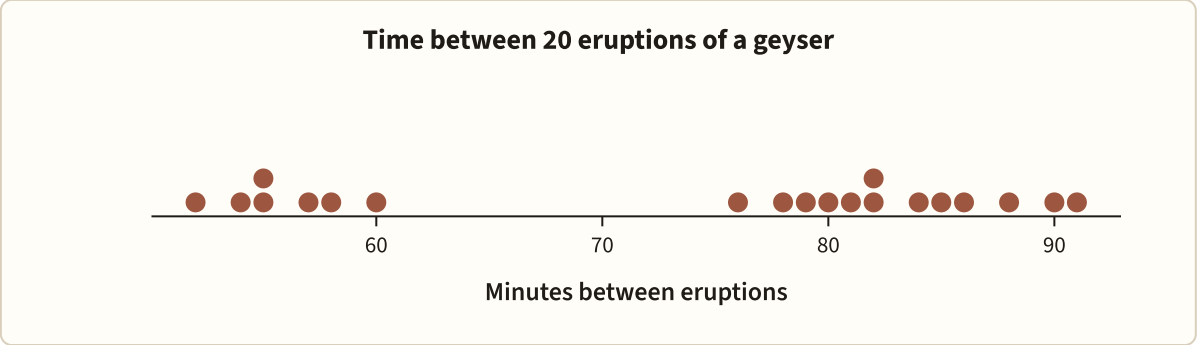
5. **MEDIUM** The dotplot shows the number of books borrowed by 20 members of a summer reading club. (a) Find the median. (b) Find the range.



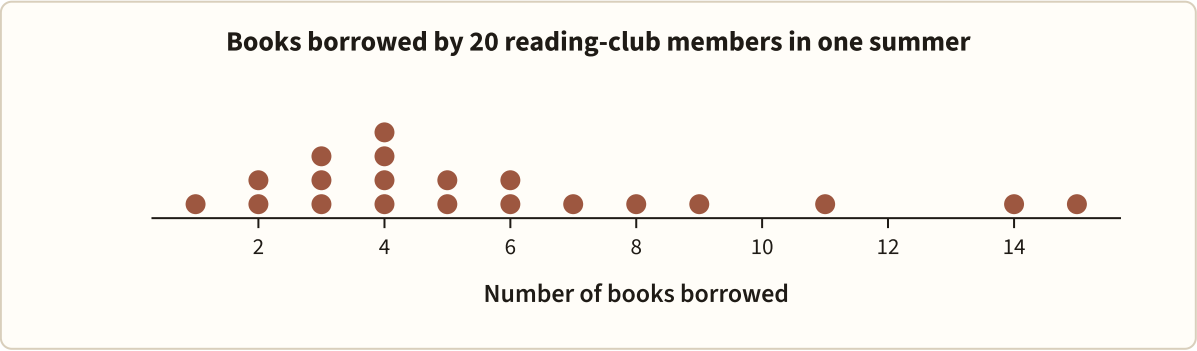
6. MEDIUM Using the church-distance histogram, what percent of the families drive less than 10 miles? Round to one decimal place.



7. MEDIUM The dotplot shows the time between 20 eruptions of a geyser. (a) How many of the times were shorter than 65 minutes? (b) Describe the shape of the distribution.



8. MEDIUM Henry described the reading-club dotplot this way: "The distribution is skewed right with a center of 4.5 and a spread of 14." What is missing from his description? Rewrite it.



9. HARD The weights, in pounds, of 9 watermelons from one garden were 14, 15, 15, 16, 17, 17, 18, 19, 31. (a) Which value is an outlier? (b) Find the median of all 9 weights. (c) Find the median of the other 8 weights, leaving out the outlier.

10. **HARD** Write a complete one-paragraph description (shape, outliers, center, spread, in context) of the bakery loaf weights.

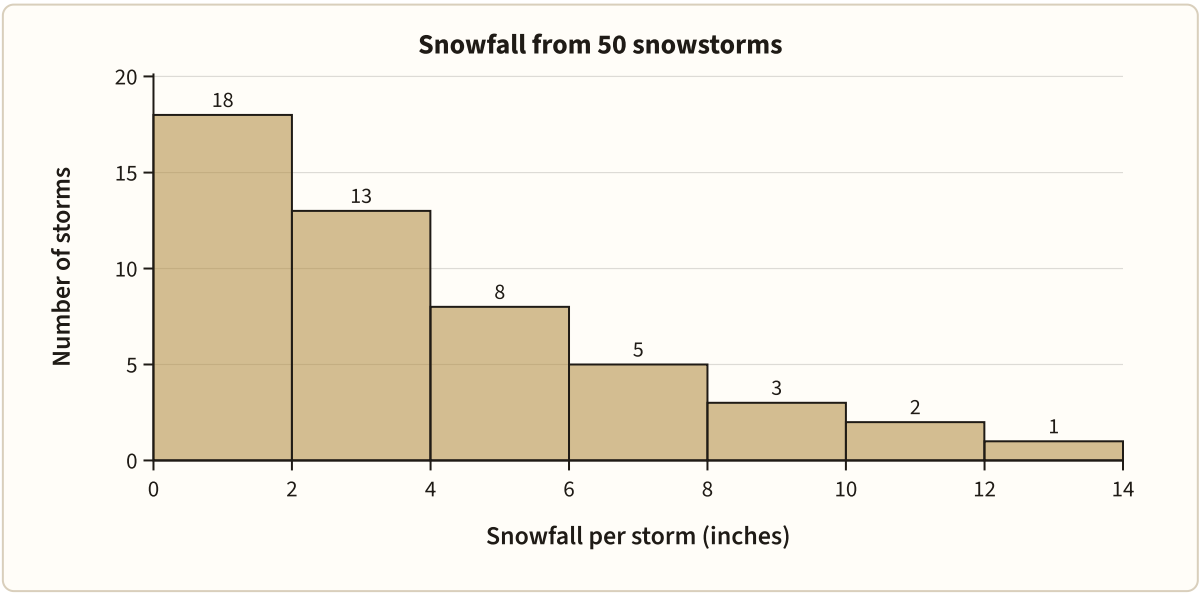
11. **HARD** The back-to-back stemplot shows the weights of nine 60-day-old lambs on each of two pastures. (a) Find the median weight on the North pasture. (b) Find the median weight on the South pasture. (c) Write a short comparison of the two distributions.

Weights of 60-day-old lambs on two pastures

South pasture								North pasture							
							3	8							
					9	5	4	2	4	5	7	8			
9	7	6	4	3	2		5	0	1	5					
					3		6								

Key: 4 | 5 means 45 pounds

12. **CHALLENGE** Using the snowfall histogram from the lesson, (a) what percent of the 50 storms brought at least 8 inches of snow? (b) Explain why the median snowfall is well below the middle of the range of the data.



13. **CHALLENGE** A farmer makes one dotplot of the birth weights of all the animals born on the farm one spring: 12 lambs (weighing about 8 to 12 pounds) and 6 calves (weighing about 70 to 90 pounds). Describe what the dotplot would look like, explain why, and say what a statistician should do instead of describing the whole graph as one distribution.

COMMON MISTAKES

- Naming the skew for the peak instead of the tail. A distribution with its peak on the left and a long tail to the right is skewed RIGHT.
- Describing a graph without context, for example "the center is 23". Always name the variable and the units: "a median of 23 visitors per Saturday".
- Listing two groups' numbers without comparing them. Use comparison words: "Variety B's yields are typically about 11.5 bushels per acre higher than Variety A's."
- Ignoring an outlier or quietly deleting it. Point it out and look for a reason (a recording error or a special case).

ANSWERS TO PRACTICE

1. Roughly symmetric: the counts rise to a peak at 490 to 510 grams and fall off about equally on both sides.
2. Skewed right: most families live within 10 miles, and a few who live farther away form a long tail toward larger distances.
3. Skewed left: most scores are in the 80s and 90s, with a long tail toward the low scores of 52 and 64.

4. With 15 scores, the median is the 8th in order: 87.
5. (a) With 20 values, the median is the average of the 10th and 11th values: $\frac{4+5}{2} = 4.5$ books. (b) $15 - 1 = 14$ books.
6. $\frac{12+8}{30} = \frac{20}{30} \approx 0.667$, so about 66.7%.
7. (a) 7 times. (b) The times form two clusters, one from 52 to 60 minutes and a larger one from 76 to 91 minutes, with a gap from 61 to 75 minutes where no times fell. The distribution has two peaks, which suggests two different kinds of eruption cycles (a short wait and a long wait).
8. He gave no context or units and did not mention outliers. Better: "The number of books borrowed by the 20 club members is skewed right, with a median of 4.5 books; most members borrowed between 1 and 9 books, but a few borrowed many more (11, 14 and 15 books), forming the long right tail. The largest values, 14 and 15 books, stand a little apart from the rest and are worth checking as possible outliers."
9. (a) 31 pounds stands far above the others (which run from 14 to 19 pounds). (b) The 5th of 9 values: 17 pounds. (c) Of the 8 values 14, 15, 15, 16, 17, 17, 18, 19, the median is $\frac{16+17}{2} = 16.5$ pounds. The outlier barely moves the median.
10. The weights of the 40 loaves are roughly symmetric and single-peaked, with no outliers. The center is about 500 grams: the 20th and 21st weights both fall in the 500-to-510-gram class. The weights range from somewhere in the 470s to somewhere in the 520s, so all 40 loaves were within about 30 grams of 500 grams.
11. (a) The 5th of 9 values: 47 pounds. (b) 54 pounds. (c) Both distributions are roughly symmetric with no outliers. The South pasture lambs were typically heavier, by about $54 - 47 = 7$ pounds at the median. The spreads are about the same: North ranges from 38 to 55 pounds (range 17) and South from 45 to 63 pounds (range 18).
12. (a) $\frac{3+2+1}{50} = \frac{6}{50} = 0.12$, so 12%. (b) The distribution is skewed right. The median is the 25th or 26th value in order, and since 18 storms had less than 2 inches and 31 had less than 4 inches, the median lies between 2 and 4 inches. The few big storms stretch the range out to 14 inches without moving the middle storms, so the median sits far below the middle of the range (about 7 inches).
13. The dotplot would show two clusters, 12 dots between about 8 and 12 pounds and 6 dots between about 70 and 90 pounds, with a wide gap between them. The clusters appear because two very different kinds of animals are mixed in one graph. A statistician should separate the groups and describe (or compare) the lamb weights and the calf weights on their own, since a single "center" for the combined data (somewhere in the gap) would describe no animal at all.

Unit 1 test

You have 55 minutes. You may use a scientific calculator; no tables are needed. Show your work for every problem that involves a calculation, and answer questions about data in complete sentences in the context of the problem, with units. Round percents to one decimal place unless the problem says otherwise. The test is worth 100 points.

1. **5 PTS** At a county fair, the judges record three things about each pie entered: the kind of fruit in it, the number of judges who tasted it, and its weight in pounds. (a) Is kind of fruit categorical or quantitative? (b) Is the number of judges who tasted it discrete or continuous? (c) Is the weight discrete or continuous?

2. **6 PTS** A beekeeper has 80 hives. To estimate the mean honey harvest per hive this year, the beekeeper weighs the honey from 12 hives chosen at random and finds a mean of 52 pounds. (a) What is the population? (b) What is the sample size? (c) Is 52 pounds a parameter or a statistic? Explain.

3. **6 PTS** The table shows the 80 birds a family counted at their feeder one morning. (a) What percent of the birds were sparrows? Give the exact percent. (b) In a pie chart of these data, what angle (in degrees) would the cardinal slice have?

   Birds seen at a farmhouse feeder one morning

Species	Number seen
Chickadee	18
Cardinal	12
Sparrow	27
Blue jay	9
Finch	14
Total	80

4. **6 PTS** A bakery sold 90 loaves on Monday and 120 loaves on Saturday. Its newsletter shows these as two bars on a vertical axis that starts at 80 loaves. (a) How many times as tall as Monday's bar does Saturday's bar look? (b) How many times as many loaves were actually sold on Saturday? Round to two decimal places. (c) How should the graph be fixed?

5. **8 PTS** The table shows the activity chosen by each of 140 campers. (a) What percent of all the campers chose archery? (b) Of the younger campers, what percent chose canoeing? (c) Of the older campers, what percent chose canoeing? Round to one decimal place where needed.

Activities chosen at a summer camp

Age group	Canoeing	Archery	Hiking	Total
Younger (10 to 12)	20	35	25	80
Older (13 to 15)	36	15	9	60
Total	56	50	34	140

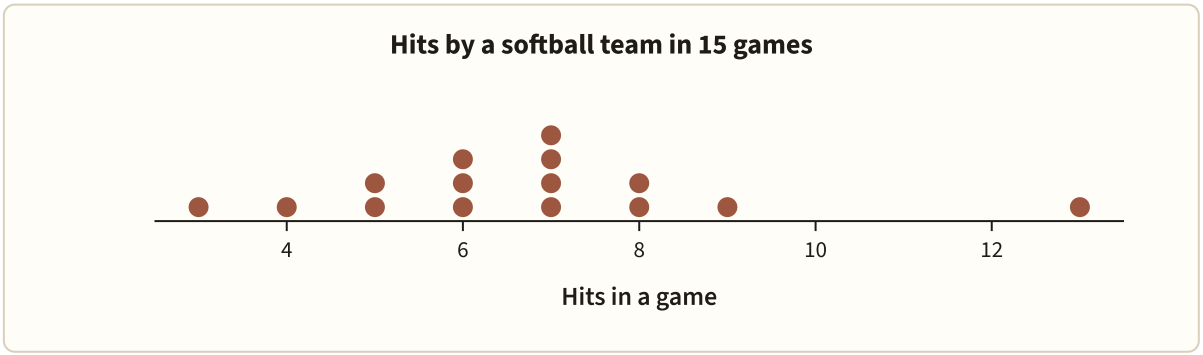
6. **6 PTS** Use the camp table again. (a) Of the campers who chose canoeing, what percent were older campers? Round to one decimal place. (b) Is there an association between age group and activity chosen? Explain using conditional percents.

Activities chosen at a summer camp

Age group	Canoeing	Archery	Hiking	Total
Younger (10 to 12)	20	35	25	80
Older (13 to 15)	36	15	9	60
Total	56	50	34	140

7. **6 PTS** Hatchery A hatched 180 of the 240 eggs it set, and Hatchery B hatched 120 of the 150 eggs it set. (a) What percent of its eggs did Hatchery A hatch? (b) What percent did Hatchery B hatch? (c) Which hatchery did a better job? Explain why comparing the counts 180 and 120 would mislead.

8. **6 PTS** The dotplot shows the number of hits a softball team made in each of 15 games. (a) In how many games did the team make 7 or more hits? (b) Which value appears to be an outlier?



9. **7 PTS** The stemplot shows the heights of 11 tomato seedlings. (a) Find the median height. (b) How many seedlings were at least 5.0 inches tall?

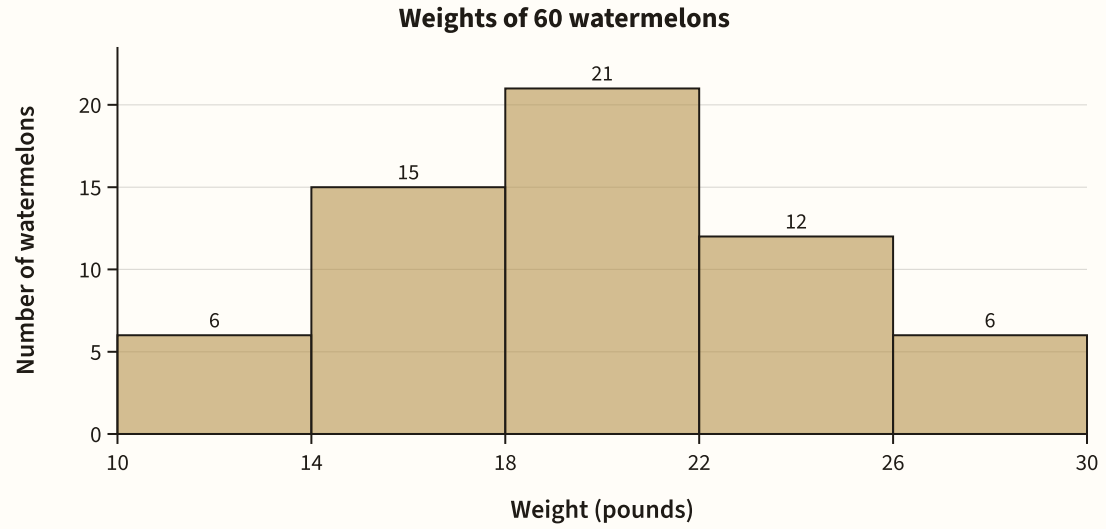
Heights of 11 tomato seedlings

3	2 8
4	1 4 5 7
5	0 2 3 9
6	6

Key: 3 | 2 means 3.2 inches

10. **7 PTS** The wind speeds, in miles per hour, at a farm at noon on 18 days were 3, 5, 6, 7, 8, 8, 9, 10, 11, 11, 12, 13, 14, 16, 18, 21, 24, 29. Give the counts for a histogram with the classes 0 to under 5, 5 to under 10, 10 to under 15, 15 to under 20, 20 to under 25, and 25 to under 30.

11. **6 PTS** The histogram shows the weights of 60 watermelons. (a) What percent of the watermelons weighed at least 22 pounds? (b) Is the distribution roughly symmetric, skewed left or skewed right?



12. **6 PTS** Samuel graphed the watermelon weights from Problem 11 by drawing the five bars with spaces between them, arranged from tallest to shortest. Explain what is wrong with his graph and why.

13. **8 PTS** The stemplot shows how many minutes it took each of 20 volunteers to sort one box of donations at a food pantry. (a) Find the median time. (b) Describe the distribution (shape, outliers, center, spread) in context.

Minutes for 20 volunteers to sort one box at a food pantry

1	2 3 3 4 4 5 5 5 6 6 7 8 8 9
2	1 2 4 6 9
3	
4	1

Key: 1 | 2 means 12 minutes

14. **9 PTS** The back-to-back stemplot shows the December snowfall in two towns over the same 10 years. (a) Find the median December snowfall in Ridgeton. (b) Find the median December snowfall in Valley Falls. (c) Write a short paragraph comparing the two distributions.

December snowfall in two towns over 10 years

Valley Falls		Ridgeton
9 8 7 6 6 5 4 2	0	8
4 1	1	2 4 5 7 8
	2	0 1 3 9

Key: 1 | 2 means 12 inches

15. **8 PTS** At a county fair, 150 jars of jam were entered: 90 were strawberry and the rest were peach. Thirty of the strawberry jars and 24 of the peach jars won ribbons. (a) What percent of the strawberry jars won ribbons? (b) What percent of the peach jars won ribbons? (c) Of the jars that won ribbons, what percent were peach? Round to one decimal place where needed.

UNIT 1 TEST: ANSWER KEY

1. (a) Categorical. (b) Discrete (a count). (c) Continuous (a measurement).
2. (a) All 80 of the beekeeper's hives (their honey harvests this year). (b) 12 hives. (c) A statistic: it is computed from the sample of 12 hives, and it estimates the parameter, the mean harvest of all 80 hives.
3. (a) $\frac{27}{80} = 0.3375$, so 33.75%. (b) $\frac{12}{80} = 0.15$ and $0.15 \times 360^\circ = 54^\circ$.
4. (a) The bars are $90 - 80 = 10$ and $120 - 80 = 40$ units tall, so Saturday's looks $\frac{40}{10} = 4$ times as tall. (b) $\frac{120}{90} \approx 1.33$ times as many. (c) Start the vertical axis at 0 so the bar heights are in the true proportion.
5. (a) $\frac{50}{140} \approx 35.7\%$. (b) $\frac{20}{80} = 25\%$. (c) $\frac{36}{60} = 60\%$.
6. (a) $\frac{36}{56} \approx 64.3\%$. (b) Yes. Older campers chose canoeing much more often (60% versus 25%) and archery much less often ($\frac{15}{60} = 25\%$ versus $\frac{35}{80} \approx 43.8\%$) than younger campers, so knowing a camper's age group tells you something about the activity chosen.
7. (a) $\frac{180}{240} = 75\%$. (b) $\frac{120}{150} = 80\%$. (c) Hatchery B hatched a higher percent of its eggs (80% versus 75%). Hatchery A hatched more chicks only because it set more eggs; percents remove the difference in size and compare the success rates fairly.
8. (a) $4 + 2 + 1 + 1 = 8$ games. (b) 13 hits, which stands well apart from the other games (3 to 9 hits).
9. (a) The 6th of 11 values in order: 4.7 inches. (b) 5.0, 5.2, 5.3, 5.9 and 6.6, so 5 seedlings.
10. 0–5: 3 (1). 5–10: 5, 6, 7, 8, 8, 9 (6). 10–15: 10, 11, 11, 12, 13, 14 (6). 15–20: 16, 18 (2). 20–25: 21, 24 (2). 25–30: 29 (1). Check: $1 + 6 + 6 + 2 + 2 + 1 = 18$.
11. (a) $\frac{12 + 6}{60} = \frac{18}{60} = 0.30$, so 30%. (b) Roughly symmetric: the counts rise to a peak at 18 to 22 pounds and fall off about equally on both sides.
12. Weight is a quantitative variable, so the classes must stay in number-line order (10–14, 14–18, and so on); rearranging them by height destroys the shape of the distribution. The bars should also touch, because the classes cover the number line with no gaps. Samuel has drawn a bar graph, which is for categorical data, instead of a histogram.
13. (a) The average of the 10th and 11th times: $\frac{16 + 17}{2} = 16.5$ minutes. (b) The sorting times are skewed right, with most volunteers taking 12 to 19 minutes and a tail toward longer times. One time, 41 minutes, is an outlier, far above the rest. The median time is 16.5 minutes, and apart from the outlier the times range from 12 to 29 minutes.

14. (a) $\frac{17+18}{2} = 17.5$ inches. (b) $\frac{6+7}{2} = 6.5$ inches. (c) Ridgeton typically gets much more December snow than Valley Falls, about $17.5 - 6.5 = 11$ inches more at the median. Ridgeton's snowfall is also more variable, ranging from 8 to 29 inches (range 21) compared with 2 to 14 inches (range 12) in Valley Falls. Ridgeton's distribution is roughly symmetric and Valley Falls' is slightly skewed right; neither has a clear outlier.
15. Two-way table: strawberry 30 ribbons, 60 no ribbon (90); peach 24 ribbons, 36 no ribbon (60); totals 54 ribbons, 96 no ribbon (150). (a) $\frac{30}{90} \approx 33.3\%$. (b) $\frac{24}{60} = 40\%$. (c) $\frac{24}{54} \approx 44.4\%$.

UNIT 2

Describing Data with Numbers

In Unit 1 you learned to draw a distribution and describe it in words. In this unit you learn to describe it with numbers: the mean and median for center; the range, quartiles, IQR and standard deviation for spread; boxplots and the 1.5 × IQR rule for outliers. Just as important, you learn why each measure works the way it does: why the median resists outliers and the mean does not, why we square deviations and divide by n − 1, and why changing units shifts or stretches a distribution without changing its shape.

These ideas are the foundation for the rest of the course. The mean and standard deviation lead directly to z-scores and normal distributions in Unit 3, to correlation and regression in Unit 4, and to every method of inference later in the year. The habit of choosing the right summary for the shape of the data, and of comparing groups clearly in context, is one you will use in every unit and in your statistics project.

PREREQUISITES

- Describe the shape, center, spread and outliers of a distribution in words (Lesson 1.5).
- Read and draw dotplots, stemplots (including back-to-back) and histograms (Lesson 1.4).
- Compute with fractions, decimals, percents, squares and square roots.
- Use order of operations and evaluate a formula such as $C = \frac{5}{9}(F - 32)$.
- Distinguish a sample from a population (Lesson 1.1).

MASTERY CHECKLIST

☐ I can compute the mean and median and explain the mean as a fair share and a balance point.
☐ I can explain why the median is resistant and the mean is not, and predict which is larger in a skewed distribution.
☐ I can compute a weighted mean, such as an average price per bushel.
☐ I can find the quartiles by the course method, write the five-number summary and interpret the IQR in context.
☐ I can use the 1.5 × IQR rule to identify outliers and draw a modified boxplot.
☐ I can compare groups with parallel boxplots and say what a boxplot hides.
☐ I can compute a standard deviation by hand with a table and with a calculator, and interpret it as a typical distance from the mean.
☐ I can explain why we square deviations and why a sample standard deviation divides by n − 1.

☐ I can choose the median and IQR or the mean and standard deviation depending on the shape of the data.
☐ I can write a comparison paragraph in context that covers shape, center, spread and outliers.
☐ I can find the new center and spread after adding a constant, multiplying by a constant, or changing units.

LESSON 2.1 · 2 SESSIONS OF 45 MINUTES

Measures of Center

OBJECTIVES

- Compute the mean and the median of a small data set and interpret each in context.
- Explain the mean as a fair share and as a balance point, and use $\sum$ to mean "add them up".
- Explain why the median is resistant to outliers and the mean is not, and predict which is larger in a skewed distribution.
- Compute a weighted mean, such as a grade average or an average price per bushel.

VIDEO

WATCH

Measures of Central Tendency (1 of 2: Comparing Mean & Median)

Eddie Woo · 15:26

When to watch: Before Example 1: mean versus median, and why an outlier pulls the mean.

https://www.youtube.com/watch?v=NLNf5W6g6jc

BACKUP

Finding Mean, Median, and Mode | Math with Mr. J

Math with Mr. J · 6:31

When to watch: Review: finding the mean, median and mode.

https://www.youtube.com/watch?v=oatwXlZBPw0

KEY TERMS

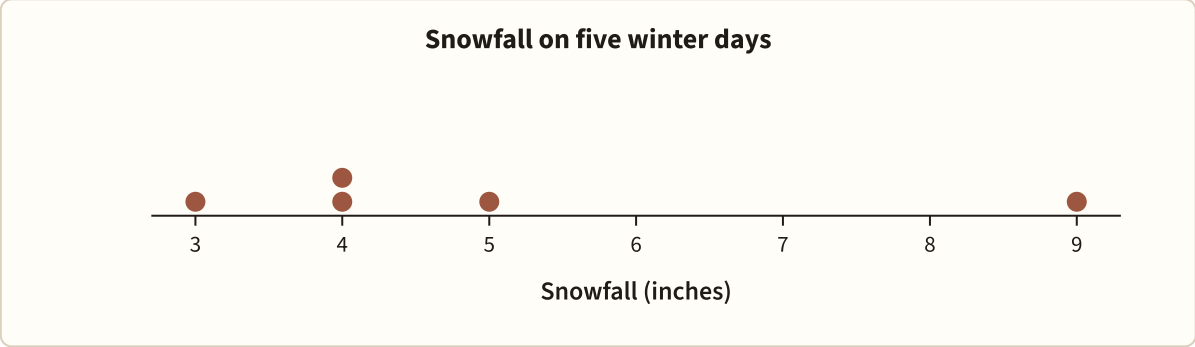
Measure of center	A single number that describes a typical or middle value of a quantitative data set, such as the mean or the median.
Mean	The sum of the values divided by the number of values, written $\bar{x}$ ("x-bar"). It is the fair-share value and the balance point of the data.
Median	The middle value of the data when they are put in order; with an even number of values, the average of the two middle values. Half the data lie at or below it and half at or above it.
$\sum$ (sigma)	The Greek capital letter S, used to mean "add them up". $\sum x$ is the sum of all the values of x.
Resistant	A statistic is resistant if a few extreme values (outliers) cannot change it very much. The median is resistant; the mean is not.

Weighted mean A mean in which each value counts according to its weight (how many items or how much importance it carries): multiply each value by its weight, add, and divide by the total weight.

CORE IDEAS

In Unit 1 you described the center of a distribution by looking at a graph. Now we measure it with a number. A **measure of center** answers the question "what is a typical value?" The two most important are the mean and the median, and they answer that question in different ways.

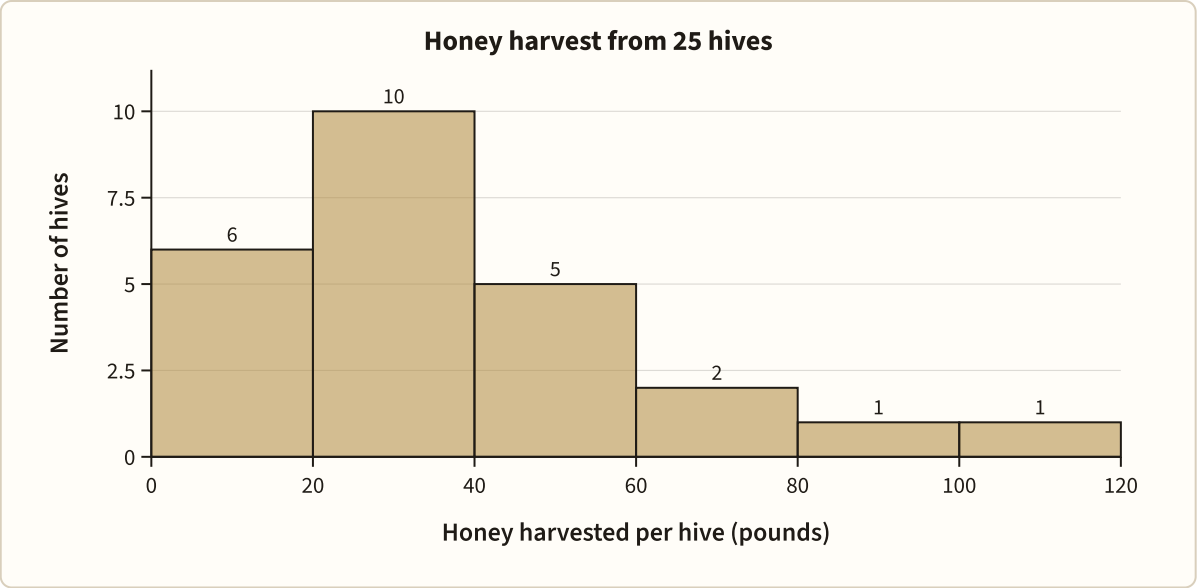
The **mean** is what most people call the average: add up all the values and divide by how many there are. Statisticians write the sum with the Greek letter $\sum$ (sigma), which simply means "add them up", so the mean is $\bar{x} = \dfrac{\sum x}{n}$, where n is the number of values. One way to think of the mean is as a **fair share**. Suppose the snowfall on five winter days was 3, 4, 4, 5 and 9 inches. The total is 25 inches; if that snow had been shared out evenly, each day would have received $25 \div 5 = 5$ inches. So $\bar{x} = 5$ inches.



The mean is also the **balance point** of the data. Picture the dotplot above as a seesaw with equal weights at 3, 4, 4, 5 and 9. The values below 5 are $2 + 1 + 1 = 4$ inches short of it in total, and the one value above 5 is 4 inches beyond it. The distances on the two sides cancel exactly, so the plank balances at 5. This is why the mean is pulled toward a far-off value: a single value a long way out has a long lever arm, and the balance point must slide toward it.

The **median** is the middle of the ordered data. Put the values in order. If n is odd, the median is the single middle value, in position $\dfrac{n+1}{2}$. If n is even, there are two middle values, and the median is their average. For the snowfall data (3, 4, 4, 5, 9) the median is the third value, 4 inches. Half the days had 4 inches or less, and half had 4 inches or more.

Why do the two measures react so differently to outliers? A basketball player scored 12, 14, 15, 16 and 18 points in five games: the mean is $75 \div 5 = 15$ and the median is 15. Now suppose the 18 had been 48 (a wonderful night). The sum becomes 105 and the mean jumps to $105 \div 5 = 21$, higher than four of the five games. The median is still 15, because the median depends only on which value sits in the middle, not on how far away the biggest value is. We say the median is **resistant** and the mean is not. The mean uses the size of every value, which is its strength (it uses all the information) and its weakness (one wild value can drag it).



This explains how the mean and median compare in a skewed distribution. The honey harvests above are skewed right: most hives gave under 40 pounds, but a few gave much more. Those few large values pull the mean toward the long right tail, while the median stays with the bulk of the hives. So in a distribution skewed right the mean is usually larger than the median; skewed left, the mean is usually smaller; roughly symmetric, the two are close. When a distribution is strongly skewed or has outliers, the median usually gives the better picture of a typical value.

Sometimes values should not count equally. A farmer who sells 40 bushels of apples at \$5 a bushel and 60 bushels at \$6 did not get an average price of \$5.50, because more of the apples brought the higher price. Each price must be weighted by the number of bushels: $\frac{40 \times 5 + 60 \times 6}{40 + 60} = \frac{560}{100} = 5.60$, so the average price was \$5.60 per bushel. This is a **weighted mean**: multiply each value by its weight, add, and divide by the total weight. Your grade average, when tests count more than homework, is a weighted mean too.

Always report a center with its units and in context: "The mean snowfall was 5 inches per day" tells a reader something; "the mean is 5" does not.

KEY FORMULAS

$$\bar{x} = \frac{\sum x}{n}$$

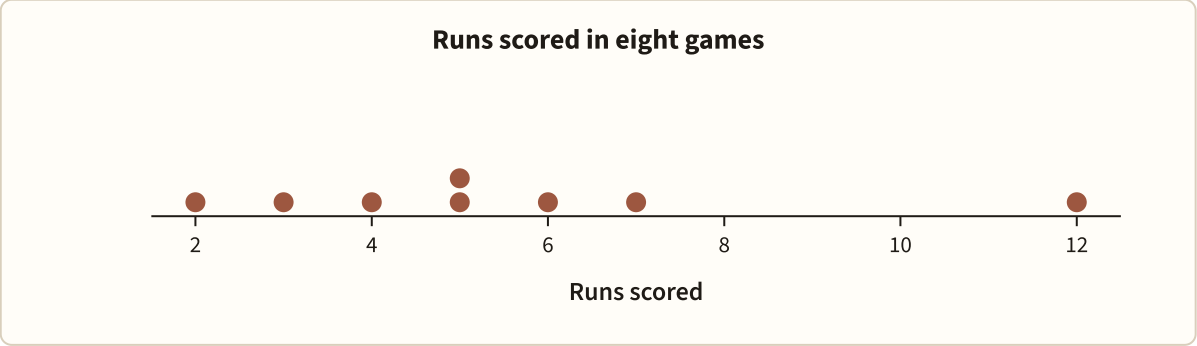
Median: the value in position $\frac{n+1}{2}$ of the ordered data (average the two middle values when n is eve

$$\text{Weighted mean} = \frac{\sum(\text{weight} \times \text{value})}{\sum \text{weight}}$$

Practice habit. Sort the data and glance at a quick dotplot before you choose a measure of center. If you see strong skewness or an outlier, report the median (or both, and say why they differ).

WORKED EXAMPLES

Example 1. A church-league baseball team scored these numbers of runs in its eight games: 3, 5, 2, 7, 4, 12, 5, 6. (a) Find the mean number of runs. (b) Find the median. (c) Which better describes a typical game, and why?



1. Add the values: $\sum x = 3 + 5 + 2 + 7 + 4 + 12 + 5 + 6 = 44$ runs.
2. Divide by $n = 8$: $\bar{x} = \dfrac{44}{8} = 5.5$ runs.
3. Sort the data: 2, 3, 4, 5, 5, 6, 7, 12.
4. Since n = 8 is even, the median is the average of the 4th and 5th values: $\dfrac{5 + 5}{2} = 5$ runs.
5. The dotplot shows one game (12 runs) far to the right of the others. It pulls the mean up; the median ignores how far out it is.
6. Both are reasonable here because the outlier is only one game, but the median of 5 runs better describes a typical game; the mean of 5.5 runs is nudged up by the 12-run game.

Answer: (a) $\bar{x} = 5.5$ runs. (b) Median = 5 runs. (c) The median, 5 runs, because the 12-run game pulls the mean up while the median resists it.

Example 2. Henry sold apples at three places last fall, as shown in the table. Find his average price per bushel for all the apples he sold, and explain why it is not simply the average of the three prices.

Henry's apple sales

Where sold	Bushels sold	Price per bushel (dollars)
Farm stand	30	18
Town market	50	20
Grocery store	20	25
Total	100	

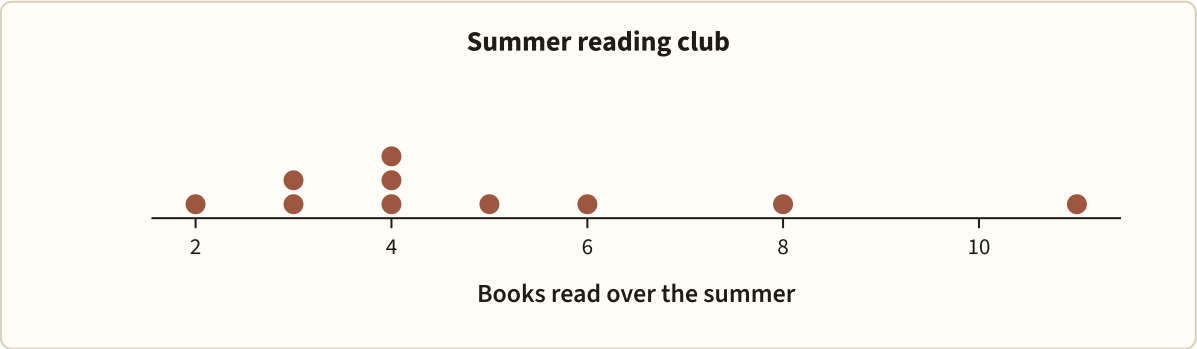
1. Each price must be weighted by the number of bushels sold at that price.
2. Money taken in at each place: $30 \times 18 = 540$, $50 \times 20 = 1000$, $20 \times 25 = 500$ dollars.
3. Total money: $540 + 1000 + 500 = 2040$ dollars.
4. Total bushels: $30 + 50 + 20 = 100$.
5. Weighted mean price: $\dfrac{2040}{100} = 20.40$ dollars per bushel.
6. The plain average of the three prices is $\dfrac{18 + 20 + 25}{3} = 21$ dollars, which is too high: it counts the \$25 price as heavily as the others, but only 20 of the 100 bushels sold at that price.
7. Conclusion: Henry received an average of \$20.40 for each bushel of apples he sold.

Answer: \$20.40 per bushel, since $\dfrac{30(18) + 50(20) + 20(25)}{100} = \dfrac{2040}{100}$. The plain average of the prices (\$21) gives the small grocery-store sale too much weight.

PRACTICE

1. EASY Grace collected these numbers of eggs from her hens on five days: 6, 8, 7, 9, 10. Find the mean number of eggs per day.
2. EASY The rainfall in inches at a farm for seven months was 2.1, 3.4, 1.8, 4.0, 2.9, 3.1, 2.5. Find the median monthly rainfall.
3. EASY Samuel hit six golf drives of 215, 240, 228, 232, 250 and 221 yards. Find the median drive.
4. EASY A baker weighs the 12 loaves from one batch and finds $\sum x = 252$ ounces. What is the mean weight of a loaf?

5. MEDIUM The dotplot shows how many books each of the 10 members of a summer reading club read. Find (a) the mean and (b) the median number of books.

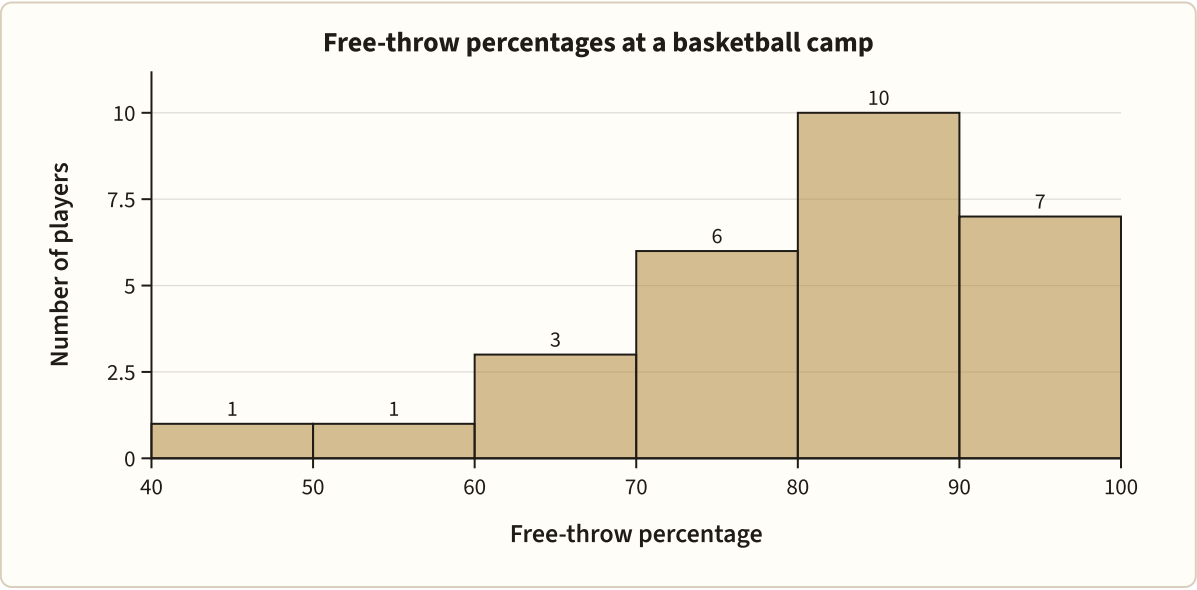


6. MEDIUM Five pumpkins from Ruth's garden weigh 12, 14, 15, 17 and 22 pounds. A neighbor adds a prize pumpkin weighing 94 pounds to the group. Find (a) the new mean and (b) the new median of the six pumpkins.

7. MEDIUM Samuel's first four quiz scores are 85, 90, 82 and 91. What score does he need on the fifth quiz for his mean on all five to be exactly 88?

8. MEDIUM In Grace's course, homework counts 25%, quizzes 25% and tests 50% of the grade. Her averages are homework 92, quizzes 85 and tests 80. Find her course grade.

9. MEDIUM The histogram shows the free-throw percentages of 28 players at a basketball camp. (a) Describe the shape of the distribution. (b) Without calculating, would you expect the mean or the median to be larger? Explain.



10. HARD Two church youth groups did a service project. The 12 members of the first group worked a mean of 8 hours each, and the 18 members of the second group worked a mean of 13 hours each. Find the mean number of hours for all 30 young people.

11. **HARD** Mr. Miller drives the 60 miles to the county fair at 30 miles per hour and the 60 miles home at 60 miles per hour. What is his average speed for the whole trip? (Hint: average speed is total distance divided by total time.)

12. **HARD** For the 200 farms in a county, the median size is 180 acres and the mean size is 420 acres. (a) What does this suggest about the shape of the distribution of farm sizes? (b) Which measure better describes the size of a typical farm in the county? Explain.

13. **HARD** A carpenter measured 10 fence posts and found a mean length of 7.2 feet. He then noticed that one post recorded as 9.1 feet is really 6.1 feet. What is the correct mean length?

14. **CHALLENGE** Seven pumpkins at a harvest festival have a mean weight of 12 pounds. When one pumpkin is sold, the remaining six have a mean weight of 11 pounds. How much did the sold pumpkin weigh?

COMMON MISTAKES

- Finding the median without sorting the data first. Always put the values in order before you look for the middle.
- Taking the plain average of prices or rates when the amounts differ. If 50 bushels sold at one price and 20 at another, weight each price by its number of bushels (a weighted mean).
- Thinking the mean is resistant. One outlier can move the mean a long way; check the data for outliers and report the median when the distribution is strongly skewed.
- Reporting a center with no units or context. Write "the median drive was 230 yards", not "the median is 230".

ANSWERS TO PRACTICE

1. $\bar{x} = \dfrac{6+8+7+9+10}{5} = \dfrac{40}{5} = 8$ eggs per day.
2. Sorted: 1.8, 2.1, 2.5, 2.9, 3.1, 3.4, 4.0. The 4th value is the median: 2.9 inches.
3. Sorted: 215, 221, 228, 232, 240, 250. Median $= \dfrac{228+232}{2} = 230$ yards.
4. $\bar{x} = \dfrac{252}{12} = 21$ ounces.
5. (a) Values 2, 3, 3, 4, 4, 4, 5, 6, 8, 11 add to 50, so $\bar{x} = \dfrac{50}{10} = 5$ books. (b) The 5th and 6th values are 4 and 4, so the median is 4 books.
6. (a) $\dfrac{12+14+15+17+22+94}{6} = \dfrac{174}{6} = 29$ pounds (the mean was 16 pounds before). (b) Sorted: 12, 14, 15, 17, 22, 94; median $= \dfrac{15+17}{2} = 16$ pounds (it was 15 before). The mean jumped by 13 pounds; the median moved only 1 pound.
7. The five scores must total $5 \times 88 = 440$. He has $85 + 90 + 82 + 91 = 348$, so he needs $440 - 348 = 92$.
8. $0.25(92) + 0.25(85) + 0.50(80) = 23 + 21.25 + 40 = 84.25$.

9. (a) Skewed left: most players shoot between 70% and 100%, with a tail of a few weaker shooters toward 40%. (b) The median should be larger. The few low percentages in the left tail pull the mean down toward them, while the median stays with the bulk of the players.
10. Total hours: $12 \times 8 + 18 \times 13 = 96 + 234 = 330$. Mean $= \dfrac{330}{30} = 11$ hours. (Not 10.5: the larger group counts more.)
11. Time there $= 60 \div 30 = 2$ hours; time home $= 60 \div 60 = 1$ hour. Average speed $= \dfrac{120}{3} = 40$ miles per hour, not 45, because he spent twice as long driving at the slower speed.
12. (a) Skewed right: a few very large farms or ranches pull the mean far above the median. (b) The median, 180 acres. Half the farms are 180 acres or smaller; the mean of 420 acres is inflated by the few huge farms and is larger than most farms in the county.
13. Recorded total $= 10 \times 7.2 = 72$ feet. Correct total $= 72 - 9.1 + 6.1 = 69$ feet. Correct mean $= \dfrac{69}{10} = 6.9$ feet.
14. Total of seven $= 7 \times 12 = 84$ pounds; total of six $= 6 \times 11 = 66$ pounds. The sold pumpkin weighed $84 - 66 = 18$ pounds.

LESSON 2.2 · 2 SESSIONS OF 45 MINUTES

Quartiles, the Five-Number Summary and IQR

OBJECTIVES

- Compute the range and explain why it is not resistant.
- Find the quartiles of a data set by the course method, for both odd and even numbers of values.
- Write the five-number summary and compute the interquartile range (IQR).
- Interpret the quartiles and the IQR in context as describing the middle half of the data.

VIDEO

WATCH

Box Plots (1 of 2: Five-Number Summary)

Eddie Woo · 8:41

When to watch: Before Example 1: the five-number summary.

https://www.youtube.com/watch?v=drKEoMcYW3Q

BACKUP

Interquartile Range (IQR) | Math with Mr. J

Math with Mr. J · 7:33

When to watch: Extra practice: finding the IQR step by step, for an odd and an even number of values.

https://www.youtube.com/watch?v=VABsJBw1JqA

> **Note:** Both videos leave the median out of each half when n is odd, as the course does.

KEY TERMS

Range	The maximum minus the minimum. It measures the spread of the whole data set, from the very smallest value to the very largest.
First quartile (Q1)	The median of the lower half of the ordered data. About a quarter of the values lie below it.
Third quartile (Q3)	The median of the upper half of the ordered data. About three quarters of the values lie below it.
Five-number summary	The minimum, Q1, median, Q3 and maximum, written in that order. It splits the data into four parts with about a quarter of the values in each.
Interquartile range (IQR)	IQR = Q3 − Q1, the spread of the middle half of the data.

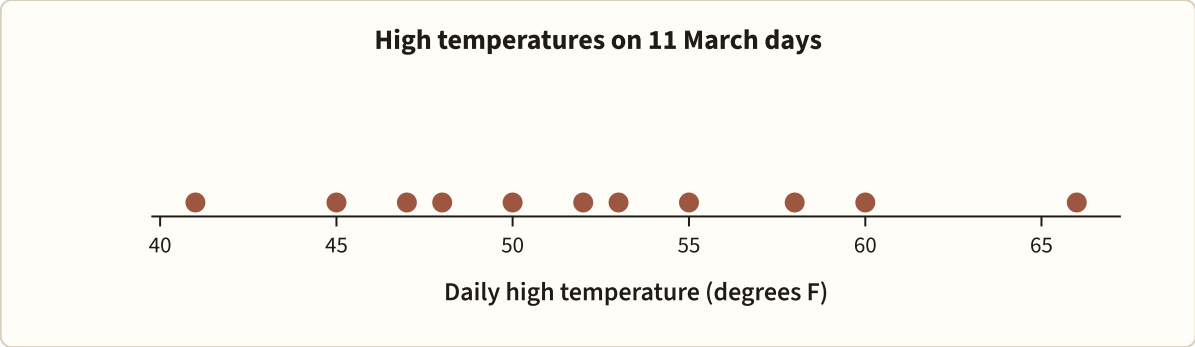
CORE IDEAS

Two farms can have the same mean yield and still be very different: one gets nearly the same harvest every year, and the other swings from feast to famine. A center alone is not enough; we also need a measure of **spread**, which tells how much the values vary.

The simplest measure of spread is the **range** = maximum − minimum. It is easy to find, but it depends on only the two most extreme values. If one value is unusual (a record cold night, a mis-typed number), the range changes completely. Like the mean, the range is **not resistant**, and it tells us nothing about how the values in between are spread out.

A better idea is to describe the spread of the **middle half** of the data, which ignores the extremes. To do that we cut the ordered data into quarters. The median splits the data into a lower half and an upper half. The **first quartile Q1** is the median of the lower half, and the **third quartile Q3** is the median of the upper half. In this course, when n is odd, the overall median is left out of both halves.

Here are the daily high temperatures (°F) on 11 days in March, in order: 41, 45, 47, 48, 50, 52, 53, 55, 58, 60, 66. With n = 11, the median is the 6th value, 52°F. Leave it out; the lower half is 41, 45, 47, 48, 50, with median Q1 = 47°F, and the upper half is 53, 55, 58, 60, 66, with median Q3 = 58°F.



The **five-number summary** lists minimum, Q1, median, Q3, maximum: here 41, 47, 52, 58, 66. These five numbers cut the data into four parts, each holding about a quarter of the values. The **interquartile range** is IQR = Q3 − Q1 = 58 − 47 = 11°F: the middle half of the March days had highs within an 11-degree span, from 47°F to 58°F. Compare the range, 66 − 41 = 25°F.

Why is the IQR resistant? Because Q1 and Q3 are medians, and medians care only about which value is in the middle position, not how far away the extreme values are. If the 66°F day had been recorded as 86°F, the range would jump from 25 to 45 degrees, but the upper half would still have 58 in the middle, so Q3 and the IQR would not change at all. The extreme quarter at each end can do whatever it likes; the IQR does not see it.

When n is even, nothing is left out: the lower half is the first n/2 values and the upper half is the last n/2 values. With 10 values, the halves have 5 values each, and Q1 and Q3 are the 3rd and 8th values. Calculators and computer programs sometimes use slightly different quartile rules, so their answers can differ a little from the course method; on tests, use the course method.

KEY FORMULAS

$$\text{Range} = \text{maximum} - \text{minimum}$$

$$\text{Five-number summary: Min, } Q_1, \text{ Median, } Q_3, \text{ Max}$$

$$\text{IQR} = Q_3 - Q_1$$

Practice habit. Write the data in order, mark the median, and when n is odd cross the median out before you split the data into halves. Then find the median of each half.

WORKED EXAMPLES

Example 1. The birth weights (pounds) of 10 calves on the Miller farm were 75, 62, 81, 70, 94, 68, 77, 85, 72, 80. Find the five-number summary and the IQR, and interpret the IQR.

1. Sort: 62, 68, 70, 72, 75, 77, 80, 81, 85, 94.
2. n = 10 is even, so the median is the average of the 5th and 6th values: $\frac{75 + 77}{2} = 76$ pounds.
3. Lower half (first 5 values): 62, 68, 70, 72, 75. Its median is Q1 = 70 pounds.
4. Upper half (last 5 values): 77, 80, 81, 85, 94. Its median is Q3 = 81 pounds.
5. Minimum = 62 and maximum = 94 pounds.
6. IQR $= Q_3 - Q_1 = 81 - 70 = 11$ pounds.
7. Interpretation: the middle half of the calves weighed between 70 and 81 pounds at birth, a spread of 11 pounds.

Answer: Five-number summary: 62, 70, 76, 81, 94 pounds. IQR = 11 pounds: the middle half of the calves' birth weights lie within an 11-pound span, from 70 to 81 pounds.

Example 2. The stemplot shows the corn yields (bushels per acre) on 15 test plots. (a) Find the five-number summary. (b) Find the range and the IQR. (c) Interpret the IQR in context.

Corn yields on 15 plots

14	2 8
15	1 3 5 8
16	0 1 3 5 7
17	0 2 8
18	
19	1

Key: 14 | 2 means 142 bushels per acre

1. Read the ordered values: 142, 148, 151, 153, 155, 158, 160, 161, 163, 165, 167, 170, 172, 178, 191 (n = 15).
2. n is odd, so the median is the 8th value: 161 bushels per acre.
3. Leave out the median. Lower half: 142, 148, 151, 153, 155, 158, 160. Its middle (4th) value is Q1 = 153.
4. Upper half: 163, 165, 167, 170, 172, 178, 191. Its middle (4th) value is Q3 = 170.
5. Five-number summary: 142, 153, 161, 170, 191.
6. Range $= 191 - 142 = 49$ bushels per acre; IQR $= 170 - 153 = 17$ bushels per acre.
7. The range is stretched by the single high plot (191); the IQR describes the middle half and is not affected by it.

Answer: (a) 142, 153, 161, 170, 191 bushels per acre. (b) Range = 49, IQR = 17 bushels per acre. (c) The middle half of the plots had yields within a 17-bushel span, between 153 and 170 bushels per acre.

PRACTICE

1. EASY The wind speeds (miles per hour) at noon on six days were 8, 12, 5, 15, 9, 11. Find the range.
2. EASY Henry counted meteors on seven clear nights: 3, 5, 6, 8, 9, 12, 14. Find Q1, the median and Q3.
3. EASY A sawmill worker measured eight boards (inches): 96, 94, 98, 95, 97, 99, 95, 97. Find the five-number summary.
4. EASY For the eggs from a flock of hens, Q1 = 54 grams and Q3 = 62 grams. Find the IQR and say what it tells you.

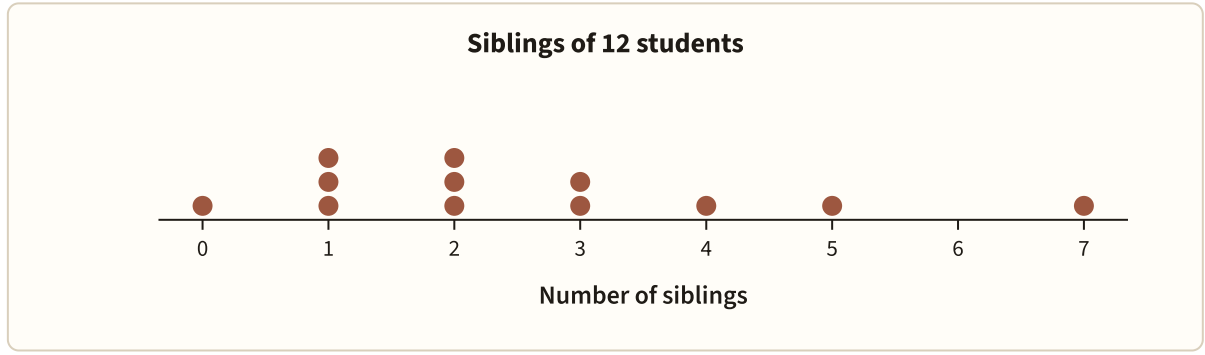
5. **MEDIUM** The stemplot shows the ages of the 14 members of a church choir. Find the five-number summary and the IQR.

Ages of 14 choir members

1	6 8
2	2 5
3	1 4 8
4	1 5 7
5	2 8
6	3
7	0

Key: 3 | 1 means 31 years

6. **MEDIUM** The dotplot shows the number of siblings of 12 students. Find the five-number summary.



7. **MEDIUM** Nine lambs weighed 38, 41, 43, 44, 46, 48, 50, 51, 55 pounds. (a) Find the range and the IQR. (b) Suppose the 55 had been mis-recorded as 85. Find the range and IQR now. (c) What does this show?

8. **MEDIUM** The table gives five-number summaries of apple weights from two orchards. (a) Find the IQR for each orchard. (b) Which orchard produces apples of more consistent size? Explain.

Five-number summaries of apple weights (grams)

Orchard	Min	Q1	Median	Q3	Max
Hillside	140	165	178	190	215
Valley	150	172	180	186	200

9. **MEDIUM** For the honey harvests of a beekeeping club, Q1 = 22 pounds and Q3 = 41 pounds per hive. Ruth says, "Q1 = 22 means 22% of the hives made less honey than that." Explain her mistake and interpret Q1 and Q3 correctly.

10. **HARD** The monthly average sunspot numbers for 13 months were 5, 8, 12, 12, 15, 20, 24, 27, 31, 33, 40, 52, 66. Find the five-number summary and the IQR.

11. HARD A lumberyard records the lengths of 40 boards, all different. About how many of the boards have lengths between Q1 and Q3?

12. HARD A weather reporter wants to describe how much the January low temperatures in a town vary. One January included a record cold night of −31°F, far colder than any other night. Explain why the IQR is a better measure of the typical spread than the range for that January.

13. CHALLENGE The five-number summary of eight 400-meter race times is 50, 52, 55, 59, 63 seconds. Seven of the times are 50, 51, 53, 54, 56, 58, 63 seconds. Find the eighth time.

COMMON MISTAKES

- Including the median in both halves when n is odd. In this course, leave the overall median out of both halves before finding Q1 and Q3.
- Forgetting to sort the data before finding quartiles. Always write the values in order first.
- Treating Q1 as a percent. Q1 is a data value with units (pounds, inches); about 25% of the data lie below it.
- Reporting the IQR as an interval. The IQR is one number, Q3 − Q1; the interval from Q1 to Q3 is where the middle half of the data lie.

ANSWERS TO PRACTICE

1. Range $= 15 - 5 = 10$ miles per hour.
2. Median = 8 (4th value). Lower half 3, 5, 6 gives Q1 = 5; upper half 9, 12, 14 gives Q3 = 12.
3. Sorted: 94, 95, 95, 96, 97, 97, 98, 99. Median $= \dfrac{96 + 97}{2} = 96.5$. Lower half 94, 95, 95, 96 gives Q1 $= \dfrac{95 + 95}{2} = 95$. Upper half 97, 97, 98, 99 gives Q3 $= \dfrac{97 + 98}{2} = 97.5$. Summary: 94, 95, 96.5, 97.5, 99 inches.
4. IQR $= 62 - 54 = 8$ grams: the middle half of the eggs weigh within an 8-gram span, between 54 and 62 grams.
5. n = 14: median $= \dfrac{38 + 41}{2} = 39.5$ years. Lower half 16, 18, 22, 25, 31, 34, 38 gives Q1 = 25; upper half 41, 45, 47, 52, 58, 63, 70 gives Q3 = 52. Summary: 16, 25, 39.5, 52, 70 years; IQR $= 52 - 25 = 27$ years.
6. Ordered: 0, 1, 1, 1, 2, 2, 2, 3, 3, 4, 5, 7. Median $= \dfrac{2 + 2}{2} = 2$. Lower half 0, 1, 1, 1, 2, 2 gives Q1 $= \dfrac{1 + 1}{2} = 1$. Upper half 2, 3, 3, 4, 5, 7 gives Q3 $= \dfrac{3 + 4}{2} = 3.5$. Summary: 0, 1, 2, 3.5, 7.
7. (a) Range $= 55 - 38 = 17$ pounds. Median 46; Q1 $= \dfrac{41 + 43}{2} = 42$, Q3 $= \dfrac{50 + 51}{2} = 50.5$, IQR = 8.5 pounds. (b) Range $= 85 - 38 = 47$ pounds; IQR still 8.5 pounds. (c) The range is not resistant: one wrong value nearly tripled it. The IQR, which uses only the middle half, did not change.

8. (a) Hillside: $190 - 165 = 25$ grams; Valley: $186 - 172 = 14$ grams. (b) Valley: the middle half of its apples vary over only 14 grams, compared with 25 grams at Hillside (its range, 50 grams, is also smaller than Hillside's 75 grams).
9. Q1 is a value in pounds, not a percent. About 25% of the hives produced less than 22 pounds, about 25% produced more than 41 pounds, and the middle half of the hives produced between 22 and 41 pounds of honey.
10. Median = 7th value = 24. Lower half 5, 8, 12, 12, 15, 20 gives Q1 $= \frac{12 + 12}{2} = 12$. Upper half 27, 31, 33, 40, 52, 66 gives Q3 $= \frac{33 + 40}{2} = 36.5$. Summary: 5, 12, 24, 36.5, 66; IQR $= 36.5 - 12 = 24.5$.
11. The quartiles mark off the middle half of the data, so about $\frac{40}{2} = 20$ boards.
12. The range uses only the coldest and warmest nights, so the single record night makes the range much larger than the usual night-to-night variation. The IQR is the spread of the middle half of the nights; the record night is in the bottom quarter, so it does not affect Q1 or Q3, and the IQR still describes the ordinary variation in that January's lows.
13. With n = 8, Q3 is the average of the 6th and 7th ordered values. The known values give a median $\frac{54 + 56}{2} = 55$ and Q1 $= \frac{51 + 53}{2} = 52$ only if the missing time is above 56. For Q3 = 59 we need $\frac{58 + x}{2} = 59$, so x = 60. Check: 50, 51, 53, 54, 56, 58, 60, 63 has summary 50, 52, 55, 59, 63. The missing time is 60 seconds.

LESSON 2.3 · 2 SESSIONS OF 45 MINUTES

Boxplots and Outliers

OBJECTIVES

- Draw a boxplot from a five-number summary and read a five-number summary from a boxplot.
- Identify outliers with the 1.5 × IQR rule and draw a modified boxplot.
- Compare two or more distributions with parallel boxplots, in context.
- Explain what a boxplot hides (clusters, gaps, the number of values) and when another graph is needed.

VIDEO

WATCH

Box Plots (2 of 2: Constructing the plot)

Eddie Woo · 7:12

When to watch: Before Example 1: drawing a boxplot from the five-number summary.

https://www.youtube.com/watch?v=c0cTHthV5GM

BACKUP

Outliers - Box and Whisker Plot (1.5 IQR)

Mario's Math Tutoring · 4:23

When to watch: Before Example 1(b): the 1.5 × IQR rule for outliers.

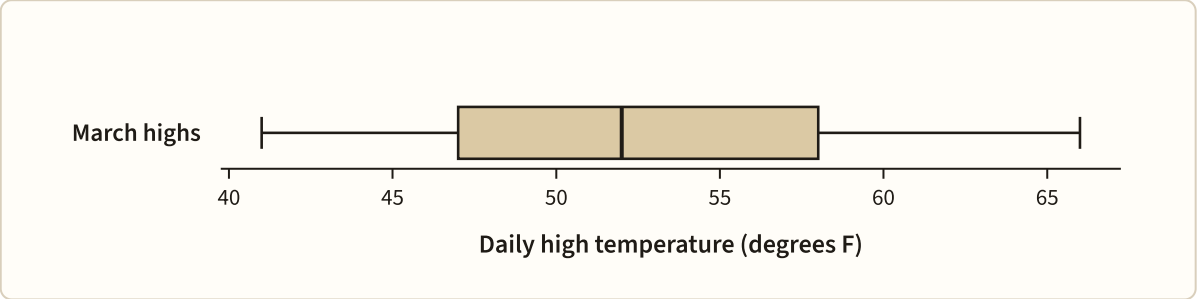
https://www.youtube.com/watch?v=vQ0L-gSUyvc

KEY TERMS

Boxplot	A graph of the five-number summary: a box from Q1 to Q3 with a line at the median, and whiskers reaching out toward the smallest and largest values.
Outlier (1.5 × IQR rule)	A value below Q1 − 1.5·IQR or above Q3 + 1.5·IQR. It is unusually far from the middle half of the data.
Fences	The two cutoffs Q1 − 1.5·IQR and Q3 + 1.5·IQR. They are not drawn on the graph; they only decide which values are outliers.
Modified boxplot	A boxplot in which outliers are drawn as separate dots and the whiskers reach only to the most extreme values that are not outliers.
Parallel boxplots	Two or more boxplots drawn on the same scale so that groups can be compared.

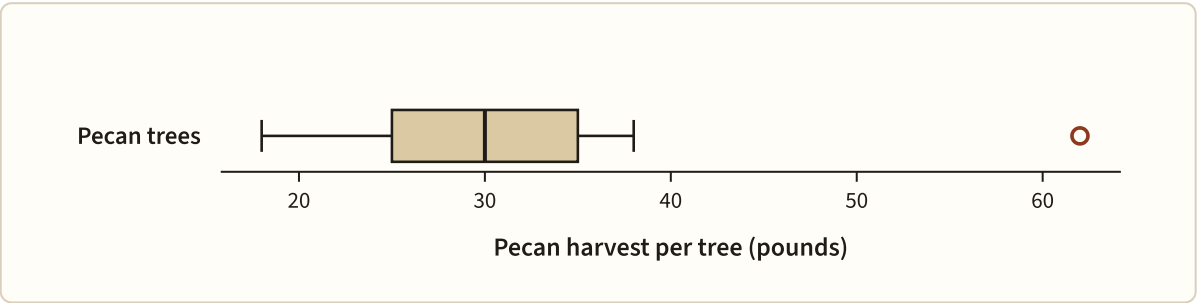
CORE IDEAS

A **boxplot** turns the five-number summary into a picture. Draw a number line with the units. Draw a box from Q1 to Q3 and a line across it at the median. Then draw whiskers from the box out toward the minimum and the maximum. The box holds the middle half of the data, so its length is the IQR; each whisker covers about a quarter of the data. Here is the boxplot of the 11 March high temperatures from Lesson 2.2 (41, 47, 52, 58, 66).



What counts as an **outlier**? We need a rule, so that different people looking at the same data agree. The course uses the **1.5 × IQR rule**: a value is an outlier if it lies more than 1.5 IQRs below Q1 or above Q3, that is, below Q1 − 1.5·IQR or above Q3 + 1.5·IQR. The yardstick is the IQR because the IQR is resistant: if we measured "far" with the range, the outlier itself would stretch the yardstick and hide itself. The number 1.5 is a convention, chosen because it flags only values that are really unusual compared with the middle half of the data.

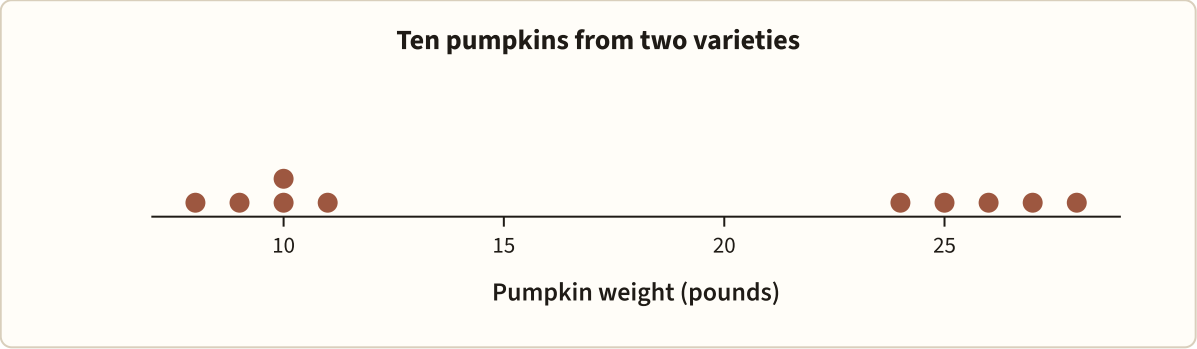
Example: the pecan harvests (pounds) from 11 trees were 18, 22, 25, 26, 28, 30, 31, 33, 35, 38, 62. The median is 30, Q1 = 25, Q3 = 35, so IQR = 10 and 1.5·IQR = 15. The fences are 25 − 15 = 10 and 35 + 15 = 50. Nothing is below 10, but 62 is above 50, so the 62-pound tree is an outlier.

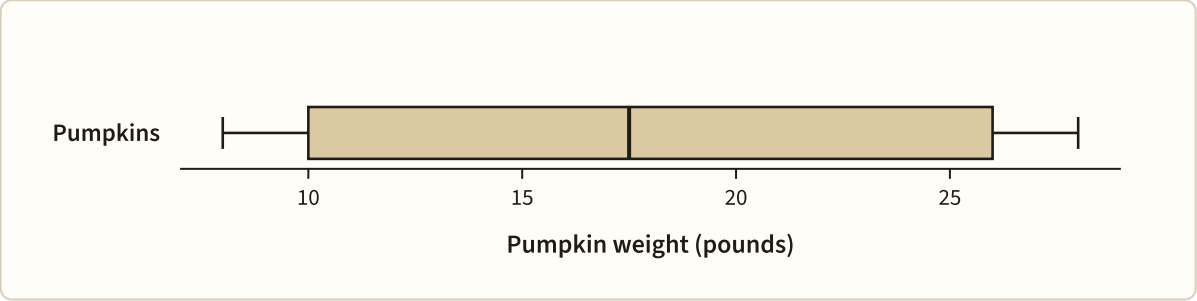


In a **modified boxplot** (the kind this course uses), each outlier is drawn as its own dot, and the whisker stops at the most extreme value that is NOT an outlier. For the pecans, the upper whisker ends at 38, not at 62 and not at the fence 50. Why? If the whisker ran out to 62, the picture would suggest trees were spread all the way from 38 to 62, when really one tree stands alone. The separate dot tells the truth: "here is one unusual tree."

An outlier is a value to **investigate**, not to throw away. Is it a mistake in recording? Then fix it. Is it real (a tree in an especially good spot, an Easter Sunday crowd)? Then keep it and mention it, because it may be the most interesting thing in the data.

Parallel boxplots drawn on one scale make comparisons quick: whose median is higher, whose box (IQR) is longer, who has outliers. But a boxplot hides a lot. It does not show how many values there are, and it can hide clusters and gaps. The 10 pumpkins below come from two varieties, one small and one large. The dotplot shows two clear clusters with nothing between 11 and 24 pounds; the boxplot shows a wide box with a median of 17.5 pounds, a weight that no pumpkin actually has. When you have the raw data, look at a dotplot, stemplot or histogram too.





KEY FORMULAS

$$\text{Lower fence} = Q_1 - 1.5 \times \text{IQR} \qquad \text{Upper fence} = Q_3 + 1.5 \times \text{IQR}$$

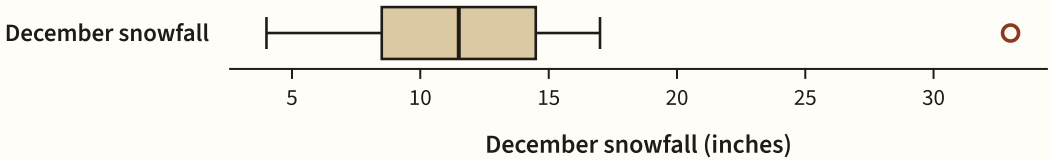
Outlier: a value below the lower fence or above the upper fence

From history. The boxplot was introduced by the American statistician John Tukey in the 1970s as part of what he called exploratory data analysis: quick, simple pictures that let you look at data before doing any formal calculations. He also made the stemplot popular.

Practice habit. Always compute both fences before you draw the whiskers. Write them down: "fences 10 and 50; outliers: 62; whiskers end at 18 and 38."

WORKED EXAMPLES

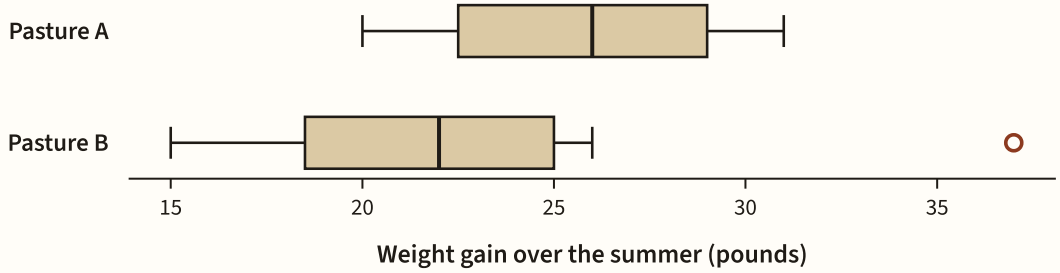
Example 1. A town's December snowfall (inches) over 12 years was 4, 7, 8, 9, 10, 11, 12, 12, 14, 15, 17, 33. Find the five-number summary, use the 1.5 × IQR rule to check for outliers, and describe the modified boxplot.



1. The data are in order and n = 12. Median $= \dfrac{11 + 12}{2} = 11.5$ inches.
2. Lower half: 4, 7, 8, 9, 10, 11, so Q1 $= \dfrac{8 + 9}{2} = 8.5$ inches.
3. Upper half: 12, 12, 14, 15, 17, 33, so Q3 $= \dfrac{14 + 15}{2} = 14.5$ inches.
4. IQR $= 14.5 - 8.5 = 6$ inches, and $1.5 \times 6 = 9$ inches.
5. Lower fence $= 8.5 - 9 = -0.5$; upper fence $= 14.5 + 9 = 23.5$ inches.
6. No value is below −0.5, but 33 is above 23.5, so 33 inches is an outlier.
7. Modified boxplot: box from 8.5 to 14.5 with the median at 11.5; lower whisker to 4; upper whisker to 17 (the largest value that is not an outlier); a separate dot at 33.
8. In context: a typical December brings about 8 to 15 inches of snow, and the 33-inch December was unusually snowy.

Answer: Five-number summary: 4, 8.5, 11.5, 14.5, 33 inches. Fences −0.5 and 23.5; the 33-inch December is an outlier. Whiskers end at 4 and 17 inches.

Example 2. Nine lambs grazed on Pasture A and nine on Pasture B all summer. Their weight gains (pounds) were: A: 20, 22, 23, 25, 26, 27, 28, 30, 31; B: 15, 18, 19, 21, 22, 23, 24, 26, 37. Find each five-number summary, identify any outliers, and compare the pastures using the parallel boxplots.



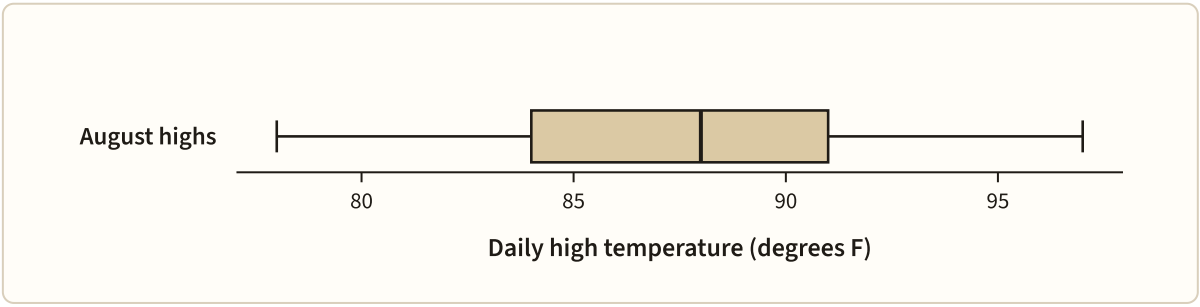
1. A (n = 9): median = 5th value = 26. Lower half 20, 22, 23, 25 gives Q1 = 22.5; upper half 27, 28, 30, 31 gives Q3 = 29.
2. A: IQR $= 29 - 22.5 = 6.5$; $1.5 \times 6.5 = 9.75$; fences $22.5 - 9.75 = 12.75$ and $29 + 9.75 = 38.75$. No outliers.
3. B: median = 22. Lower half 15, 18, 19, 21 gives Q1 = 18.5; upper half 23, 24, 26, 37 gives Q3 = 25.
4. B: IQR $= 25 - 18.5 = 6.5$; fences $18.5 - 9.75 = 8.75$ and $25 + 9.75 = 34.75$. Since $37 > 34.75$, the 37-pound gain is an outlier; B's upper whisker ends at 26.
5. Compare centers: the median gain on Pasture A (26 pounds) is 4 pounds more than on Pasture B (22 pounds).
6. Compare spreads: both IQRs are 6.5 pounds, so the middle halves are equally variable.
7. Conclusion: lambs on Pasture A typically gained more weight, with similar consistency; Pasture B had one lamb with an unusually large gain of 37 pounds that is worth checking.

Answer: A: 20, 22.5, 26, 29, 31 pounds, no outliers. B: 15, 18.5, 22, 25, 37 pounds, with 37 an outlier. Lambs on Pasture A typically gained about 4 pounds more (medians 26 versus 22), with the same IQR of 6.5 pounds.

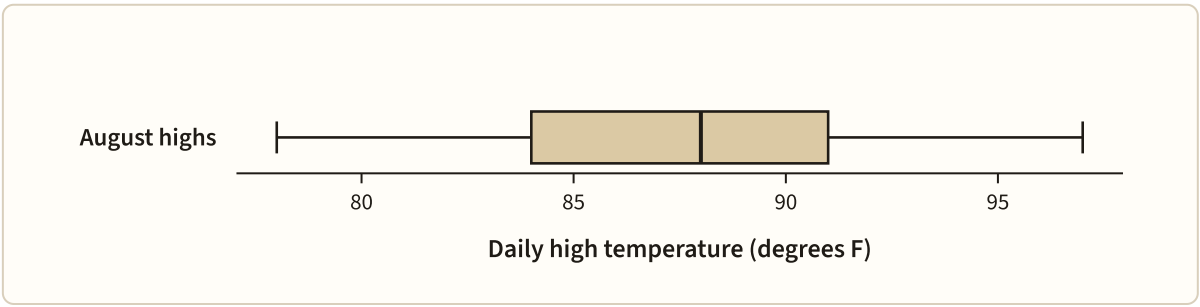
PRACTICE

1. EASY The five-number summary of the egg weights from a flock is 48, 54, 58, 62, 71 grams. Find the lower and upper fences. Are there any outliers?

2. **EASY** The boxplot shows the daily high temperatures in a town for one August. Read the median and find the IQR.



3. **EASY** Use the August boxplot. About what percent of the days had a high temperature above 91°F?



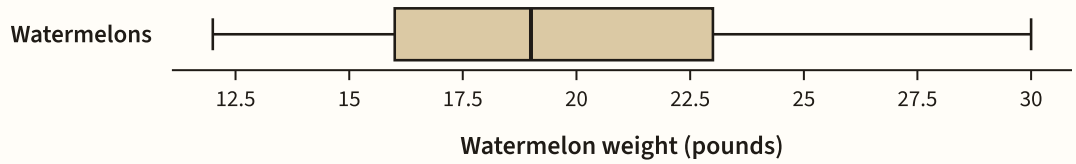
4. **EASY** On a hay farm, the numbers of bales per field have Q1 = 120 and Q3 = 136. Is a field that produced 165 bales an outlier?

5. **MEDIUM** Grace scored these points in nine basketball games: 4, 8, 9, 10, 11, 12, 13, 15, 27. (a) Find the five-number summary. (b) Which values, if any, are outliers? (c) Where do the whiskers of the modified boxplot end?

6. **MEDIUM** Ten hens laid these numbers of eggs in one week: 1, 5, 5, 6, 6, 6, 7, 7, 7, 7. (a) Find the fences. (b) Which value is an outlier?

7. **MEDIUM** The parallel boxplots compare the golf drives of Samuel and Henry over a season. (a) Find each golfer's median and IQR. (b) Which golfer is more consistent? Explain.

8. MEDIUM The boxplot shows the weights of the watermelons at a farm stand. From the boxplot alone, can you tell (a) how many watermelons were weighed? (b) whether any watermelon weighs exactly 20 pounds? (c) whether the weights form two separate clusters? Answer yes or no for each.



9. MEDIUM On a dairy farm, the 1.5 × IQR rule flags one cow whose daily milk yield is far above the rest of the herd. Mr. Miller wants to delete that cow's number from his records so that his summaries "look normal". Explain what he should do instead.

10. HARD The five-number summary of a batch of boards is 91.2, 95.8, 96.4, 97.0, 99.5 inches. (a) Find the fences. (b) Is the shortest board an outlier? (c) Is the longest board an outlier?

11. HARD Sunday attendance at a small country church over 15 weeks was 58, 61, 62, 64, 65, 66, 67, 68, 69, 70, 71, 72, 74, 75, 112. (a) Find the five-number summary. (b) Identify any outliers. (c) Where do the whiskers of the modified boxplot end?

12. HARD Explain why, in a modified boxplot, the whisker is drawn to the most extreme value that is not an outlier, rather than all the way to the maximum or to the fence.

13. CHALLENGE Seven workers picked these numbers of bushels of apples: 10, 12, 13, 15, 16, 18 and x, where x is at least 18. What is the smallest whole number x for which x is an outlier?

COMMON MISTAKES

- Drawing the whisker all the way to an outlier. Find the fences first; whiskers end at the most extreme values that are inside the fences, and outliers are separate dots.
- Calling a value an outlier because it "looks far". Use the 1.5 × IQR rule so your answer can be checked.
- Thinking a longer whisker or a bigger box means more data values. Each part of a boxplot holds about a quarter of the data whatever its length; length shows spread, not count.
- Deleting outliers without a reason. Check for errors; if the value is real, keep it and comment on it.

ANSWERS TO PRACTICE

1. IQR $= 62 - 54 = 8$; $1.5 \times 8 = 12$. Fences: $54 - 12 = 42$ and $62 + 12 = 74$ grams. The minimum 48 and maximum 71 are inside the fences, so there are no outliers.
2. Median = 88°F. IQR $= 91 - 84 = 7$°F.
3. 91°F is Q3, so about 25% of the days were hotter than 91°F.

4. IQR = 16, $1.5 \times 16 = 24$, upper fence $= 136 + 24 = 160$. Since 165 > 160, yes, it is an outlier.
5. (a) Median 11; Q1 $= \dfrac{8+9}{2} = 8.5$; Q3 $= \dfrac{13+15}{2} = 14$. Summary: 4, 8.5, 11, 14, 27. (b) IQR = 5.5, $1.5 \times 5.5 = 8.25$, fences 0.25 and 22.25: 27 is an outlier. (c) Lower whisker at 4, upper whisker at 15.
6. (a) Median $= \dfrac{6+6}{2} = 6$; lower half 1, 5, 5, 6, 6 gives Q1 = 5; upper half 6, 7, 7, 7, 7 gives Q3 = 7. IQR = 2, $1.5 \times 2 = 3$; fences 2 and 10. (b) The hen that laid 1 egg is an outlier (1 < 2). Perhaps she was sick or molting.
7. (a) Samuel: median 240 yards, IQR $= 250 - 225 = 25$ yards. Henry: median 236 yards, IQR $= 241 - 230 = 11$ yards. (b) Henry: the middle half of his drives fall within 11 yards, less than half Samuel's spread, even though Samuel typically hits a little farther.
8. (a) No: a boxplot of 10 melons and of 1000 melons can look the same. (b) No: only the five summary values are shown. (c) No: a boxplot can hide clusters and gaps; a dotplot or histogram would show them.
9. He should first check whether the number is a recording error; if it is, correct it. If it is real, he should keep it: it is a true value, and an unusually productive cow is worth knowing about (her breed, feed or health might explain it). He can report the summaries with a note about the outlier, and use the median and IQR, which it hardly affects.
10. (a) IQR $= 97.0 - 95.8 = 1.2$; $1.5 \times 1.2 = 1.8$; fences $95.8 - 1.8 = 94.0$ and $97.0 + 1.8 = 98.8$ inches. (b) Yes: 91.2 < 94.0. (c) Yes: 99.5 > 98.8.
11. (a) Median = 8th value = 68. Lower half 58, 61, 62, 64, 65, 66, 67 gives Q1 = 64; upper half 69, 70, 71, 72, 74, 75, 112 gives Q3 = 72. Summary: 58, 64, 68, 72, 112. (b) IQR = 8, $1.5 \times 8 = 12$, fences 52 and 84; 112 is an outlier (most likely Easter Sunday). (c) Whiskers end at 58 and 75.
12. Drawing to the maximum would make one unusual value look like part of a long, continuous spread of data. Drawing to the fence would show a stretch of the number line where there may be no data at all. Stopping at the last non-outlier shows exactly how far the ordinary data reach, and the separate dot shows the unusual value honestly.
13. Sorted: 10, 12, 13, 15, 16, 18, x. Median 15; lower half 10, 12, 13 gives Q1 = 12; upper half 16, 18, x gives Q3 = 18 (x is the largest, so the middle of the upper half is 18 whatever x is). IQR = 6; upper fence $= 18 + 9 = 27$. The value x is an outlier when x > 27, so the smallest whole number is 28 bushels.

LESSON 2.4 · 2 SESSIONS OF 45 MINUTES

Standard Deviation

OBJECTIVES

- Explain why deviations from the mean add to 0 and why we square them.
- Compute the variance and standard deviation of a small data set by hand with a table and with a calculator.
- Explain why the sample standard deviation divides by n − 1 and distinguish s from the population standard deviation σ.
- Interpret the standard deviation in context as a typical distance from the mean.

VIDEO

WATCH

Measures of Variability (Variance, Standard Deviation, Range, Mean Absolute Deviation)

jbstatistics · 12:11

When to watch: Before Example 1: variance and standard deviation, dividing by n − 1.

https://www.youtube.com/watch?v=Cx2tGUze60s

BACKUP

The Sample Variance: Why Divide by n-1?

jbstatistics · 6:53

When to watch: After Example 1: why the sample variance divides by n − 1.

https://www.youtube.com/watch?v=9ONRMymR2Eg

> **Note:** The main video also shows the mean absolute deviation, which the course does not use.

KEY TERMS

Deviation	How far a value is from the mean, with a sign: $x - \bar{x}$. Values above the mean have positive deviations; values below have negative deviations.
Variance s^2	The sum of the squared deviations divided by n − 1: roughly the average squared distance from the mean. Its units are squared units.
Standard deviation s	The square root of the variance. It measures a typical distance of the values from their mean, in the original units.

Population standard deviation σ	The standard deviation of a whole population, written σ (sigma); its variance divides by n instead of n − 1, and the mean is written μ (mu).
Degrees of freedom	The number of deviations that are free to vary: n − 1, because the deviations must add to 0, so the last one is fixed by the others.

CORE IDEAS

The range and IQR each use just two numbers. We also want a measure of spread that uses **every** value and goes naturally with the mean. The idea is to ask how far each value is from the mean. That distance, with its sign, is the **deviation** $x - \bar{x}$.

Take the snowfall data from Lesson 2.1: 3, 4, 4, 5, 9 inches, with $\bar{x} = 5$. The deviations are −2, −1, −1, 0 and 4. Could we simply average them? No: they add to exactly 0. This always happens, because the mean is the balance point: the negative deviations exactly cancel the positive ones. (In symbols, $\sum(x - \bar{x}) = \sum x - n\bar{x} = n\bar{x} - n\bar{x} = 0$.) So the average deviation is always 0 and tells us nothing.

To stop the cancelling we **square** each deviation. Squares are never negative, so every value now adds to the total, and a value far from the mean counts much more than a value close to it (a deviation of 4 contributes 16, a deviation of 1 contributes only 1). Squares also have mathematical properties that make later work much simpler; in Unit 7 you will see that variances add in a way that distances do not.

Deviations for the snowfall data (mean 5 inches)

Day	x (inches)	x − x̄	(x − x̄)²
1	3	-2	4
2	4	-1	1
3	4	-1	1
4	5	0	0
5	9	4	16
Sum	25	0	22

Add the squared deviations: $4 + 1 + 1 + 0 + 16 = 22$. Divide by $n - 1 = 4$ to get the **variance** $s^2 = \dfrac{22}{4} = 5.5$ square inches. Square inches are awkward units for snowfall, so take the square root to return to inches: the **standard deviation** is $s = \sqrt{5.5} \approx 2.3$ inches. The daily snowfalls typically differ from the mean of 5 inches by about 2.3 inches.

Why divide by n − 1 and not n? Our data are a sample, and what we really want is the spread around the population mean μ. But we measure deviations from $\bar{x}$, and $\bar{x}$ is computed from these very values, so it sits right in the middle of them. In fact the sum of squared deviations is as small as possible when measured from $\bar{x}$: from 5 it is 22, but if the true mean were 6, the deviations would be −3, −2, −2, −1, 3 and their squares would add to 27. Measuring from $\bar{x}$ therefore tends to make the data look a little less spread out than they really are, and dividing by the smaller number n − 1 corrects for this. Another way to see it: the deviations must add to 0, so once you know n − 1 of them, the last one is forced. Only n − 1 deviations are free; we say there are n − 1 **degrees of freedom**.

If the data are an entire **population** (every tree in a small orchard, every member of a choir), there is nothing to estimate: we use the population mean μ and divide by n, and call the result σ. For the snowfall data treated as a population, $\sigma = \sqrt{22 \div 5} = \sqrt{4.4} \approx 2.1$ inches. In this course "standard deviation" means s unless the problem says the data are a whole population. On a calculator with one-variable statistics, s is usually labeled **Sx** and σ is labeled **σx**; choose Sx.

Facts to remember about s: s is never negative; s = 0 only when every value is the same (no spread at all); larger s means more spread. And s is **not resistant**: because deviations are squared, one outlier far from the mean can make s much larger. Like the mean, s is best for distributions that are roughly symmetric without outliers.

KEY FORMULAS

$$s^2 = \frac{\sum (x - \bar{x})^2}{n - 1} \qquad s = \sqrt{\frac{\sum (x - \bar{x})^2}{n - 1}}$$

$$\sigma = \sqrt{\frac{\sum (x - \mu)^2}{n}} \quad \text{(whole population)}$$

$$\sum (x - \bar{x}) = 0 \quad \text{always}$$

From history. The name "standard deviation" was introduced by the English statistician Karl Pearson in the 1890s. The idea of measuring spread with squared errors is older: astronomers and mathematicians such as Gauss used it in the early 1800s when combining repeated measurements of the positions of planets and stars.

Practice habit. Make a three-column table (x, x − x̄, (x − x̄)²) and check that the deviations add to 0 before you square them. If they do not, your mean or a subtraction is wrong.

Example 1. Grace weighed six eggs from her hens: 56, 58, 60, 61, 62, 63 grams. Find the mean and the standard deviation, and interpret the standard deviation.

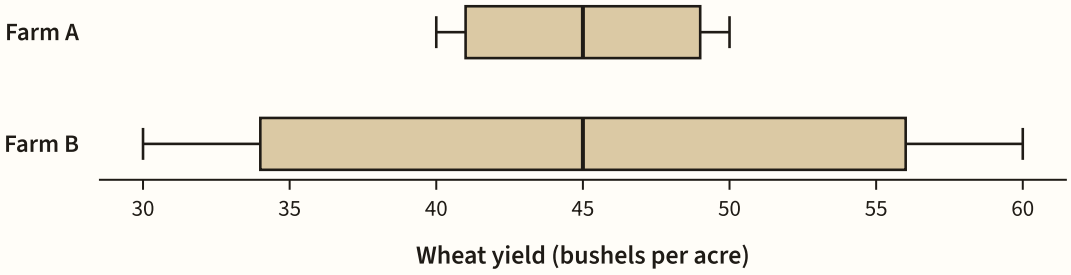
Egg weights (mean 60 grams)

Egg	x (grams)	$x - \bar{x}$	$(x - \bar{x})^2$
1	56	-4	16
2	58	-2	4
3	60	0	0
4	61	1	1
5	62	2	4
6	63	3	9
Sum	360	0	34

1. Mean: $\bar{x} = \dfrac{56 + 58 + 60 + 61 + 62 + 63}{6} = \dfrac{360}{6} = 60$ grams.
2. Deviations: −4, −2, 0, 1, 2, 3. Check: $-4 - 2 + 0 + 1 + 2 + 3 = 0$.
3. Squared deviations: 16, 4, 0, 1, 4, 9. Sum $= 34$.
4. Variance: $s^2 = \dfrac{34}{6 - 1} = \dfrac{34}{5} = 6.8$ square grams.
5. Standard deviation: $s = \sqrt{6.8} \approx 2.6077$, so $s \approx 2.6$ grams (one more decimal place than the data).
6. Interpretation: the weights of these eggs typically differ from the mean of 60 grams by about 2.6 grams.

Answer: $\bar{x} = 60.0$ grams and $s \approx 2.6$ grams: the egg weights typically differ from the mean of 60 grams by about 2.6 grams.

Example 2. Over five years, Farm A's wheat yields were 40, 42, 45, 48, 50 bushels per acre and Farm B's were 30, 38, 45, 52, 60 bushels per acre. Find the mean and standard deviation for each farm, and compare the farms in context. Round standard deviations to one decimal place.



1. Farm A: $\bar{x} = \dfrac{225}{5} = 45$. Deviations −5, −3, 0, 3, 5; squares 25, 9, 0, 9, 25; sum 68.
2. Farm A: $s^2 = \dfrac{68}{4} = 17$, $s = \sqrt{17} \approx 4.1$ bushels per acre.
3. Farm B: $\bar{x} = \dfrac{225}{5} = 45$. Deviations −15, −7, 0, 7, 15; squares 225, 49, 0, 49, 225; sum 548.
4. Farm B: $s^2 = \dfrac{548}{4} = 137$, $s = \sqrt{137} \approx 11.7$ bushels per acre.
5. Same center: both farms averaged 45 bushels per acre.
6. Different spread: Farm A's yields typically differ from 45 by about 4 bushels per acre, Farm B's by about 12.
7. Conclusion: the farms have the same mean yield, but Farm B's harvest is far less predictable from year to year.

Answer: Farm A: $\bar{x} = 45$, $s \approx 4.1$ bushels per acre. Farm B: $\bar{x} = 45$, $s \approx 11.7$ bushels per acre. Same mean, but Farm B's yields vary almost three times as much from year to year.

PRACTICE

1. EASY Ruth collected 10, 12 and 14 eggs on three days. Find the standard deviation.
2. EASY The deviations from the mean of four board lengths are −0.3, 0.1, 0.4 inches and one more. What is the fourth deviation?
3. EASY Five sacks of flour each weigh exactly 50 pounds. What is the standard deviation of their weights?
4. EASY The variance of a runner's 100-meter times is 2.25 square seconds. Find the standard deviation.
5. MEDIUM The noon wind speeds on four days were 6, 8, 12 and 14 miles per hour. Find the mean and the standard deviation (one decimal place).

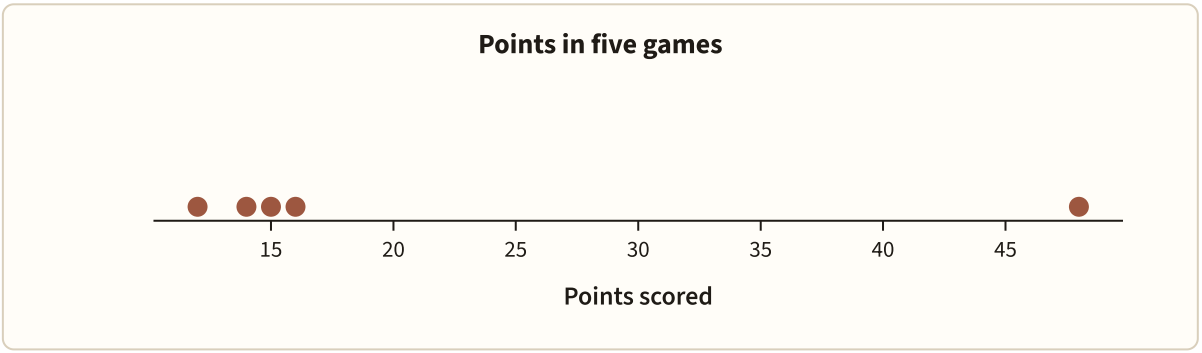
6. MEDIUM Samuel's times for a 3-mile run on five days were 21, 23, 24, 26 and 31 minutes. Find the mean and the standard deviation (one decimal place).

7. MEDIUM A tiny orchard has exactly four apple trees, which yielded 30, 34, 38 and 42 bushels. (a) Treating the four trees as the whole population, find σ. (b) Find s for the same numbers. Round to one decimal place. (c) Which is correct to report for this orchard?

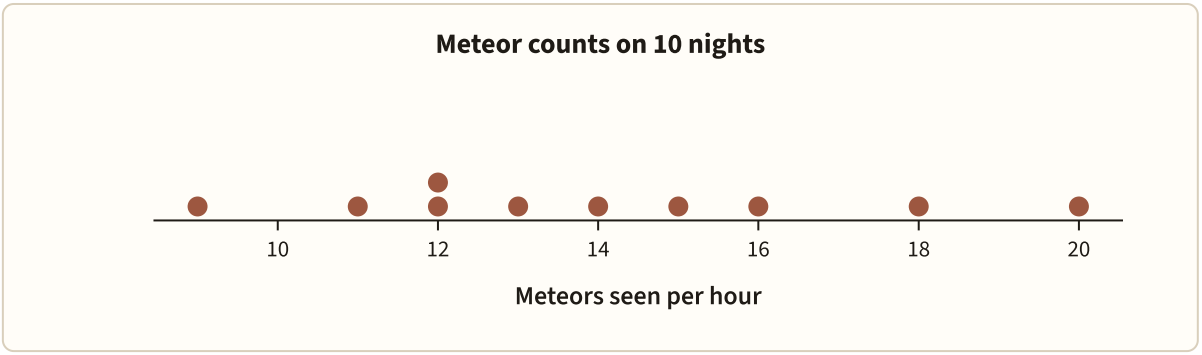
8. MEDIUM A farmer reports that the eggs from her flock have a mean weight of 58 grams with a standard deviation of 3.5 grams. Interpret the standard deviation in context.

9. MEDIUM The high temperatures (°F) on six winter days were: Town A: 10, 10, 10, 20, 20, 20; Town B: 10, 14, 15, 15, 16, 20. Both means are 15°F. (a) Without calculating, which town has the larger standard deviation? Explain. (b) Compute both standard deviations (one decimal place) to check.

10. HARD A basketball player scored 12, 14, 15, 16 and 18 points in five games. (a) Find s. (b) Suppose the 18 had been 48, as shown in the dotplot. Find s now. Round to one decimal place. What does this show?



11. HARD Henry counted the meteors he saw in one hour on each of 10 nights (dotplot). Use your calculator's one-variable statistics (or a table) to find the mean and standard deviation. Round to one decimal place.



12. HARD Explain in your own words why the sample standard deviation divides by n − 1 rather than by n.

13. CHALLENGE Samuel's five run times (21, 23, 24, 26, 31 minutes) have mean 25 and $s \approx 3.8$ minutes. On a sixth day he runs exactly 25 minutes. (a) Find the new standard deviation (one decimal place). (b) Explain why it went down even though nothing farther from the mean was added.

COMMON MISTAKES

- Dividing by n instead of n − 1 for a sample (or choosing σx instead of Sx on the calculator). Use n − 1 unless the data are a whole population.
- Forgetting the square root and reporting the variance as the standard deviation. Check the units: the standard deviation has the same units as the data.
- Squaring a negative deviation incorrectly, such as $(-4)^2 = -16$. A squared deviation is never negative.
- Rounding the mean or the deviations early. Keep full precision until the final square root, then round.

ANSWERS TO PRACTICE

1. $\bar{x} = 12$; deviations −2, 0, 2; squares 4, 0, 4, sum 8. $s^2 = \dfrac{8}{2} = 4$, so $s = 2$ eggs.
2. Deviations add to 0: $-0.3 + 0.1 + 0.4 = 0.2$, so the fourth is −0.2 inches.
3. Every deviation is 0, so $s = 0$ pounds. There is no spread at all.
4. $s = \sqrt{2.25} = 1.5$ seconds.
5. $\bar{x} = \dfrac{40}{4} = 10$. Deviations −4, −2, 2, 4; squares 16, 4, 4, 16; sum 40. $s^2 = \dfrac{40}{3} \approx 13.333$, $s \approx 3.7$ miles per hour.
6. $\bar{x} = \dfrac{125}{5} = 25$. Deviations −4, −2, −1, 1, 6; squares 16, 4, 1, 1, 36; sum 58. $s^2 = \dfrac{58}{4} = 14.5$, $s = \sqrt{14.5} \approx 3.8$ minutes.
7. Mean 36; deviations −6, −2, 2, 6; squares 36, 4, 4, 36; sum 80. (a) $\sigma = \sqrt{\dfrac{80}{4}} = \sqrt{20} \approx 4.5$ bushels. (b) $s = \sqrt{\dfrac{80}{3}} \approx 5.2$ bushels. (c) σ, because these four trees are the entire population we care about.
8. The weights of the eggs typically differ from the mean of 58 grams by about 3.5 grams; some are closer and some farther, but 3.5 grams is a typical distance from the mean.
9. (a) Town A: every day is 5 degrees from the mean, while most of Town B's days are within 1 degree of it. (b) A: squares 25 six times, sum 150, $s = \sqrt{30} \approx 5.5$°F. B: deviations −5, −1, 0, 0, 1, 5; squares sum $25 + 1 + 0 + 0 + 1 + 25 = 52$; $s = \sqrt{10.4} \approx 3.2$°F.
10. (a) Mean 15; deviations −3, −1, 0, 1, 3; squares sum 20; $s = \sqrt{5} \approx 2.2$ points. (b) Mean 21; deviations −9, −7, −6, −5, 27; squares 81, 49, 36, 25, 729, sum 920; $s = \sqrt{230} \approx 15.2$ points. One outlier made s about seven times as large: s is not resistant.
11. $\sum x = 140$, $\bar{x} = 14.0$ meteors. Deviations −2, 1, −5, 4, 0, −3, 2, −1, 6, −2; squared deviations sum to 100. $s = \sqrt{\dfrac{100}{9}} \approx 3.3$ meteors per hour.

12. We want the spread around the population mean, but we can only measure deviations from $\bar{x}$, which is calculated from the same data and sits in their middle. Squared deviations from $\bar{x}$ are as small as they can be, so they slightly understate the true spread; dividing by the smaller number $n - 1$ makes up for this. Also, the deviations must add to 0, so only $n - 1$ of them are free to vary.
13. (a) The mean stays 25 and the new deviation is 0, so the sum of squares stays 58 but now $n - 1 = 5$: $s = \sqrt{\dfrac{58}{5}} = \sqrt{11.6} \approx 3.4$ minutes. (b) A value right at the mean adds nothing to the squared deviations but adds one more value to divide among, so the typical distance from the mean shrinks.

LESSON 2.5 · 2 SESSIONS OF 45 MINUTES

Comparing Distributions

OBJECTIVES

- Choose suitable numerical summaries: the five-number summary and IQR for skewed data or data with outliers, the mean and standard deviation for roughly symmetric data.
- Compare two or more groups in context by shape, center, spread and outliers, using numbers and graphs.
- Write a clear comparison paragraph with comparative words, numbers and units.

VIDEO

WATCH

Analysing Data (Example: comparing golfers)

Eddie Woo · 12:58

When to watch: Before Example 1: comparing two distributions by center and spread (two golfers).

https://www.youtube.com/watch?v=mOEZWKdDCfI

BACKUP

Choosing Statistical Measures

Eddie Woo · 12:42

When to watch: After Example 2: choosing which measures of center and spread to report.

https://www.youtube.com/watch?v=q0rsZPp7euY

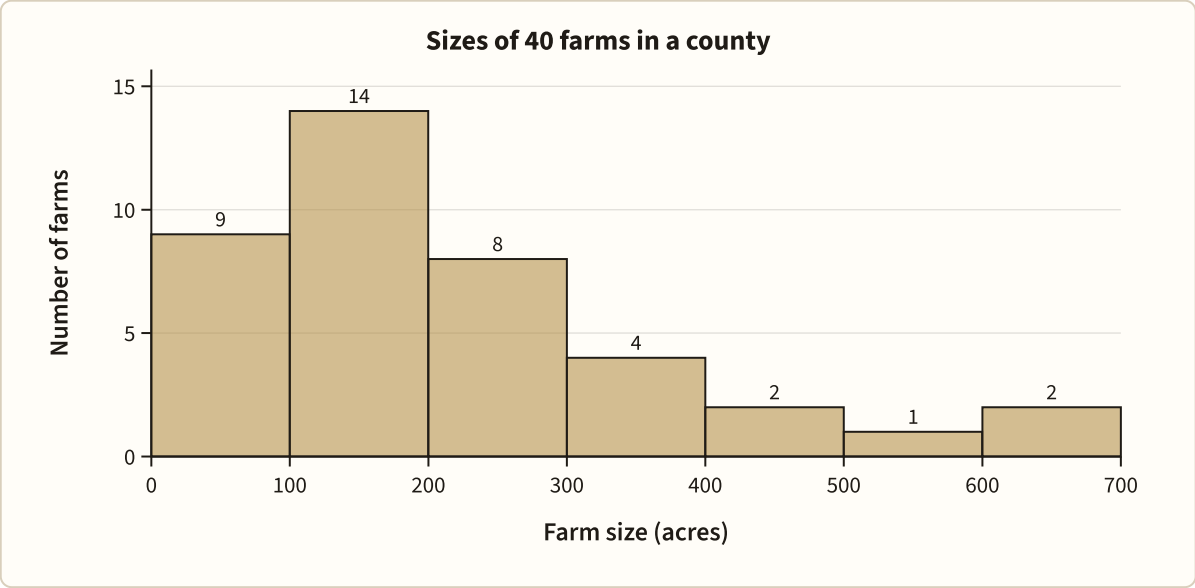
KEY TERMS

Resistant summary	A summary not much affected by outliers or long tails: the median and the IQR (and the five-number summary).
Nonresistant summary	A summary that outliers and skewness can pull strongly: the mean, the standard deviation and the range.
Comparison	A description of two or more distributions side by side that uses comparative words ("greater than", "about 6 pounds more", "less variable") rather than two separate descriptions.
Parallel graphs	Graphs of several groups drawn on the same scale, such as parallel boxplots, back-to-back stemplots or stacked histograms, so that the groups can be compared by eye.

CORE IDEAS

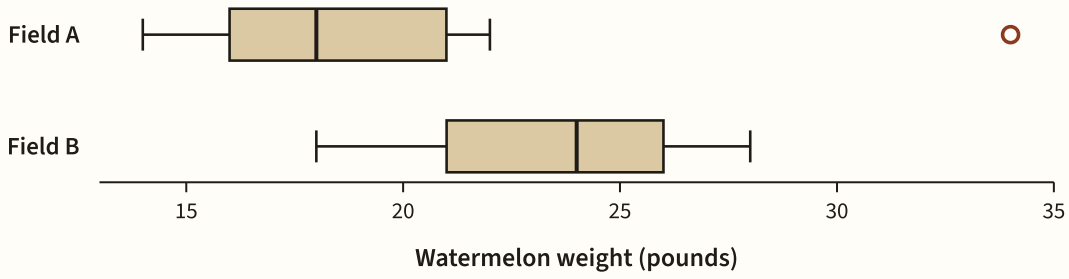
In Unit 1 you described distributions in words: shape, center, spread and outliers, always in context. Now you have numbers to go with the words. The question is which numbers to use, and how to use them to compare groups fairly.

The rule of thumb is simple. If a distribution is **roughly symmetric with no outliers**, report the **mean and standard deviation**. If it is **skewed or has outliers**, report the **median and IQR** (or the whole five-number summary). The reason is resistance. The mean and s use the size of every value, so a long tail or an outlier pulls them; the median and IQR depend only on the middle of the ordered data, so they still describe the bulk of the values. In a symmetric distribution the mean and median agree anyway, and the mean and s carry more information because they use every value.



The farm sizes above are strongly skewed right: most farms are under 300 acres, but a few are very large. The mean would be pulled toward those few big farms and s would be inflated by them, so the median and IQR give a truer picture of a typical farm and of the usual variation. Always look at a graph before you choose your numbers.

When comparing groups, use the **same** kind of summary for both, even if only one group is skewed; it is not fair to compare one group's median with another group's mean. Draw the groups on the **same scale** (parallel boxplots, a back-to-back stemplot, or histograms with the same classes). Then compare all four features, using comparison words and numbers with units.



Here is a model comparison for the watermelons from two fields shown above (Worked Example 1 finds the numbers). "Watermelons from Field B were typically heavier than those from Field A: the median weight was 24 pounds for Field B and 18 pounds for Field A, about 6 pounds more. The two fields had similar spread, with an IQR of 5 pounds each. Field B's weights are roughly symmetric with no outliers, while Field A's are roughly symmetric except for one melon of 34 pounds, which is an outlier by the 1.5 × IQR rule." Notice that every sentence compares, every number has units, and the context (watermelons, fields) never disappears.

Finally, numbers describe only the data you have. A difference between two groups may be real, or it may be the luck of which melons happened to grow in this one summer. Later in the course (Units 11 and 12) you will learn how to decide whether a difference is too large to be explained by chance alone. For now, describe carefully what the data show.

KEY FORMULAS

Roughly symmetric, no outliers: report $\bar{x}$ and s

Skewed or with outliers: report the median and IQR (five-number summary)

Practice habit. Ask "compared to what?" Every comparison sentence should name both groups, use a comparison word, and give numbers with units: "the median for Field B was 6 pounds higher than for Field A."

Example 1. The back-to-back stemplot shows the weights of 11 watermelons from each of two fields. (a) Find the median and IQR for each field. (b) Write a paragraph comparing the two distributions.

Watermelon weights from two fields

Field B		Field A
	1	4 5 6 7 7 8 9
8 7 6 5 4 4 3 2 1 0	2	0 1 2
	3	4

Key: 2 | 4 means 24 pounds

1. Field A, ordered: 14, 15, 16, 17, 17, 18, 19, 20, 21, 22, 34. Median = 6th value = 18 pounds.
2. Field A: lower half 14, 15, 16, 17, 17 gives Q1 = 16; upper half 19, 20, 21, 22, 34 gives Q3 = 21; IQR = 5 pounds.
3. Field A outlier check: $1.5 \times 5 = 7.5$; fences 8.5 and 28.5; 34 is an outlier.
4. Field B, ordered: 18, 20, 21, 22, 23, 24, 24, 25, 26, 27, 28. Median = 24 pounds; Q1 = 21, Q3 = 26; IQR = 5 pounds. Fences 13.5 and 33.5; no outliers.
5. Choose summaries: Field A has an outlier, so use medians and IQRs for both fields.
6. Center: Field B's median (24 pounds) is 6 pounds higher than Field A's (18 pounds).
7. Spread: equal IQRs of 5 pounds.
8. Shape and outliers: Field B is roughly symmetric with no outliers; Field A is roughly symmetric apart from one high outlier at 34 pounds.

Answer: (a) Field A: median 18 pounds, IQR 5 pounds. Field B: median 24 pounds, IQR 5 pounds. (b) Field B's melons were typically about 6 pounds heavier (medians 24 versus 18 pounds), with the same spread (IQR 5 pounds each). Field B's weights are roughly symmetric with no outliers; Field A's are roughly symmetric except for one unusually heavy 34-pound melon.

Example 2. Two bakeries sell loaves labeled 16 ounces. The table summarizes the weights of 20 loaves from each. (a) How much larger is Mill Road's mean? (b) Which bakery's loaves are more consistent? Explain in context.

Weights of loaves labeled 16 ounces

Bakery	Number of loaves	Mean (ounces)	Standard deviation (ounces)	Shape
Oak Street	20	16.1	0.3	Roughly symmetric
Mill Road	20	16.4	0.8	Roughly symmetric

1. Both distributions are roughly symmetric, so the mean and standard deviation are suitable summaries.
2. Difference in means: $16.4 - 16.1 = 0.3$ ounces.
3. Compare spreads: Oak Street's s is 0.3 ounces; Mill Road's is 0.8 ounces, more than twice as large.
4. Interpret s: Oak Street's loaves typically differ from their mean by about 0.3 ounces; Mill Road's by about 0.8 ounces.
5. Context: Oak Street's loaves cluster closely around 16.1 ounces, so a customer gets nearly the same weight every time. Mill Road's loaves are heavier on average, but their weights are much less predictable: a typical loaf is about 0.8 ounce away from 16.4 ounces, so some are well over and some are under the labeled 16 ounces.
6. Conclusion: Mill Road's loaves are slightly heavier on average, but Oak Street's are much more consistent.

Answer: (a) $16.4 - 16.1 = 0.3$ ounces. (b) Oak Street: its standard deviation is 0.3 ounces compared with 0.8 ounces at Mill Road, so its loaf weights stay much closer to their mean, even though Mill Road's loaves are 0.3 ounces heavier on average.

PRACTICE

1. EASY The table gives five-number summaries of July rainfall over 20 years in two towns. How much greater is Dale's median July rainfall than Hill's?

July rainfall over 20 years (inches)

Town	Min	Q1	Median	Q3	Max
Hill	1.2	2.4	3.1	3.9	6.8
Dale	0.8	2.9	3.6	4.2	5

2. EASY Use the rainfall table. Find the IQR for Hill and for Dale.

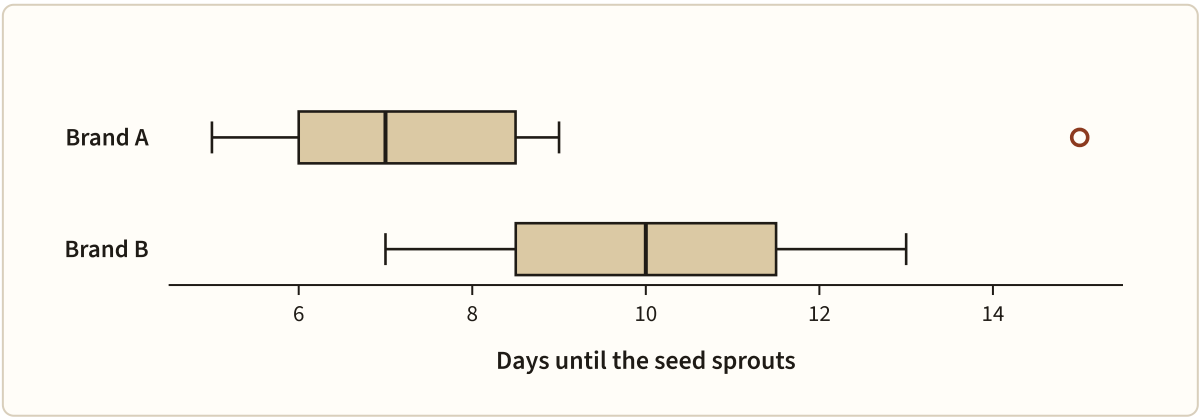
July rainfall over 20 years (inches)

Town	Min	Q1	Median	Q3	Max
Hill	1.2	2.4	3.1	3.9	6.8
Dale	0.8	2.9	3.6	4.2	5

3. EASY The amounts of honey harvested per hive in a county are strongly skewed right. Which summaries of center and spread should a beekeepers' club report, and why?

4. EASY Over a season, Ruth's golf scores had mean 82 strokes and standard deviation 3.1 strokes; Grace's had mean 85 strokes and standard deviation 5.4 strokes. Both distributions were roughly symmetric. (a) How many strokes lower was Ruth's mean? (b) Whose scores were more consistent?

5. MEDIUM The parallel boxplots show the number of days until seeds sprouted for nine bean seeds of each of two brands. The data were A: 5, 6, 6, 7, 7, 8, 8, 9, 15 and B: 7, 8, 9, 9, 10, 11, 11, 12, 13. (a) Find each brand's median and IQR. (b) Compare the two brands in context.



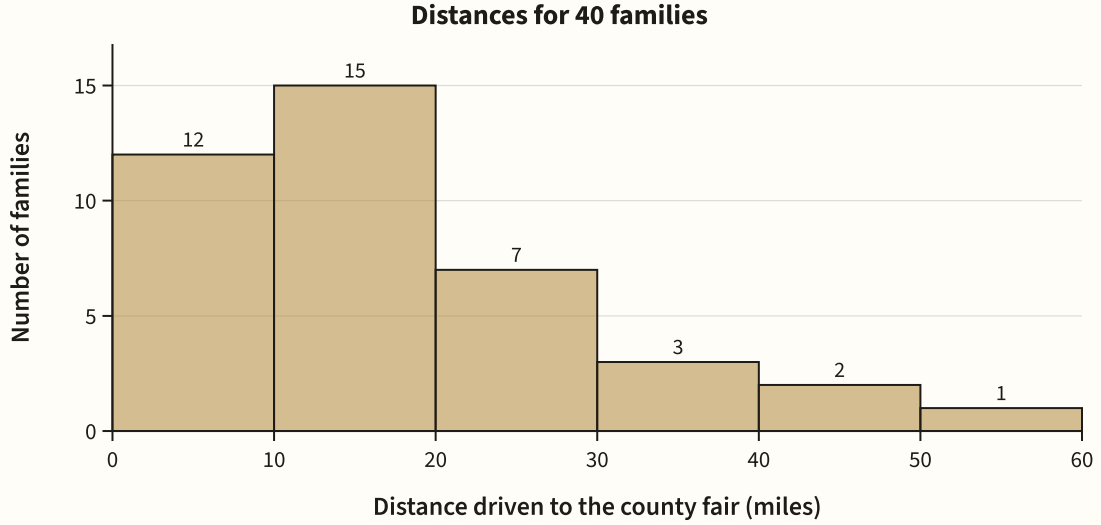
6. **MEDIUM** The back-to-back stemplot (split stems) shows the lengths of 10 trout caught in each of two streams. Both distributions are roughly symmetric. (a) Find the mean length for each stream. (b) Which stream's trout were typically longer, and by how much?

Lengths of trout from two streams

South Stream								North Stream							
							0								
						9	0	8	9						
4	4	3	3	2	1	1	1	0	0	1	1	2	2	3	4
			8	6	5		1								

Key: 1 | 3 means 13 inches

7. **MEDIUM** The histogram shows how far 40 families drove to the county fair. A fair official wants to report a typical distance and its spread. Which summaries should she use? Explain using the graph.



8. **MEDIUM** A 4-H member grew 15 sunflowers with fertilizer and 15 without. The table summarizes their heights. (a) Find the difference in mean heights. (b) Write two or three sentences comparing the groups.

Heights of sunflowers after 10 weeks (inches)

Group	Number of plants	Mean	Standard deviation	Shape
Fertilized	15	68.2	4.1	Roughly symmetric
Not fertilized	15	61.5	4.3	Roughly symmetric

9. HARD Field A's watermelon weights were 14, 15, 16, 17, 17, 18, 19, 20, 21, 22, 34 pounds. (a) Find the mean with all 11 melons. (b) Find the mean without the 34-pound melon. Round to one decimal place. (c) The median changes from 18 to 17.5 pounds when the outlier is removed. What do these results show?

10. HARD Samuel wants to compare the daily rainfall at two weather stations. One distribution is roughly symmetric and the other is strongly skewed right. He plans to report the mean for the first station and the median for the second. Explain what is wrong with his plan and what he should do.

11. HARD Henry's five 100-meter times were 13.1, 13.4, 13.2, 13.5, 13.3 seconds, and Samuel's were 12.6, 13.9, 13.0, 14.0, 12.5 seconds. (a) Find the mean and standard deviation for each runner (two decimal places for s). (b) The coach needs a runner who will reliably run near 13.3 seconds. Whom should he choose? Explain.

12. HARD A newsletter reports, "The mean yield on the new farm was 52 bushels per acre, compared with 48 on the old farm." Name two other things you would want to know before concluding that the new farm is really more productive, and explain why.

13. CHALLENGE For a church bake sale, Ruth filled 20 jars of honey with a mean weight of 16.1 ounces and Grace filled 30 jars with a mean weight of 16.4 ounces. (a) Find the mean weight of all 50 jars together. (b) If you also knew the two medians, could you find the median of all 50 jars from them alone? Answer yes or no.

COMMON MISTAKES

- Reporting the mean and standard deviation for strongly skewed data or data with outliers. Look at a graph first; use the median and IQR when you see skewness or outliers.
- Writing two separate descriptions instead of a comparison. Use comparison words ("higher than", "about 6 pounds more", "less variable") and name both groups in each sentence.
- Comparing one group's mean with another group's median. Use the same summaries for every group you compare.
- Forgetting units and context. "24 versus 18" means nothing; "24 versus 18 pounds" for the watermelons in Fields B and A does.

ANSWERS TO PRACTICE

1. $3.6 - 3.1 = 0.5$ inches.
2. Hill: $3.9 - 2.4 = 1.5$ inches. Dale: $4.2 - 2.9 = 1.3$ inches.
3. The median and IQR (or the five-number summary), because the few very large harvests in the long right tail would pull the mean up and inflate the standard deviation, while the median and IQR resist them.
4. (a) $85 - 82 = 3$ strokes (in golf, lower is better). (b) Ruth's: her standard deviation (3.1 strokes) is smaller than Grace's (5.4 strokes).

5. (a) A: median 7 days, Q1 = 6, Q3 = 8.5, IQR = 2.5 days. B: median 10 days, Q1 = 8.5, Q3 = 11.5, IQR = 3 days. (b) Brand A's seeds typically sprouted about 3 days sooner (medians 7 versus 10 days), with slightly less spread (IQR 2.5 versus 3 days). Brand A had one slow seed at 15 days, an outlier (above its fence of 12.25); Brand B had no outliers.
6. (a) North: $\frac{110}{10} = 11.0$ inches; South: $\frac{135}{10} = 13.5$ inches. (b) South Stream's trout were typically longer, by about 2.5 inches on average.
7. The median and IQR. The distribution is skewed right: most families drove under 20 miles, but a few drove 40 to 60 miles. Those long drives would pull the mean up and inflate the standard deviation, while the median and IQR describe the typical family.
8. (a) $68.2 - 61.5 = 6.7$ inches. (b) The fertilized sunflowers were typically taller: their mean height was 6.7 inches greater (68.2 versus 61.5 inches). The two groups had very similar spread (standard deviations 4.1 and 4.3 inches), and both distributions were roughly symmetric.
9. (a) $\frac{213}{11} \approx 19.4$ pounds. (b) $\frac{179}{10} = 17.9$ pounds. (c) The single outlier moved the mean by about 1.5 pounds but the median by only 0.5 pound, which is why the median is preferred when there are outliers.
10. He would be comparing two different kinds of center, so any difference might come from the choice of measure rather than from the stations. He should use the same summaries for both; since one distribution is skewed, he should use the median and IQR (five-number summaries) for both, ideally with parallel boxplots.
11. (a) Henry: $\bar{x} = \frac{66.5}{5} = 13.30$; squared deviations 0.04, 0.01, 0.01, 0.04, 0 sum to 0.10; $s = \sqrt{0.025} \approx 0.16$ seconds. Samuel: $\bar{x} = \frac{66.0}{5} = 13.20$; squared deviations 0.36, 0.49, 0.04, 0.64, 0.49 sum to 2.02; $s = \sqrt{0.505} \approx 0.71$ seconds. (b) Henry: his mean is only 0.1 second slower, but his times typically vary by about 0.16 second, compared with 0.71 second for Samuel.
12. For example: (1) the spread of yields on each farm, since a 4-bushel difference means little if yields vary by 15 bushels from year to year; (2) the shapes and any outliers, since one unusual year or field could pull a mean; (3) how many years or fields were measured and whether the conditions (weather, crop) were the same, since a difference in one season might be chance.
13. (a) $\frac{20(16.1) + 30(16.4)}{50} = \frac{322 + 492}{50} = \frac{814}{50} = 16.28$ ounces. (b) No: the combined median depends on how the individual jar weights from the two groups interleave, which the two medians do not tell you.

LESSON 2.6 · 2 SESSIONS OF 45 MINUTES

Transforming Data

OBJECTIVES

- Describe the effect of adding a constant to every value: centers shift and spreads stay the same.
- Describe the effect of multiplying every value by a constant: centers and spreads are multiplied (the standard deviation by |b|).
- Convert summary statistics between units, including Fahrenheit to Celsius with $C = \frac{5}{9}(F - 32)$.
- Explain why the shape of a distribution does not change under a linear change of units.

VIDEO

WATCH

Modifying Statistical Scale & Origin (1 of 2: Summary)

Eddie Woo · 8:30

When to watch: Before Example 1: how adding a constant and multiplying by a constant change center and spread.

https://www.youtube.com/watch?v=DUt0v_BZpFc

BACKUP

Adding Vs. Multiplying Effect on Median and Standard Deviation

Mario's Math Tutoring · 4:16

When to watch: Review: adding versus multiplying, and their effect on the median and standard deviation.

https://www.youtube.com/watch?v=raoTRsdAc3Y

KEY TERMS

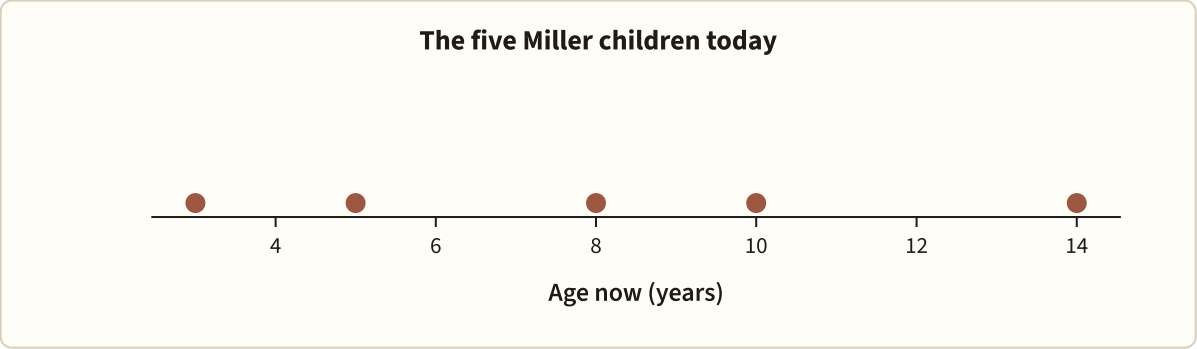
Linear transformation	A change of every value x to a new value $y = a + bx$: multiply by a constant b and then add a constant a. Changes of units, such as feet to inches or Fahrenheit to Celsius, are linear transformations.
Measures of center and position	Numbers that locate values on the number line: the mean, the median, the quartiles, the minimum and the maximum. Adding a constant moves all of them.
Measures of spread	Numbers that measure how far apart the values are: the range, the IQR and the standard deviation. Adding a constant does not change them.

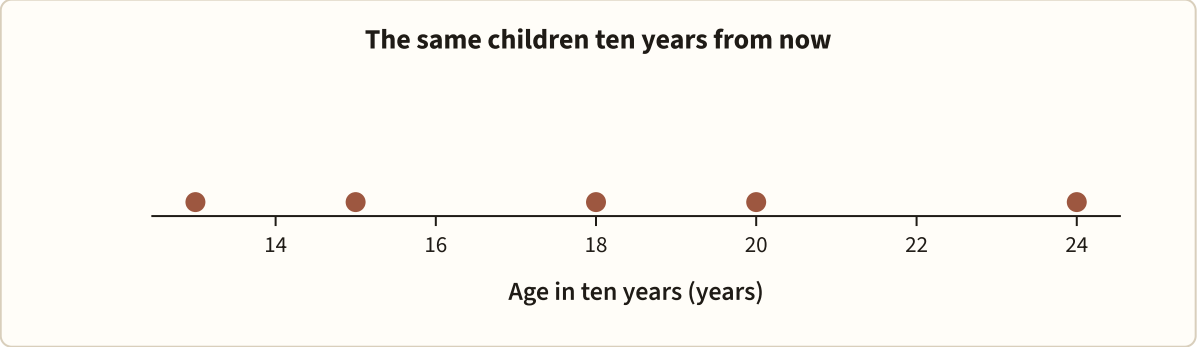
Shift	Adding the same constant to every value. The whole distribution slides along the number line without changing its shape or spread.
Rescale	Multiplying every value by the same constant, as when changing units. The distribution is stretched or shrunk, but its shape stays the same.

CORE IDEAS

Data often need to be changed in a simple way: a scale that reads 2 pounds heavy must be corrected, feet must be changed to inches, Fahrenheit to Celsius, bushels picked to wages earned. Each of these replaces every value x by $y = a + bx$, a **linear transformation**. Rather than recomputing every summary from scratch, you can predict exactly what happens to each one, and it is worth understanding why.

Adding a constant. The five Miller children are 3, 5, 8, 10 and 14 years old (mean 8, median 8, range 11, standard deviation about 4.3 years). In ten years every child will be 10 years older: 13, 15, 18, 20 and 24. The mean and median both become 18, 10 more than before. But the range is still $24 - 13 = 11$ years, and the standard deviation is still about 4.3 years. Why? Every value slid 10 to the right, so the mean slid too, and each child's deviation from the mean is unchanged (the oldest is still 6 years above the mean). The children are no more and no less spread out than before.





Multiplying by a constant. Fence posts measured in feet have mean 6.5 feet and standard deviation 0.25 foot. In inches every length is multiplied by 12, so the mean is $12 \times 6.5 = 78$ inches. The distance between any two posts is also 12 times as many inches as it was feet, so every measure of spread is

multiplied by 12 as well: $s = 12 \times 0.25 = 3$ inches. Here is the algebra for s: multiplying every value by b multiplies every deviation by b, every squared deviation by b^2, and so after taking the square root, s is multiplied by $\sqrt{b^2} = |b|$. (A spread can never be negative, which is why it is |b| and not b.)

What a linear transformation y = a + bx does

Summary	Add a	Multiply by b	Both: y = a + bx
Mean, median, quartiles	Add a	Multiply by b	a + b times old value
Standard deviation, IQR, range	No change	Multiply by |b|	|b| times old value
Shape	No change	No change (b > 0)	No change (b > 0)

Both together. For $y = a + bx$, apply the two rules in order: measures of center and position become $a + b\times$ (old value), and measures of spread become $|b|\times$ (old value); the added a has no effect on spread. The temperature formula $C = \frac{5}{9}(F - 32)$ subtracts 32 and then multiplies by $\frac{5}{9}$. So a mean of 68°F becomes $\frac{5}{9}(68 - 32) = 20$°C, but a standard deviation of 9°F becomes simply $\frac{5}{9} \times 9 = 5$°C. Do NOT subtract 32 from a spread: a 9-degree difference in Fahrenheit is a 5-degree difference in Celsius, wherever on the scale it happens.

Why the shape does not change. A change of units only relabels the number line: adding slides every value the same amount, and multiplying by a positive number stretches every gap by the same factor. Values that were bunched together are still bunched; a long right tail is still a long right tail; an outlier is still an outlier. A histogram of temperatures in Celsius, with matching classes, has exactly the same bars as the histogram in Fahrenheit. (Multiplying by a negative number would also flip the picture over, so a distribution skewed right would become skewed left.)

KEY FORMULAS

$$y = a + bx: \quad \bar{y} = a + b\bar{x}, \quad \text{median}_y = a + b \cdot \text{median}_x$$

$$s_y = |b|\, s_x, \quad \text{IQR}_y = |b| \cdot \text{IQR}_x, \quad \text{range}_y = |b| \cdot \text{range}_x$$

$$C = \frac{5}{9}(F - 32) \qquad F = \frac{9}{5}C + 32$$

Practice habit. Before calculating, ask: "Is this a shift, a rescale, or both?" Centers feel both the added and the multiplied constants; spreads feel only the multiplier. Label every new number with its new units.

WORKED EXAMPLES

Example 1. The table summarizes the daily high temperatures in Fahrenheit for one May. Find the mean, standard deviation, median and IQR in degrees Celsius, using $C = \frac{5}{9}(F - 32)$.

Daily high temperatures in May (degrees F)

Summary	Value
Mean	77
Standard deviation	9
Median	75.2
IQR	10.8

1. Rewrite the formula as $C = -\frac{160}{9} + \frac{5}{9}F$: a shift and a rescale with $b = \frac{5}{9}$.
2. Mean (a center): $\frac{5}{9}(77 - 32) = \frac{5}{9}(45) = 25$°C.
3. Median (a center): $\frac{5}{9}(75.2 - 32) = \frac{5}{9}(43.2) = 24$°C.
4. Standard deviation (a spread): only the multiplier matters, $\frac{5}{9} \times 9 = 5$°C.
5. IQR (a spread): $\frac{5}{9} \times 10.8 = 6$°C.
6. Conclusion: in Celsius the May highs had mean 25°C and median 24°C, and they typically varied from the mean by about 5°C, with the middle half spread over 6°C.

Answer: Mean 25°C, standard deviation 5°C, median 24°C, IQR 6°C.

Example 2. An orchard pays each picker \$12 a day plus \$3 for each bushel of apples picked. The table summarizes the number of bushels the crew members pick per day. Find the mean, standard deviation, median and IQR of the daily pay, and explain why the \$12 does not affect the spread.

Bushels of apples picked per day by the orchard crew

Summary	Bushels
Mean	20
Standard deviation	4
Median	19
IQR	6

1. Pay $= 12 + 3x$ dollars, where x is the number of bushels: $a = 12$ and $b = 3$.
2. Mean pay: $12 + 3(20) = 12 + 60 = 72$ dollars.
3. Median pay: $12 + 3(19) = 12 + 57 = 69$ dollars.
4. Standard deviation of pay: $|3| \times 4 = 12$ dollars.
5. IQR of pay: $3 \times 6 = 18$ dollars.
6. Every picker receives the same \$12, so it raises everyone's pay equally; it moves the center but does not change how far apart any two pickers' pay is.
7. Conclusion: mean daily pay \$72 (median \$69), and pay typically varies from the mean by about \$12.

Answer: Mean \$72, standard deviation \$12, median \$69, IQR \$18. The flat \$12 is added to everyone's pay, so it shifts the center but leaves the differences between pickers, and therefore the spread, unchanged.

PRACTICE

1. EASY A farm scale reads 2 pounds too heavy. The recorded weights of the spring calves have mean 84 pounds and standard deviation 6 pounds. Find the mean and standard deviation of the correct weights.
2. EASY Boards at a lumberyard have mean length 8 feet and standard deviation 0.2 foot. Find the mean and standard deviation in inches.
3. EASY Hay bales have Q1 = 50 pounds and Q3 = 58 pounds. After a light rain, every bale weighs 3 pounds more. What is the new IQR?

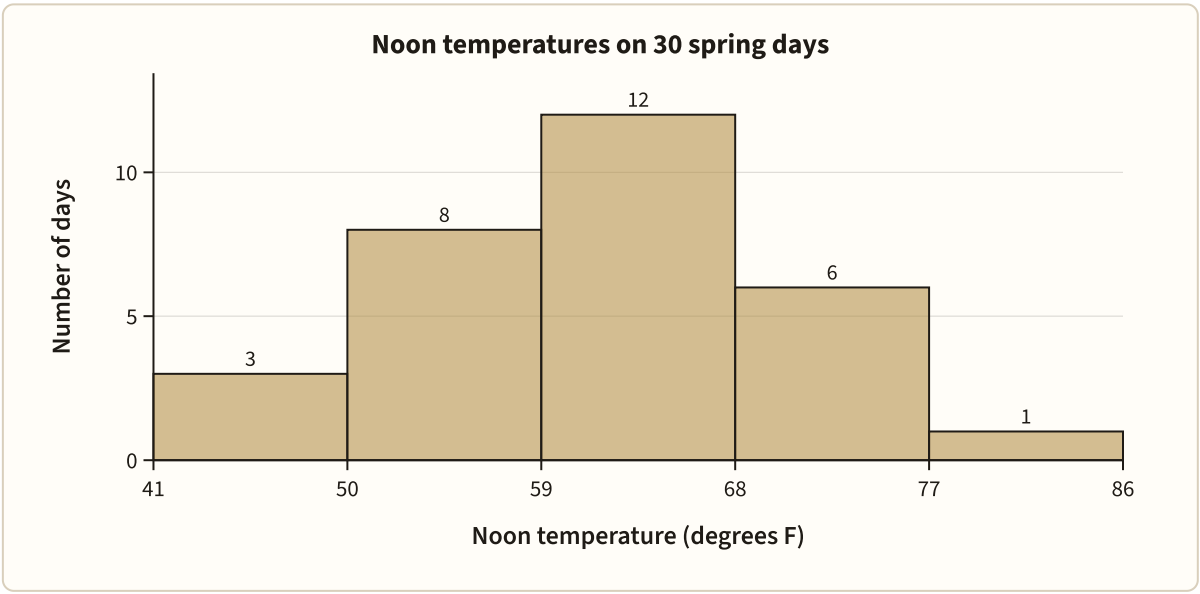
4. EASY The median monthly rainfall in a town is 2.5 inches. One inch is 2.54 centimeters. What is the median in centimeters?

5. MEDIUM Winter temperatures in a mountain town have mean 50°F and standard deviation 18°F. Find the mean and standard deviation in degrees Celsius.

6. MEDIUM Night temperatures at a campground have mean 15°C and standard deviation 4°C. Using $F = \frac{9}{5}C + 32$, find the mean and standard deviation in degrees Fahrenheit.

7. MEDIUM At a weekly bake sale, pies sell for \$8 each and the supplies cost \$20 a week, so the profit is $8x - 20$ dollars, where x is the number of pies sold. The number sold per week has mean 14 and standard deviation 3. Find the mean and standard deviation of the weekly profit.

8. MEDIUM The histogram shows noon temperatures on 30 spring days in degrees Fahrenheit. (a) Convert the class edges 41, 50, 59, 68, 77, 86 to Celsius. (b) Explain why a histogram of the same temperatures in Celsius would have exactly the same shape.



9. MEDIUM On a golf course with par 72, Henry's scores have mean 80.5 strokes and standard deviation 4.2 strokes. His score "relative to par" is his score minus 72. Find the mean and standard deviation of his scores relative to par.

10. HARD Jars of honey are meant to hold 16 ounces. The actual weights have mean 15.8 ounces and standard deviation 0.3 ounce. The shortfall of a jar is $16 - x$ ounces. Find the mean and standard deviation of the shortfall.

11. HARD At a county fair, one judge's scores for 4-H projects have mean 60 and standard deviation 8, while the other judges' scores have mean 80 and standard deviation 10. Find a and b so that the new scores $y = a + bx$ have mean 80 and standard deviation 10.

12. HARD Ruth says, "If I add 10 to every value, the standard deviation must go up, because all the numbers got bigger." Explain why she is wrong.

13. **CHALLENGE** Carrots from a garden have lengths with standard deviation 3 inches (variance 9 square inches). (a) Find the standard deviation in centimeters (1 inch = 2.54 cm). (b) Find the variance in square centimeters, rounded to two decimal places. (c) By what factor did the variance change?

COMMON MISTAKES

- Adding the constant to the standard deviation or IQR. Adding shifts only centers and positions; spreads stay the same.
- Subtracting 32 when converting a standard deviation from Fahrenheit to Celsius. For a spread, only multiply by $\frac{5}{9}$.
- Giving a negative standard deviation when b is negative. Spreads are multiplied by |b|.
- Forgetting that the variance is multiplied by b^2, not by b. If you work with variances, square the multiplier.

ANSWERS TO PRACTICE

1. Subtract 2 from every value: mean $84 - 2 = 82$ pounds; the standard deviation stays 6 pounds.
2. Multiply by 12: mean 96 inches, standard deviation $12 \times 0.2 = 2.4$ inches.
3. Adding 3 pounds shifts both quartiles (53 and 61) but not their difference: IQR = 8 pounds.
4. $2.5 \times 2.54 = 6.35$ centimeters.
5. Mean: $\frac{5}{9}(50 - 32) = \frac{5}{9}(18) = 10$°C. Standard deviation: $\frac{5}{9} \times 18 = 10$°C.
6. Mean: $\frac{9}{5}(15) + 32 = 27 + 32 = 59$°F. Standard deviation: $\frac{9}{5} \times 4 = 7.2$°F.
7. Mean: $8(14) - 20 = 112 - 20 = 92$ dollars. Standard deviation: $8 \times 3 = 24$ dollars.
8. (a) 5, 10, 15, 20, 25, 30°C (for example, $\frac{5}{9}(41 - 32) = 5$). (b) Every day stays in the same class, because the conversion keeps the order of the values and stretches all gaps by the same factor $\frac{5}{9}$. So the bars have the same counts (3, 8, 12, 6, 1) and the same shape; only the labels on the axis change.
9. Mean $80.5 - 72 = 8.5$ strokes over par; standard deviation still 4.2 strokes.
10. Here $a = 16$ and $b = -1$. Mean shortfall $= 16 - 15.8 = 0.2$ ounce. Standard deviation $= |-1| \times 0.3 = 0.3$ ounce (not −0.3: a spread is never negative).
11. Spread: $b \times 8 = 10$, so $b = 1.25$. Center: $a + 1.25(60) = 80$, so $a = 80 - 75 = 5$. Use $y = 5 + 1.25x$.
12. The standard deviation measures how far the values are from their mean. Adding 10 to every value also adds 10 to the mean, so every deviation $x - \bar{x}$ stays exactly the same. The values are bigger, but no more spread out, so s does not change.

13. (a) $2.54 \times 3 = 7.62$ cm. (b) $7.62^2 = 58.0644 \approx 58.06$ square cm. (c) The variance is multiplied by $2.54^2 = 6.4516$, the square of the multiplier, because variance is in squared units.

Unit 2 test

Time: 60 minutes. You may use a scientific calculator with one-variable statistics; no notes. Show your work for every calculation, label every answer with its units, and write interpretations in complete sentences in context. Use the course methods: quartiles as medians of the lower and upper halves (leaving out the median when n is odd), the 1.5 × IQR rule for outliers, and n − 1 for a sample standard deviation. Round as each problem states; otherwise give means and standard deviations to one more decimal place than the data.

1. **7 PTS** A weather station recorded the rainfall (inches) from six spring storms: 0.4, 1.2, 0.8, 2.5, 0.6, 0.9. (a) Find the mean (two decimal places). (b) Find the median. (c) Explain why the mean is larger than the median.

2. **6 PTS** Grace bought 300 pounds of chicken feed at \$0.40 per pound and later 200 pounds at \$0.55 per pound. What was her average price per pound for all the feed?

3. **5 PTS** Nine gardens in a community plot produced a mean of 42 tomatoes each. A tenth garden produced 62 tomatoes. What is the mean for all ten gardens?

4. **8 PTS** The lengths (centimeters) of nine ears of sweet corn were 17, 19, 20, 20, 21, 22, 23, 25, 26. Find the five-number summary and the IQR.

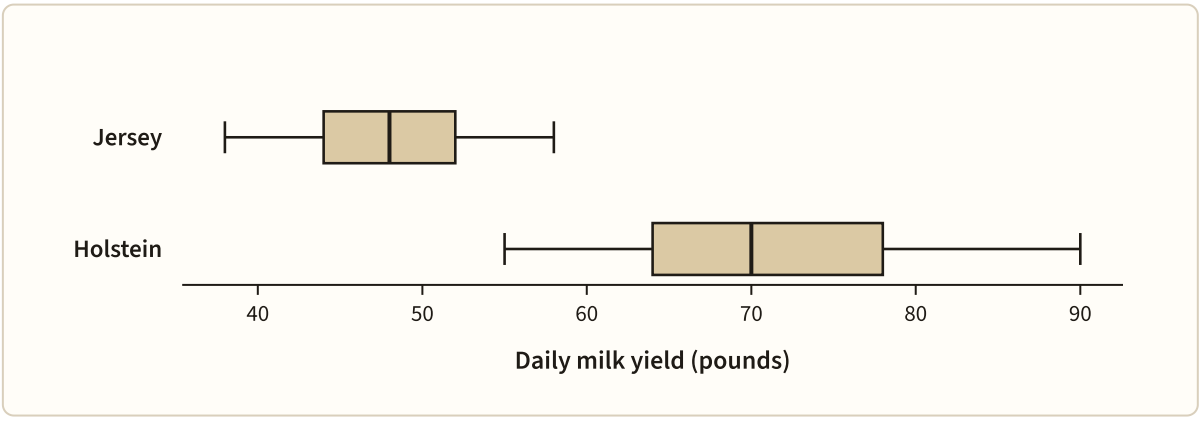
5. **9 PTS** The stemplot shows the number of visitors to a parish library on 12 Saturdays. (a) Find the fences for the 1.5 × IQR rule. (b) Identify any outliers. (c) Where would the whiskers of a modified boxplot end?

Parish library visitors on 12 Saturdays

1	4 8 9
2	1 2 3 4 5 6 8
3	0
4	7

Key: 2 | 4 means 24 visitors

6. **7 PTS** The parallel boxplots compare the daily milk yields of the Jersey and Holstein cows on a dairy farm. (a) How much larger is the Holstein median? (b) Find each breed's IQR. (c) Write one sentence comparing the spreads in context.



7. **4 PTS** A boxplot of the weights of 30 apples at a farm stand has its median at 150 grams. Samuel concludes, "Lots of the apples must weigh close to 150 grams." Explain why the boxplot alone cannot support his conclusion, and name a graph that could.

8. **8 PTS** Five tomato plants produced 26, 29, 32, 34 and 39 tomatoes. Find the mean and the standard deviation, showing a table of deviations. Round the standard deviation to one decimal place.

9. **5 PTS** The sunflowers in a 4-H plot have a mean height of 64 inches and a standard deviation of 5 inches. Interpret the standard deviation in context.

10. **6 PTS** A family keeps exactly four goats, and these four are the only goats of interest. Their daily milk yields are 5, 7, 8 and 12 pounds. (a) Find the standard deviation σ of this population (one decimal place). (b) Explain why σ, dividing by n, is the right choice here rather than s.

11. **6 PTS** The histogram shows the finishing times of 40 runners in a community 5K race. (a) Describe the shape. (b) Which measures of center and spread should the race organizers report? Explain.

12. **8 PTS** A gardener grew 20 plants of each of two potato varieties. The table summarizes the yield per plant. (a) How much greater is the Yukon Gold mean? (b) Write a short paragraph comparing the two varieties in context.

Potato yield per plant (pounds)

Variety	Number of plants	Mean	Standard deviation	Shape
Red Pontiac	20	4.6	0.7	Roughly symmetric
Yukon Gold	20	5.3	1.4	Roughly symmetric

13. **8 PTS** The table summarizes the daily high temperatures for one June in degrees Celsius. Using $F = \frac{9}{5}C + 32$, find the mean, standard deviation, median and IQR in degrees Fahrenheit.

Daily high temperatures in June (degrees C)

Summary	Value
Mean	20
Standard deviation	5
Median	18
IQR	7.5

14. **6 PTS** A church youth group runs a car wash. They charge \$5 per car and spend \$15 on supplies each Saturday, so the profit is $5x - 15$ dollars for x cars. The number of cars per Saturday has mean 30 and standard deviation 6. Find the mean and standard deviation of the profit.

15. **7 PTS** Daily high temperatures in a town have Q1 = 10°C and Q3 = 18°C. (a) Find the fences for the 1.5 × IQR rule in Celsius. (b) The temperatures are converted to Fahrenheit with $F = 1.8C + 32$. Find the new fences. (c) A day with a high of 31°C was an outlier. Is it still an outlier after the conversion? Answer yes or no.

UNIT 2 TEST: ANSWER KEY

1. (a) $\bar{x} = \dfrac{6.4}{6} \approx 1.07$ inches. (b) Sorted: 0.4, 0.6, 0.8, 0.9, 1.2, 2.5; median $= \dfrac{0.8 + 0.9}{2} = 0.85$ inch. (c) The one large 2.5-inch storm pulls the mean up toward it; the median depends only on the middle values and resists it.
2. Total cost $= 300(0.40) + 200(0.55) = 120 + 110 = 230$ dollars for 500 pounds; average $= \dfrac{230}{500} = \$0.46$ per pound.
3. $\dfrac{9(42) + 62}{10} = \dfrac{378 + 62}{10} = \dfrac{440}{10} = 44$ tomatoes.
4. Median = 5th value = 21. Lower half 17, 19, 20, 20 gives Q1 $= \dfrac{19 + 20}{2} = 19.5$; upper half 22, 23, 25, 26 gives Q3 $= \dfrac{23 + 25}{2} = 24$. Summary: 17, 19.5, 21, 24, 26 cm; IQR $= 24 - 19.5 = 4.5$ cm.
5. (a) Median $= \dfrac{23 + 24}{2} = 23.5$. Lower half 14, 18, 19, 21, 22, 23 gives Q1 = 20; upper half 24, 25, 26, 28, 30, 47 gives Q3 = 27. IQR = 7, $1.5 \times 7 = 10.5$; fences $20 - 10.5 = 9.5$ and $27 + 10.5 = 37.5$. (b) 47 visitors is an outlier. (c) Whiskers end at 14 and 30.
6. (a) $70 - 48 = 22$ pounds. (b) Jersey: $52 - 44 = 8$ pounds; Holstein: $78 - 64 = 14$ pounds. (c) The Holsteins' daily yields are more variable: the middle half of the Holsteins spread over 14 pounds, compared with 8 pounds for the Jerseys.
7. A boxplot shows only five numbers; it hides clusters and gaps. The apples might form two clusters (say, small and large varieties) with few or none near 150 grams. A dotplot, stemplot or histogram of the 30 weights would show whether many apples are near 150 grams.
8. $\bar{x} = \dfrac{160}{5} = 32$. Deviations −6, −3, 0, 2, 7 (sum 0); squares 36, 9, 0, 4, 49, sum 98. $s^2 = \dfrac{98}{4} = 24.5$, $s = \sqrt{24.5} \approx 4.9$ tomatoes.
9. The heights of the sunflowers typically differ from the mean height of 64 inches by about 5 inches.
10. (a) Mean 8; deviations −3, −1, 0, 4; squares sum to 26; $\sigma = \sqrt{\dfrac{26}{4}} = \sqrt{6.5} \approx 2.5$ pounds. (b) The four goats are the whole population, so the mean is the true mean μ and nothing is being estimated; we divide by n − 1 only when a sample's deviations from $\bar{x}$ must stand in for deviations from an unknown μ.
11. (a) Skewed right: most runners finished between 24 and 36 minutes, with a tail of slower runners out to 54 minutes. (b) The median and IQR, because the long right tail would pull the mean up and inflate the standard deviation, while the median and IQR describe the typical runner.

12. (a) $5.3 - 4.6 = 0.7$ pound. (b) Both distributions are roughly symmetric, so the means and standard deviations are suitable. Yukon Gold plants yielded more on average, about 0.7 pound more per plant (5.3 versus 4.6 pounds). But Yukon Gold yields were much less consistent: their standard deviation (1.4 pounds) is twice that of Red Pontiac (0.7 pound).
13. Mean: $\frac{9}{5}(20) + 32 = 68$°F. Standard deviation: $\frac{9}{5} \times 5 = 9$°F. Median: $\frac{9}{5}(18) + 32 = 64.4$°F. IQR: $\frac{9}{5} \times 7.5 = 13.5$°F. (The 32 is added only to the centers.)
14. Mean: $5(30) - 15 = 150 - 15 = 135$ dollars. Standard deviation: $5 \times 6 = 30$ dollars (the \$15 does not affect spread).
15. (a) IQR = 8, $1.5 \times 8 = 12$; fences $10 - 12 = -2$°C and $18 + 12 = 30$°C. (b) Q1 = 50°F, Q3 = 64.4°F, IQR = 14.4°F, $1.5 \times 14.4 = 21.6$; fences 28.4°F and 86°F (the same as converting −2 and 30). (c) Yes: 31°C is $1.8(31) + 32 = 87.8$°F, which is above 86°F. A change of units does not change which values are outliers.

UNIT 3

Position and the Normal Distribution

In Unit 2 you learned to describe a whole distribution with numbers for its center and spread. This unit asks where a single value stands within its distribution. Percentiles and z-scores measure position, and z-scores let you compare values measured in different units, such as a sprint time and a long jump, by measuring both in standard deviations.

Next you will replace jagged histograms with smooth density curves, where area stands for proportion. The most important of these is the normal curve, the bell shape followed by egg weights, crop heights, measurement errors and many other measurements. You will use the 68–95–99.7 rule, the standard normal table and technology to find proportions and cutoff values, and you will learn to check whether a normal model really fits a data set before you use it.

The normal distribution is the foundation for much of the rest of the course: the normal approximation to the binomial (Unit 8), sampling distributions (Unit 9), and the confidence intervals and significance tests of Units 10–12 all rest on the skills you build here.

PREREQUISITES

- Compute the mean and median of a data set (Lesson 2.1).
- Compute and interpret the standard deviation s, dividing by n − 1 (Lesson 2.4).
- Describe the shape of a distribution as symmetric, skewed left or skewed right (Lesson 1.5).
- Explain how adding or multiplying by a constant changes center and spread (Lesson 2.6).
- Find areas of rectangles and solve a linear equation for an unknown, including a system of two equations (Algebra I).

MASTERY CHECKLIST

☐ I can find the percentile of a value as the percent of values less than it, and read percentiles from a cumulative relative frequency table.

☐ I can compute a z-score, explain it as the number of standard deviations above or below the mean, and use z-scores to compare values from different distributions.

☐ I can explain why area under a density curve is a proportion, and find areas for a uniform distribution.

☐ I can locate the mean (balance point) and median (equal-areas point) of a density curve and explain why the mean is pulled toward a long tail.

☐ I can state the 68–95–99.7 rule and use it, with a sketch, to find percents and values for a normal distribution.

☐ I can standardize a normal value and use the standard normal table or technology to find areas to the left, to the right and between two values.
☐ I can explain why a single exact value has probability 0 for a continuous distribution.
☐ I can work backward from a proportion to a cutoff value with $x = \mu + z\sigma$, and find μ or σ when the other is known.
☐ I can decide whether data are roughly normal using a graph, the 68–95–99.7 rule and a normal probability plot.
☐ I can state every answer in context with its units.

LESSON 3.1 · 2 SESSIONS OF 45 MINUTES

Percentiles and z-Scores

OBJECTIVES

- Find and interpret the percentile of a value, using the course definition (the percent of values less than it).
- Compute a z-score and explain it as the number of standard deviations a value lies above or below the mean.
- Use z-scores to compare values that come from different distributions, and explain why the comparison is fair.
- Read percentiles from a cumulative relative frequency table or graph.

VIDEO

WATCH

Z-Scores (As a Descriptive Measure of Relative Standing)

jbstatistics · 8:07

When to watch: Before Example 2: computing and interpreting a z-score.

https://www.youtube.com/watch?v=EhUvGRddC4M

BACKUP

Deciles and Percentiles (1 of 2: Outline of various ways of breaking data up into equal groups)

Eddie Woo · 11:25

When to watch: Before Example 1: percentiles and quartiles as positions in a distribution.

https://www.youtube.com/watch?v=0DmgFtfy57Y

KEY TERMS

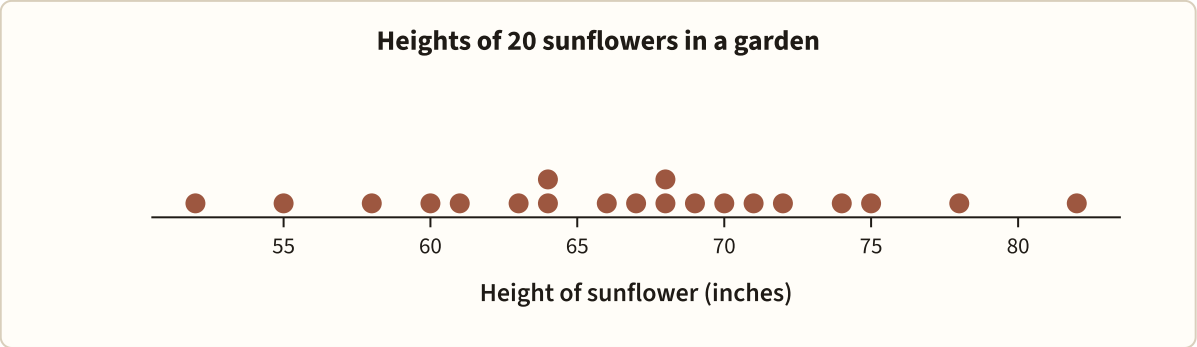
Percentile	The percentile of a value is the percent of values in the data set that are less than that value. A value at the 80th percentile is greater than 80% of the data.
z-score	The number of standard deviations a value lies above (positive) or below (negative) the mean: $z = \dfrac{x - \text{mean}}{\text{standard deviation}}$.
Standardizing	Turning a value into its z-score, so that it is measured in standard deviations from the mean instead of in its original units.

Cumulative relative frequency	For a class in a frequency table, the proportion (or percent) of all values that fall in that class or in any class below it.
Cumulative relative frequency graph	A graph that plots cumulative percent against the right endpoint of each class, so you can read the percent of values less than any number on the axis.

CORE IDEAS

In Unit 2 you described a whole distribution with a center and a spread. Now we ask a different question: where does ONE value stand in its distribution? Is a 64-gram egg large for its flock? Is a 12.6-second dash fast for the league? Statistics gives two answers: the percentile and the z-score.

A **percentile** tells what percent of the data lie below a value. In this course the percentile of a value is always the percent of values that are LESS THAN it. In the dotplot of 20 sunflower heights below, 6 sunflowers are shorter than 64 inches, so a 64-inch sunflower is at the 30th percentile, since $\frac{6}{20} \times 100\% = 30\%$. Notice that a percentile is not a test score: being at the 30th percentile means 30% of the group is below you, not that you got 30% of anything right.



A percentile uses only order, so it ignores HOW FAR a value is from the others. A **z-score** measures distance, using the standard deviation as the ruler: $z = \frac{x - \bar{x}}{s}$. If a flock's eggs have mean 58 grams and standard deviation 4 grams, a 64-gram egg has $z = \frac{64 - 58}{4} = 1.50$: it is 1.50 standard deviations above the mean. A negative z-score means below the mean, and $z = 0$ means exactly at the mean.

Why do z-scores let us compare things that are measured in different units, like seconds in a sprint and meters in a long jump? In the fraction $\frac{x - \bar{x}}{s}$ the top and the bottom have the same units, so the units cancel. What remains is a pure number that says how unusual the value is compared with the typical spread of its own group. A jump that is 1.50 standard deviations above the jumpers' mean and a dash that is 1.50 standard deviations faster than the runners' mean are equally outstanding, even though one is in meters and the other in seconds.

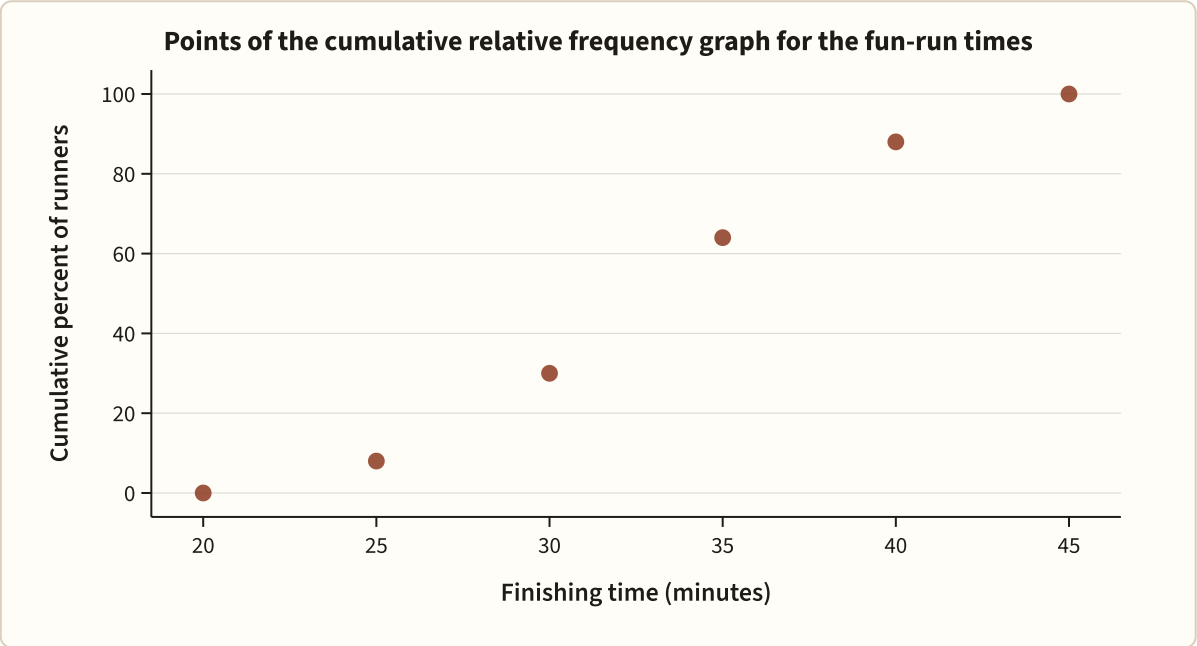
Standardizing is a transformation like those in Lesson 2.6. Subtracting the mean shifts the whole data set so its center is 0, and dividing by s shrinks (or stretches) the spread so the standard deviation

becomes 1. So a set of z-scores always has mean 0 and standard deviation 1, and its shape is the same as the shape of the original data. That is why z-scores are the common language for comparing distributions.

When data come grouped in classes, a **cumulative relative frequency** table adds up the relative frequencies from the bottom class upward. Each cumulative value is the proportion of data less than the right edge of its class, which is exactly a percentile. In the table of 50 fun-run times, 0.64 of the runners finished in less than 35 minutes, so 35 minutes is the 64th percentile. A cumulative relative frequency graph plots these points (time on the horizontal axis, cumulative percent on the vertical axis) and joins them with line segments; the graph always rises from 0% to 100%, and it is steepest where the data are most crowded.

Finishing times of 50 runners in a 5-km community fun run

Time (minutes)	Frequency	Relative frequency	Cumulative relative frequency
20 to less than 25	4	0.08	0.08
25 to less than 30	11	0.22	0.3
30 to less than 35	17	0.34	0.64
35 to less than 40	12	0.24	0.88
40 to less than 45	6	0.12	1
Total	50	1	



KEY FORMULAS

$$\text{percentile of } x = \frac{\text{number of values less than } x}{n} \times 100\%$$

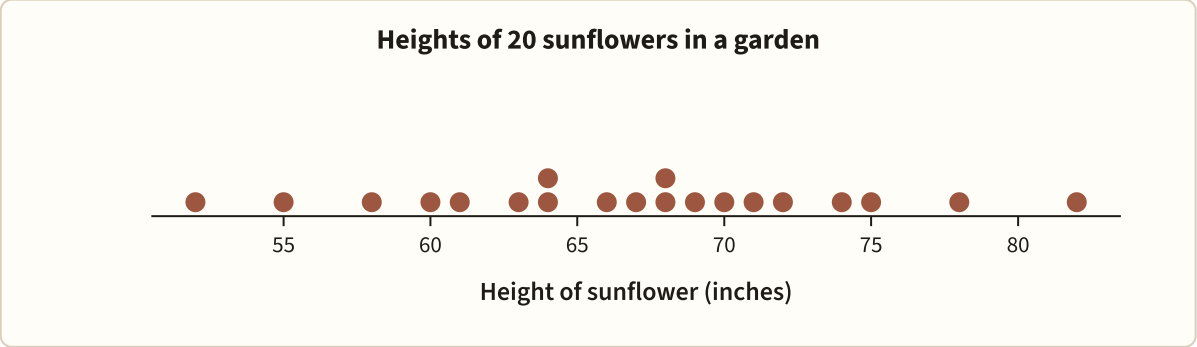
$$z = \frac{x - \bar{x}}{s} \quad \text{or, for a population,} \quad z = \frac{x - \mu}{\sigma}$$

$$x = \bar{x} + z \cdot s \quad \text{(working backward from a z-score)}$$

Practice habit. Always say a z-score in words with its direction: "1.50 standard deviations ABOVE the mean" or "0.75 standard deviations BELOW the mean." Saying it aloud keeps you from losing the sign.

WORKED EXAMPLES

Example 1. The dotplot shows the heights of 20 sunflowers in Grace's garden. (a) What percentile is a sunflower that is 72 inches tall? (b) Which of the heights shown in the dotplot is at the 50th percentile?



1. Percentile means the percent of values LESS THAN the value. Count the sunflowers shorter than 72 inches: 52, 55, 58, 60, 61, 63, 64, 64, 66, 67, 68, 68, 69, 70, 71. That is 15 sunflowers.
2. (a) Percentile $= \frac{15}{20} \times 100\% = 75\%$, so 72 inches is the 75th percentile.
3. (b) The 50th percentile needs $0.50 \times 20 = 10$ values below it.
4. List the heights in order: the 10th height is 67 and the 11th is 68. Exactly 10 sunflowers (52 through 67) are shorter than 68.
5. So 68 inches is at the 50th percentile.
6. Conclusion: a 72-inch sunflower is taller than 75% of the sunflowers in Grace's garden, and a 68-inch sunflower is taller than half of them.

Answer: (a) 75th percentile; (b) 68 inches

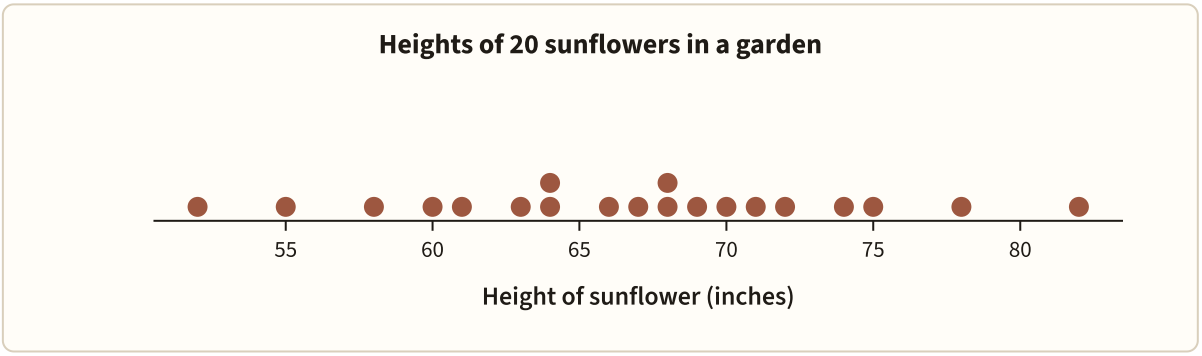
Example 2. At a county track meet, Ruth ran the 100-meter dash in 12.6 seconds. The girls in the league have a mean time of 13.4 seconds with standard deviation 0.5 second. Henry long-jumped 5.9 meters; the boys' long jumps have mean 5.3 meters and standard deviation 0.4 meter. Find each z-score and decide whose performance was more outstanding compared with the league.

1. Ruth: $z = \dfrac{x - \bar{x}}{s} = \dfrac{12.6 - 13.4}{0.5} = \dfrac{-0.8}{0.5} = -1.60.$
2. Henry: $z = \dfrac{5.9 - 5.3}{0.4} = \dfrac{0.6}{0.4} = 1.50.$
3. Be careful with direction. In a race, a LOWER time is better, so Ruth's negative z-score is good: she ran 1.60 standard deviations faster than the league mean.
4. Henry jumped 1.50 standard deviations farther than the league mean.
5. Compare the sizes: 1.60 standard deviations is more than 1.50 standard deviations.
6. Conclusion: both performances were excellent, but Ruth's dash was slightly more outstanding relative to her league, because she was 1.60 standard deviations better than average and Henry was 1.50.

Answer: Ruth: $z = -1.60$; Henry: $z = 1.50$. Ruth's dash was more outstanding (1.60 standard deviations better than average, since lower times are better).

PRACTICE

1. EASY Use the dotplot of 20 sunflower heights. What percentile is a sunflower that is 70 inches tall?



2. EASY The eggs from the Miller family's flock have mean weight 58 grams and standard deviation 4 grams. Find the z-score of an egg that weighs 64 grams.
3. EASY Calves born on a ranch this spring have mean birth weight 80 pounds and standard deviation 8 pounds. Find the z-score of a calf that weighed 74 pounds and say what it means.
4. EASY Henry's batting average has a z-score of 1.20 among all players in his summer league. Explain in one sentence what this tells you.

5. MEDIUM Use the table of 50 fun-run times. (a) What percentile is a finishing time of 30 minutes? (b) What percentile is a finishing time of 40 minutes?

Finishing times of 50 runners in a 5-km community fun run

Time (minutes)	Frequency	Relative frequency	Cumulative relative frequency
20 to less than 25	4	0.08	0.08
25 to less than 30	11	0.22	0.3
30 to less than 35	17	0.34	0.64
35 to less than 40	12	0.24	0.88
40 to less than 45	6	0.12	1
Total	50	1	

6. MEDIUM A weather station recorded July rainfall for 10 years, in inches: 2.1, 3.4, 2.8, 4.0, 1.9, 3.1, 2.5, 3.7, 2.2, 3.0. What percentile is the year with 3.1 inches?

7. MEDIUM Samuel's tallest corn stalk in Field A is 84 inches; Field A's stalks have mean 78 inches and standard deviation 4 inches. Grace's tallest stalk in Field B is 90 inches; Field B's stalks have mean 86 inches and standard deviation 5 inches. Find both z-scores. Which stalk is taller compared with its own field?

8. MEDIUM A sawmill cuts boards with mean length 96.0 inches and standard deviation 0.25 inch. One board has a z-score of -1.60. How long is it?

9. MEDIUM Five lambs weigh 40, 44, 46, 48 and 52 pounds. Find the mean and standard deviation, then the z-score of the 52-pound lamb. Keep full precision and round z to 2 decimal places.

10. HARD On a dairy farm, the Jersey cows give a mean of 40 pounds of milk per day with standard deviation 8 pounds, and the Holstein cows give a mean of 70 pounds with standard deviation 5 pounds. Daisy, a Jersey, gave 52 pounds; Bess, a Holstein, gave 80 pounds. Find both z-scores and decide which cow is more unusual for her breed.

11. HARD In a golf club's records, a drive of 265 yards has a z-score of 1.25. The mean drive is 240 yards. What is the standard deviation of the drives?

12. HARD Every value in a data set is turned into a z-score. (a) What is the mean of the z-scores? (b) What is their standard deviation? (c) Explain why, using what you know about transforming data.

13. CHALLENGE A youth group of 25 students took a Bible-knowledge quiz, and no two scores were equal. Samuel's score is at the 80th percentile. (a) How many students scored lower than Samuel? (b) Using the course definition, can any student be at the 100th percentile? Explain.

14. **CHALLENGE** July high temperatures in a town have mean 86°F and standard deviation 4°F. (a) Find the z-score of a 92°F day. (b) Convert the mean, the standard deviation and 92°F to Celsius using $C = \frac{5}{9}(F - 32)$ (a standard deviation is only multiplied by $\frac{5}{9}$), and find the z-score again. Round to 2 decimal places. (c) Explain why the two z-scores agree.

COMMON MISTAKES

- Confusing a percentile with a percent score. Being at the 90th percentile means 90% of the values are lower, not that you got 90% correct. Always say "greater than __% of the values."
- Forgetting the sign of z. Subtract in the order $x - \bar{x}$; a value below the mean must give a negative z-score.
- Comparing raw values from different groups (84 inches versus 90 inches in two fields). Standardize first, then compare the z-scores.
- Assuming a larger z is always better. For race times or golf scores, lower is better, so a strongly negative z-score is the outstanding one.

ANSWERS TO PRACTICE

1. 13 sunflowers are shorter than 70 inches; $\frac{13}{20} = 65\%$, so the 65th percentile.
2. $z = \frac{64 - 58}{4} = 1.50$; the egg is 1.50 standard deviations above the mean.
3. $z = \frac{74 - 80}{8} = -0.75$; the calf weighed 0.75 standard deviation below the mean birth weight.
4. Henry's batting average is 1.20 standard deviations above the mean batting average of the players in his league, so he hits better than a typical player in the league.
5. (a) The cumulative relative frequency up to 30 minutes is 0.30, so 30 minutes is the 30th percentile. (b) Up to 40 minutes it is 0.88, so the 88th percentile.
6. In order: 1.9, 2.1, 2.2, 2.5, 2.8, 3.0, 3.1, 3.4, 3.7, 4.0. Six values are less than 3.1, so $\frac{6}{10} = 60\%$: the 60th percentile.
7. Field A: $z = \frac{84 - 78}{4} = 1.50$. Field B: $z = \frac{90 - 86}{5} = 0.80$. Samuel's stalk is more exceptional for its field, even though Grace's stalk is taller in inches.
8. $x = \bar{x} + z \cdot s = 96.0 + (-1.60)(0.25) = 96.0 - 0.4 = 95.6$ inches.
9. $\bar{x} = \frac{230}{5} = 46$; squared deviations 36, 4, 0, 4, 36 add to 80; $s = \sqrt{\frac{80}{4}} = \sqrt{20} \approx 4.472$; $z = \frac{52 - 46}{4.472} \approx 1.34$.
10. Daisy: $z = \frac{52 - 40}{8} = 1.50$. Bess: $z = \frac{80 - 70}{5} = 2.00$. Bess is more unusual for her breed: 2.00 standard deviations above the Holstein mean.
11. $1.25 = \frac{265 - 240}{s}$, so $s = \frac{25}{1.25} = 20$ yards.

12. (a) 0. (b) 1. (c) Subtracting $\bar{x}$ from every value shifts the center from $\bar{x}$ to 0 and leaves the spread alone; dividing every result by s divides the standard deviation by s, so it becomes $\frac{s}{s} = 1$.
13. (a) $0.80 \times 25 = 20$ students. (b) No. Even the top scorer is not less than himself, so only 24 of the 25 scores are below the top score: $\frac{24}{25} = 96\%$, the 96th percentile. The 100th percentile would need every value, including itself, to be below it.
14. (a) $z = \frac{92 - 86}{4} = 1.50$. (b) Mean $\frac{5}{9}(54) = 30$°C, SD $\frac{5}{9}(4) \approx 2.222$°C, day $\frac{5}{9}(60) \approx 33.333$°C; $z = \frac{3.333}{2.222} = 1.50$. (c) A change of units moves and stretches the values and the mean together and stretches the standard deviation by the same factor, so the distance from the mean measured in standard deviations does not change.

LESSON 3.2 · 2 SESSIONS OF 45 MINUTES

Density Curves

OBJECTIVES

- Explain how a density curve is an idealized histogram, and why the area under it over an interval equals a proportion.
- Compute proportions for a uniform distribution using areas of rectangles.
- Locate the median (equal-areas point) and the mean (balance point) of a density curve.
- Explain why the mean is pulled toward the long tail of a skewed density curve while the median is not.

VIDEO

WATCH

Density Curves, Empirical Rule & Normality, Z-score Intro

ProfRobBob · 14:26

When to watch: Before Example 1: density curves, area as proportion, and where the mean and median sit.

https://www.youtube.com/watch?v=bo5STaOUOgI

BACKUP

An Introduction to Continuous Probability Distributions

jbstatistics · 5:51

When to watch: Review: area under a density curve as probability.

https://www.youtube.com/watch?v=OWSOhpS00_s

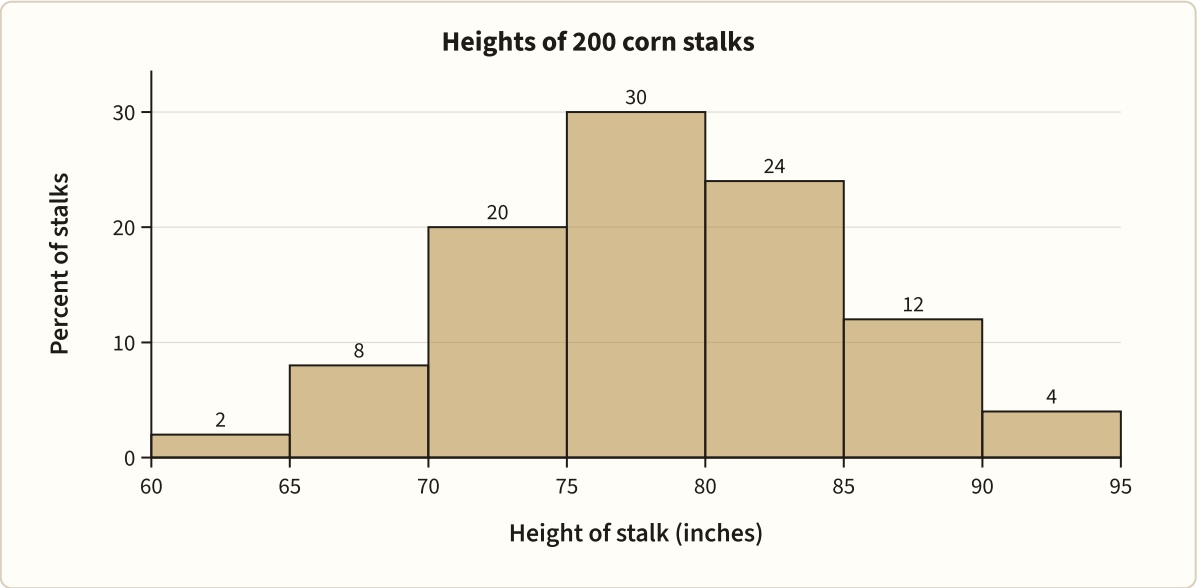
KEY TERMS

Density curve	A curve that is always on or above the horizontal axis and has total area exactly 1 underneath it. The area under the curve above any interval is the proportion of values in that interval.
Uniform distribution	A distribution whose density curve is flat (a rectangle): every interval of the same width has the same proportion of values.
Median of a density curve	The equal-areas point: the value with half the area under the curve to its left and half to its right.
Mean of a density curve	The balance point: the place where the curve would balance if it were cut out of solid material. It is written μ.

μ and σ — The Greek letters mu and sigma name the mean and standard deviation of a density curve (an idealized population), to tell them apart from $\bar{x}$ and s of actual data.

CORE IDEAS

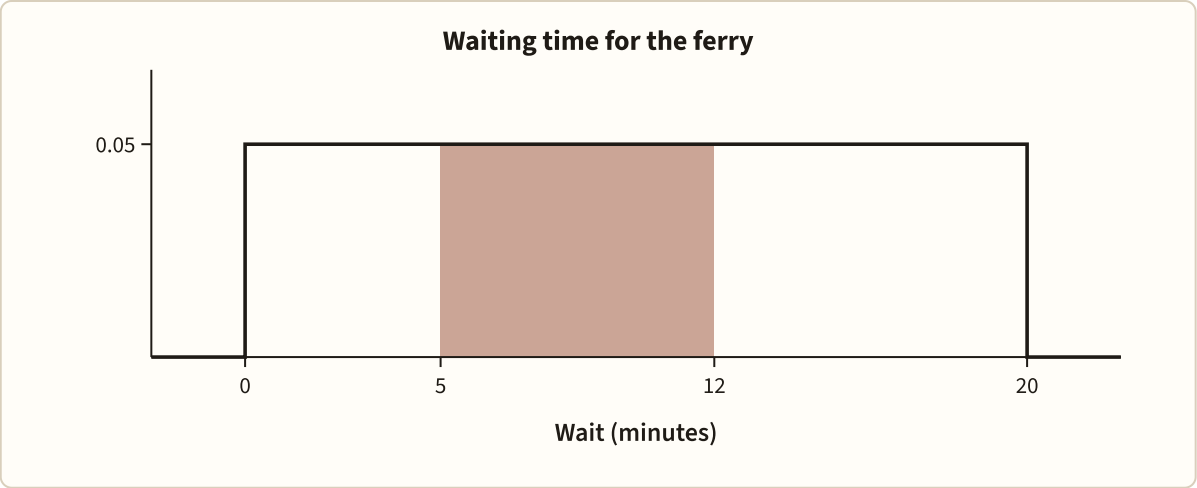
Look at the histogram of the heights of 200 corn stalks. The bars are percents, and all the classes have the same width, so the area of each bar is proportional to the percent of stalks in that class. If we measured thousands of stalks and made the classes narrower and narrower, the jagged tops of the bars would smooth out into a curve. That smooth curve is a **density curve**: an idealized picture of the whole distribution.



A density curve has two rules: it never goes below the horizontal axis, and the total area under it is exactly 1 (that is, 100% of the values). Why area and not height? Because in a histogram it is the AREA of a bar that counts the values; a bar twice as wide holds twice as many values at the same height. So on a density curve, the **area above an interval is the proportion of values in that interval**. To answer "what proportion of stalks are taller than 85 inches?" you find the area under the curve to the right of 85.

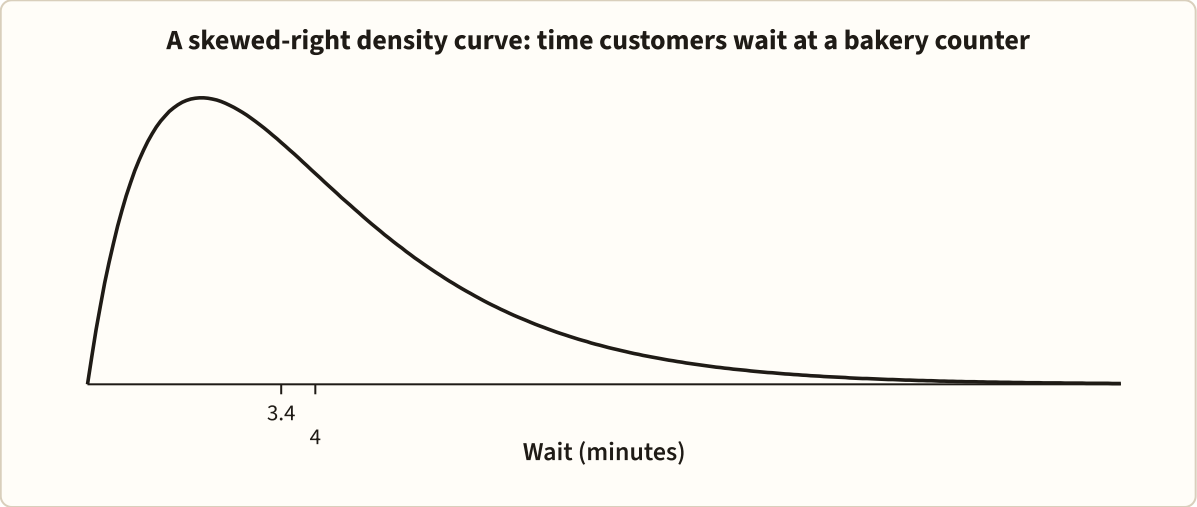
A density curve is a model. Real data never fit it perfectly, but a good model is simple and close enough to be useful. Because a density curve describes the idealized population, we write its mean as μ and its standard deviation as σ, and keep $\bar{x}$ and s for actual data.

The simplest density curve is flat. A ferry leaves every 20 minutes; if you arrive at a random time, your wait is equally likely to be anywhere from 0 to 20 minutes. Its density curve is a rectangle 20 units wide. Since the area must be 1, the height must be $\frac{1}{20} = 0.05$. This is a **uniform distribution**, and every area is just width times height. The shaded proportion of waits between 5 and 12 minutes is $(12 - 5) \times 0.05 = 0.35$.



Two centers can be read from any density curve. The **median** is the equal-areas point: half the area lies on each side. The **mean** is the balance point: imagine the curve cut from a sheet of wood and set on a knife edge. On a symmetric curve both are at the center of symmetry, so for the ferry the mean and median are both 10 minutes.

On a skewed curve they separate, and the reason is the seesaw. A small amount of area far out in the long tail has great leverage, just as a small child sitting far from the middle can balance a heavier adult sitting near it. So the balance point (the mean) is pulled out toward the long tail. The median only asks "is half the area on each side?" and does not care how far away the tail values are, so it stays closer to the peak. On the skewed-right curve below the median is about 3.4 minutes but the mean is 4 minutes: the mean is farther out in the tail. This is the same reason the median resisted outliers in Lesson 2.1.



One more consequence of area: the proportion of values EXACTLY equal to one number, such as a wait of exactly 7 minutes, is the area of a line with no width, which is 0. You will use this fact in Lesson 3.4.

KEY FORMULAS

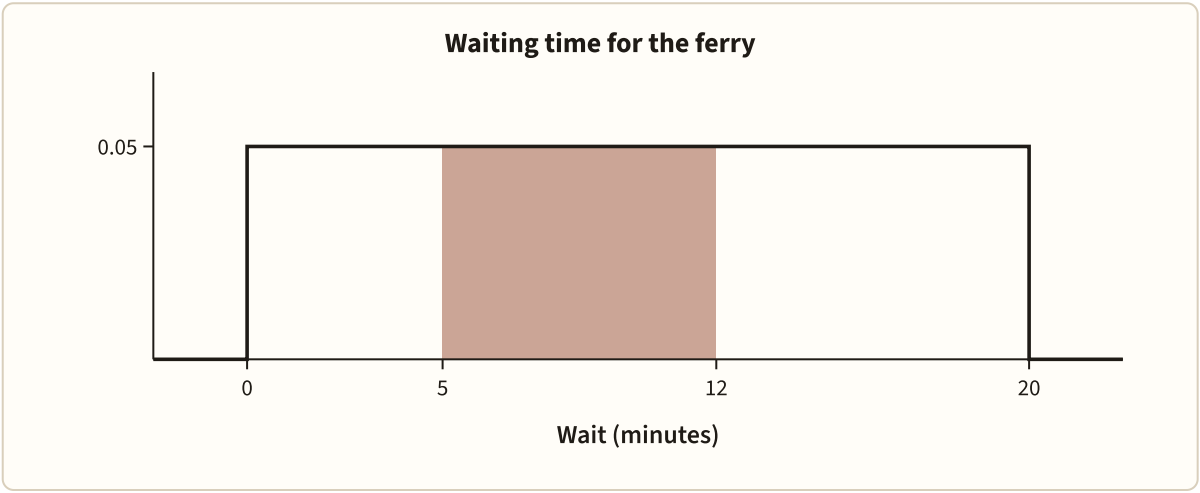
$$\text{area under a density curve over an interval} = \text{proportion of values in that interval}$$

$$\text{uniform distribution from } a \text{ to } b: \quad \text{height} = \frac{1}{b-a}$$

$$\text{uniform: proportion between } c \text{ and } d = \frac{d-c}{b-a}, \qquad \mu = \text{median} = \frac{a+b}{2}$$

Practice habit. Always sketch the density curve and shade the area that answers the question before you compute. A quick sketch tells you whether your answer should be small or large, and catches mistakes like finding the area on the wrong side.

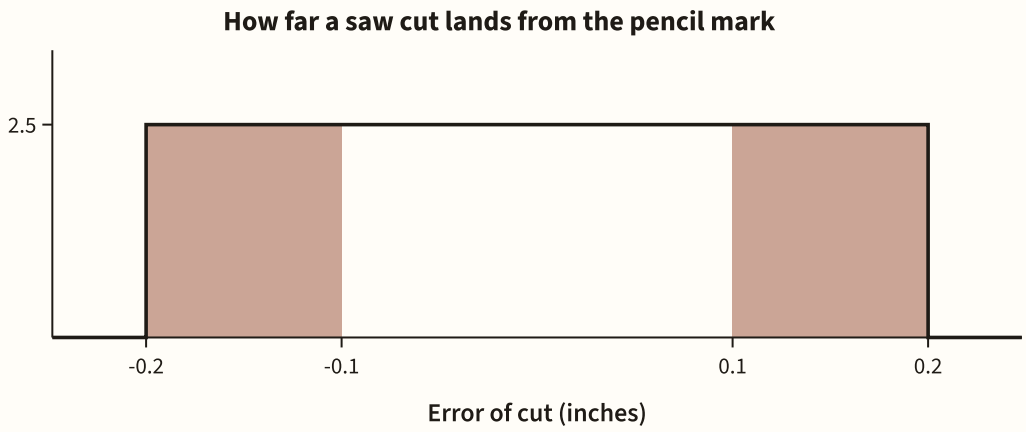
Example 1. A ferry across a lake leaves the dock every 20 minutes. If you arrive at a random time, your wait X is uniformly distributed from 0 to 20 minutes. (a) What is the height of the density curve? (b) What proportion of waits are between 5 and 12 minutes? (c) What are the mean and the median wait?



1. The density curve is a rectangle from 0 to 20, and its area must be 1.
2. (a) Height $= \dfrac{1}{b-a} = \dfrac{1}{20-0} = 0.05$.
3. (b) The shaded region is a rectangle of width $12 - 5 = 7$ and height 0.05.
4. Area $= 7 \times 0.05 = 0.35$.
5. (c) The rectangle is symmetric, so it balances in the middle and half the area lies on each side of the middle: mean $=$ median $= \dfrac{0+20}{2} = 10$ minutes.
6. Conclusion: about 35% of the time a traveler arriving at random waits between 5 and 12 minutes, and the typical wait is 10 minutes.

Answer: (a) 0.05; (b) 0.35; (c) mean = median = 10 minutes

Example 2. When a carpenter saws a board, the cut lands a little off the pencil mark. The error X (in inches; negative means short) is uniformly distributed from -0.2 to 0.2 inch. (a) What proportion of cuts miss the mark by more than 0.1 inch in either direction? (b) Find the number c so that 90% of cuts land within c inch of the mark.



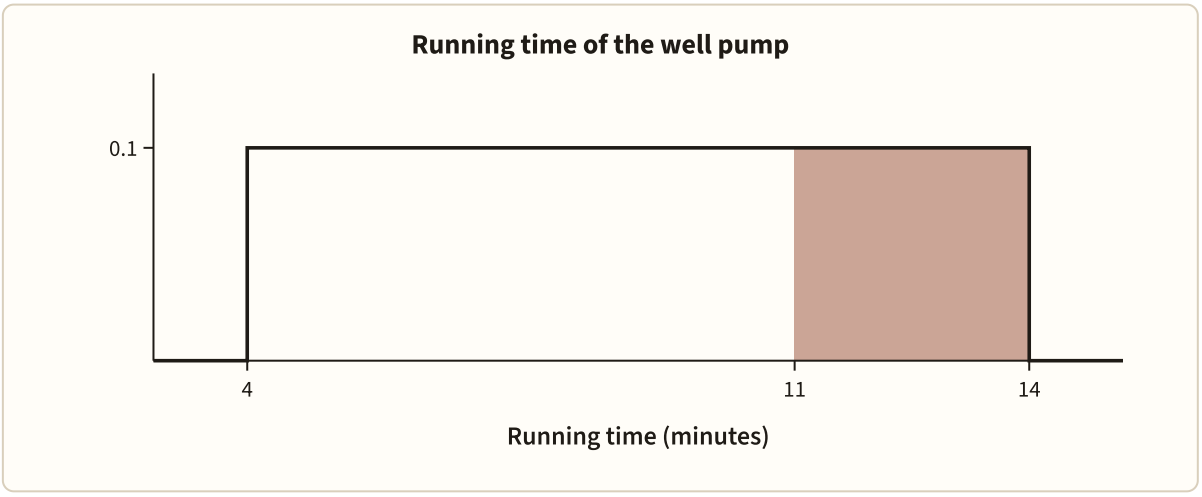
1. Width of the rectangle: $0.2 - (-0.2) = 0.4$, so the height is $\dfrac{1}{0.4} = 2.5$.
2. (a) "More than 0.1 inch off" means $X < -0.1$ or $X > 0.1$: two shaded rectangles, each of width 0.1.
3. Each has area $0.1 \times 2.5 = 0.25$, so the total is $0.25 + 0.25 = 0.50$.
4. (b) The region from $-c$ to c has width $2c$, so its area is $2c \times 2.5 = 5c$.
5. Set $5c = 0.90$, so $c = 0.18$ inch.
6. Conclusion: half of this carpenter's cuts miss the mark by more than a tenth of an inch, and 90% land within 0.18 inch of it.

Answer: (a) 0.50; (b) $c = 0.18$ inch

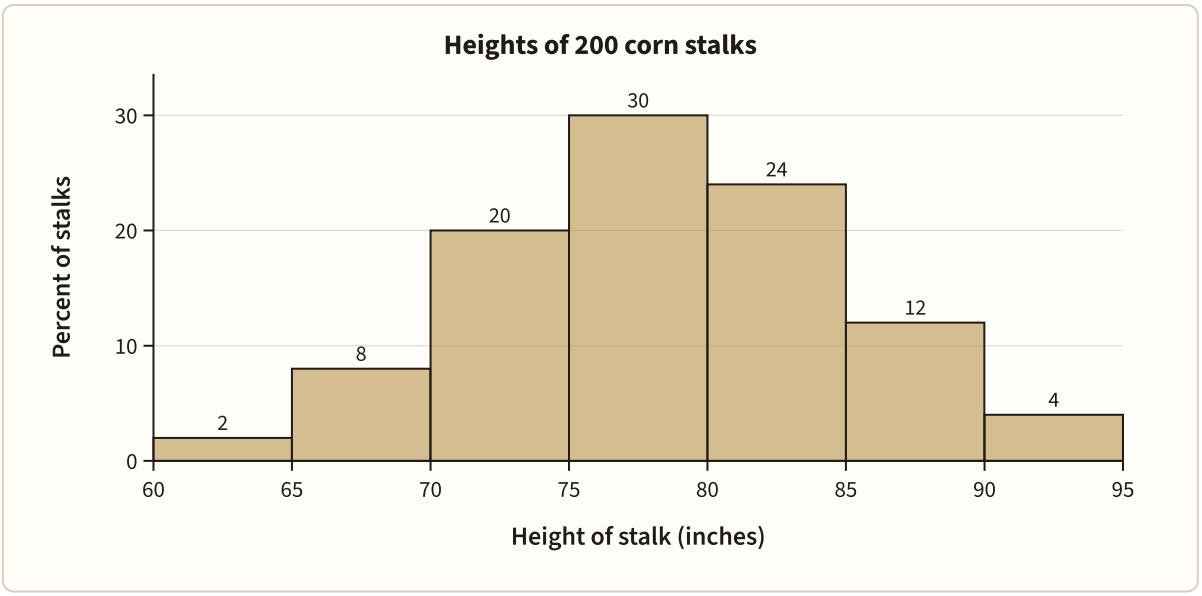
PRACTICE

1. EASY A board-game spinner gives a number X that is uniformly distributed from 0 to 8 (any decimal is possible). What is the height of its density curve?
2. EASY For the spinner in Problem 1 (uniform from 0 to 8), what proportion of spins give a number greater than 6?
3. EASY Is a flat curve of height 0.5 from 0 to 4 (and 0 everywhere else) a valid density curve? Answer yes or no and give the reason.
4. EASY Explain in your own words why the total area under every density curve must be exactly 1.
5. MEDIUM A calculator's random-number key gives numbers uniformly distributed from 0 to 1. What proportion of the numbers fall between 0.25 and 0.70?

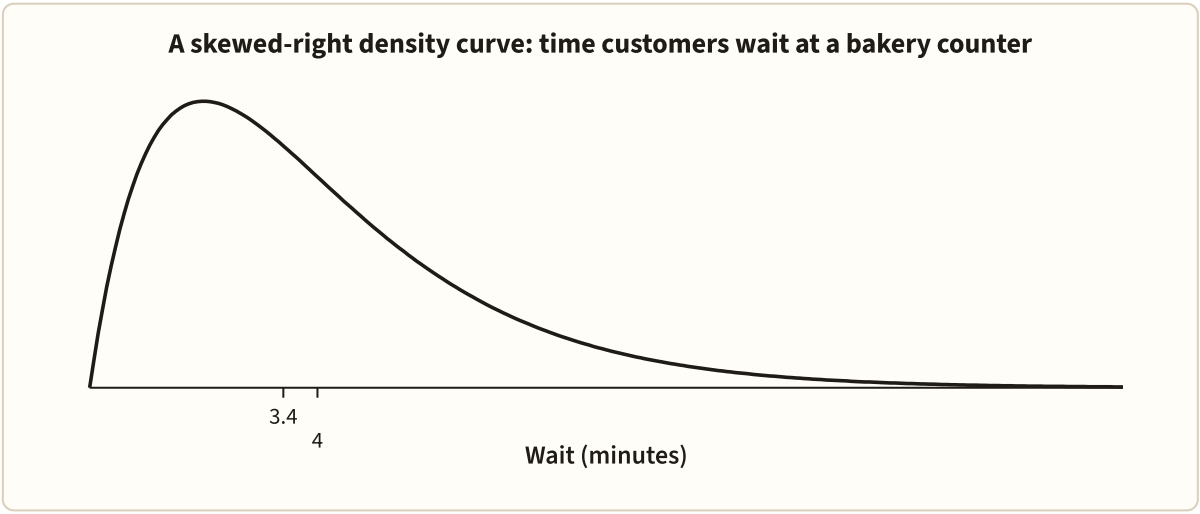
6. MEDIUM Each time the pump on the Miller farm's well switches on, it runs for a time that is uniformly distributed from 4 to 14 minutes. (a) What is the height of the density curve? (b) What proportion of runs last longer than 11 minutes? (c) What is the median running time?



7. MEDIUM Use the histogram of 200 corn stalk heights (the bars are percents). (a) What percent of stalks are at least 85 inches tall? (b) What percent are shorter than 70 inches?



8. MEDIUM Look at the density curve of waiting times at a bakery counter. (a) Describe its shape. (b) Which is larger, the mean or the median? Explain why.



9. MEDIUM For a certain density curve, the area to the left of 12 is 0.50 and the area to the left of 15 is 0.80. (a) What is the median? (b) What proportion of values lie between 12 and 15?

10. HARD A uniform density curve starts at 0 and has height 0.04. Where does it end?

11. HARD The choir director starts Thursday rehearsal late by an amount that is uniformly distributed from 0 to 6 minutes. (a) Find the 90th percentile of the lateness. (b) What proportion of rehearsals start within 1 minute of the mean lateness? Round to 4 decimal places.

12. HARD A density curve is skewed left. Where will the mean be compared with the median? Explain why, using the idea of a balance point.

13. CHALLENGE A density curve is made of two flat pieces: height 0.3 from 0 to 2, and height h from 2 to 4 (and 0 elsewhere). (a) Find h. (b) Find the median, to 2 decimal places. (c) Find the mean. (Hint: each rectangle balances at its own middle; the whole curve balances at the area-weighted mean of those middles.)

COMMON MISTAKES

- Reading proportions from the HEIGHT of a density curve instead of the AREA. Height alone is not a proportion; for a uniform curve, multiply width by height.
- Forgetting that the height of a uniform density is $\frac{1}{b-a}$, not 1. Check that width times height equals 1.
- Thinking the mean and median of a skewed curve are the same. On a skewed curve the mean is pulled toward the long tail.
- Mixing up symbols: use μ and σ for a density curve (the idealized population) and $\bar{x}$ and s for actual data.

ANSWERS TO PRACTICE

1. Height $= \dfrac{1}{8-0} = 0.125$.
2. Width $8 - 6 = 2$; area $= 2 \times 0.125 = 0.25$.
3. No. Its area is $4 \times 0.5 = 2$, but a density curve must have total area exactly 1.
4. Area under a density curve stands for the proportion of values. All of the values (100%, a proportion of 1) lie somewhere under the curve, so the whole area must be 1.
5. Height 1; area $= (0.70 - 0.25) \times 1 = 0.45$.
6. (a) $\dfrac{1}{14-4} = 0.1$. (b) $(14 - 11) \times 0.1 = 0.3$. (c) $\dfrac{4+14}{2} = 9$ minutes.
7. (a) $12\% + 4\% = 16\%$. (b) $2\% + 8\% = 10\%$.
8. (a) Skewed right: a peak at short waits and a long tail toward long waits. (b) The mean is larger (about 4 minutes, versus a median of about 3.4). The few long waits in the right tail are far from the center, so they have great leverage and pull the balance point to the right, while the median only splits the area in half.
9. (a) The median is 12, since half the area is to its left. (b) $0.80 - 0.50 = 0.30$.
10. Area $= b \times 0.04 = 1$, so $b = \dfrac{1}{0.04} = 25$.
11. Height $\dfrac{1}{6}$. (a) Area to the left of x is $\dfrac{x}{6} = 0.90$, so $x = 5.4$ minutes. (b) The mean is 3, so the interval is 2 to 4: $\dfrac{2}{6} \approx 0.3333$.
12. The mean will be to the left of (less than) the median. The long left tail puts some area far from the center, and area far from the balance point has extra leverage, so it pulls the balance point toward the left tail. The median only splits the area in half, so it is not pulled as far.
13. (a) $0.3 \times 2 + 2h = 1$, so $0.6 + 2h = 1$ and $h = 0.2$. (b) The left piece has area $0.6 > 0.5$, so the median m is in it: $0.3m = 0.5$, $m \approx 1.67$. (c) Mean $= 0.6 \times 1 + 0.4 \times 3 = 1.8$. The mean is larger than the median because more of the far-right area pulls the balance point right.

LESSON 3.3 · 2 SESSIONS OF 45 MINUTES

Normal Distributions and the 68–95–99.7 Rule

OBJECTIVES

- Describe a normal distribution N(μ, σ) and explain how μ and σ determine its center and spread, including where σ appears on the curve.
- State the 68–95–99.7 rule and use it to find percents and values for any normal distribution.
- Explain why the 68–95–99.7 rule is the same for every normal distribution.
- Explain why many measurements, such as weights of eggs and errors in measurement, are roughly normal.

VIDEO

WATCH

The Normal Distribution (2 of 3: Characteristics of the population)

Eddie Woo · 11:19

When to watch: Before Example 1: the 68–95–99.7 rule.

https://www.youtube.com/watch?v=yXAUf0ax07k

BACKUP

An Introduction to the Normal Distribution

jbstatistics · 5:27

When to watch: Review: the shape of a normal curve and the roles of μ and σ.

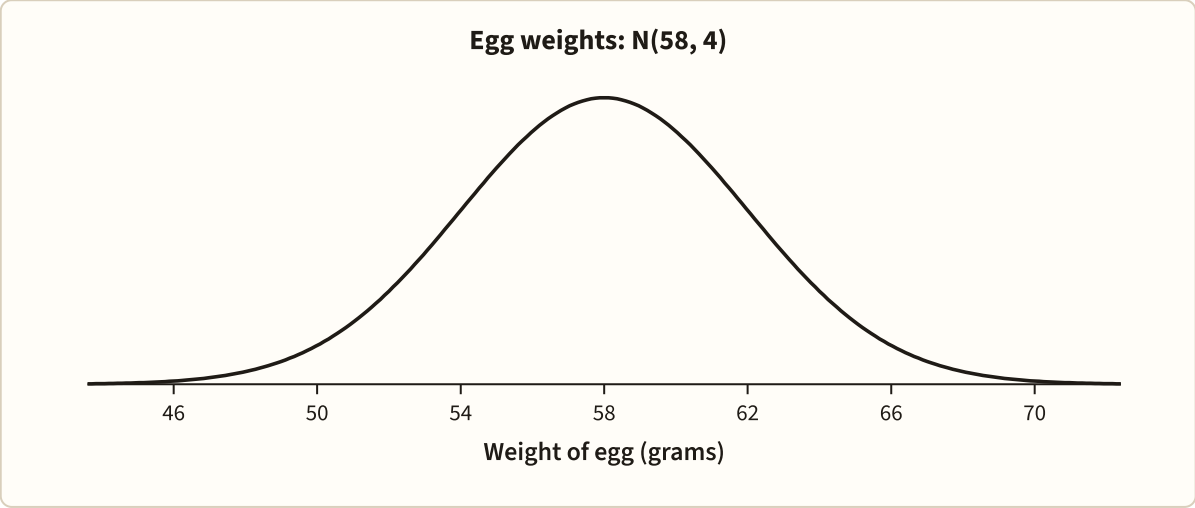
https://www.youtube.com/watch?v=iYiOVISWXS4

KEY TERMS

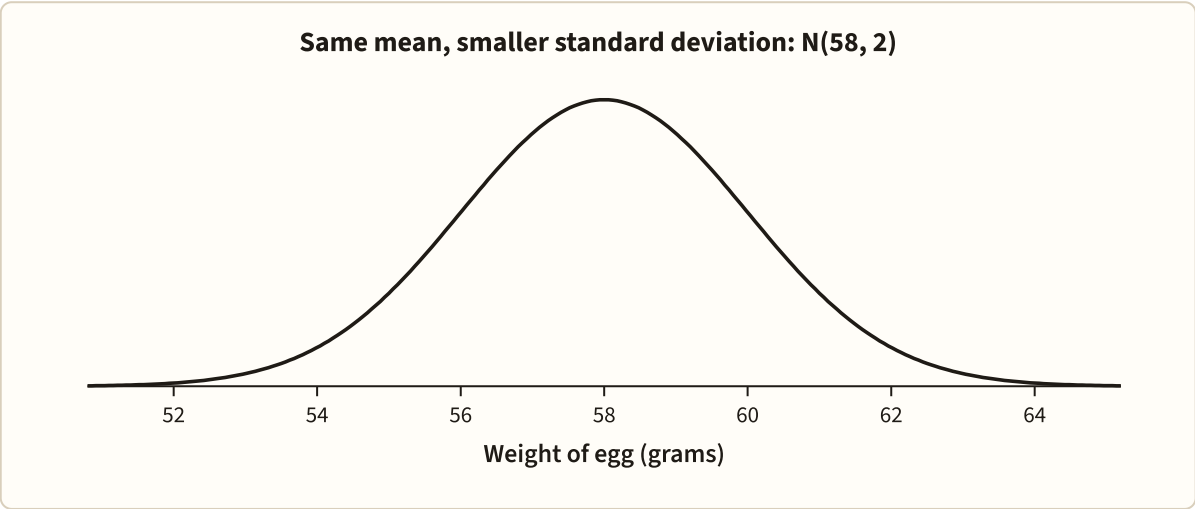
Normal distribution	A distribution described by a symmetric, single-peaked, bell-shaped density curve. It is completely determined by its mean μ and standard deviation σ, and written $N(\mu, \sigma)$.
Normal curve	The bell-shaped density curve of a normal distribution. Its peak, mean and median are all at μ.
Inflection point	A point where a curve changes from bending downward (like an upside-down bowl) to bending upward (like a bowl). On a normal curve these points are exactly one standard deviation on each side of the mean.
68–95–99.7 rule	In any normal distribution, about 68% of values lie within 1σ of the mean, about 95% within 2σ, and about 99.7% within 3σ. It is also called the empirical rule.

CORE IDEAS

The most important density curves in statistics are the **normal curves**: symmetric, single-peaked and bell-shaped. The notation $N(\mu, \sigma)$ names one of them. The mean μ sits at the center of symmetry, under the peak, and because the curve is symmetric the median is there too. The standard deviation σ sets how spread out the curve is. Here is the curve for the weights of eggs from a flock, $N(58, 4)$ in grams, with μ and $\mu \pm 1\sigma, 2\sigma, 3\sigma$ marked.



You can see σ on a normal curve. Start at the peak and move outward: near the top the curve bends downward like the top of a hill, and farther out it bends upward as it flattens toward the axis. The points where the bending changes direction, the **inflection points**, are exactly σ away from μ on each side. For the eggs they are at 54 and 62 grams. A curve with the same mean but smaller σ is taller and narrower, because its total area must still be 1. Compare the egg curve with the $N(58, 2)$ curve below.



The 68–95–99.7 rule. In every normal distribution, about 68% of the values lie within one standard deviation of the mean, about 95% within two, and about 99.7% within three. For the eggs: about 68% weigh between 54 and 62 grams, about 95% between 50 and 66 grams, and about 99.7% between 46 and 70 grams.

Why is the rule the same for every normal curve, whether it describes eggs in grams or corn in inches? Because all normal curves are really one curve drawn with different scales. If you measure in standard deviations from the mean (z-scores, Lesson 3.1), every normal distribution becomes exactly the same curve. So the area within "one standard deviation" depends only on that curve, not on the units.

To use the rule, cut the curve into pieces and use symmetry. Within 1σ there is 68%, so 34% on each side of the mean. Between 1σ and 2σ there is $\frac{95-68}{2} = 13.5\%$ on each side. Between 2σ and 3σ there is $\frac{99.7-95}{2} = 2.35\%$ on each side, and beyond 3σ only 0.15% on each side. Half the area (50%) is on each side of the mean. Always draw the curve, mark μ and $\mu \pm 1\sigma, 2\sigma, 3\sigma$, shade the region you want and add up its pieces.

Why are so many measurements roughly normal? The weight of an egg is the result of many small influences added together: the hen's breed and age, her feed, the weather, the day. Some push the weight up and some push it down, and they mostly cancel, so most eggs land near the middle and only rarely do nearly all the pushes go the same way. Errors in measuring the position of a star behave the same way: small errors from the air, the instrument and the observer's eye add up. Totals of many small, independent effects tend to pile up in a bell shape (you will see why in Unit 9). But not everything is normal: waiting times, rainfall from single storms and incomes are usually skewed right, and the 68–95–99.7 rule does not fit them.

KEY FORMULAS

$$N(\mu, \sigma): \text{ a normal distribution with mean } \mu \text{ and standard deviation } \sigma$$

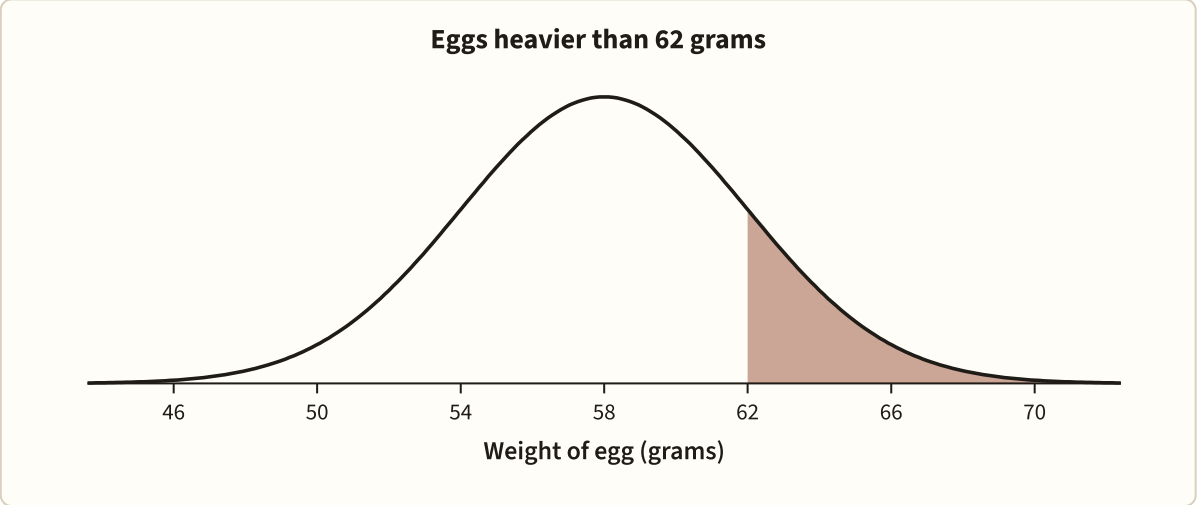
$$\text{about } 68\% \text{ within } \mu \pm \sigma, \quad 95\% \text{ within } \mu \pm 2\sigma, \quad 99.7\% \text{ within } \mu \pm 3\sigma$$

$$\text{pieces on each side of } \mu: \ 34\%, \ 13.5\%, \ 2.35\%, \ 0.15\%$$

From history. Abraham de Moivre found the normal curve in 1733 as an approximation to probabilities for tossing many coins. Carl Friedrich Gauss used it in 1809 to describe errors in astronomical measurements, which is why it is sometimes called the Gaussian curve, and Pierre-Simon Laplace extended the theory.

Practice habit. Know the 68–95–99.7 rule by heart, along with its pieces 34, 13.5, 2.35 and 0.15 on each side. Every time, sketch the bell curve and label μ and $\mu \pm 1\sigma, 2\sigma, 3\sigma$ before you add any areas.

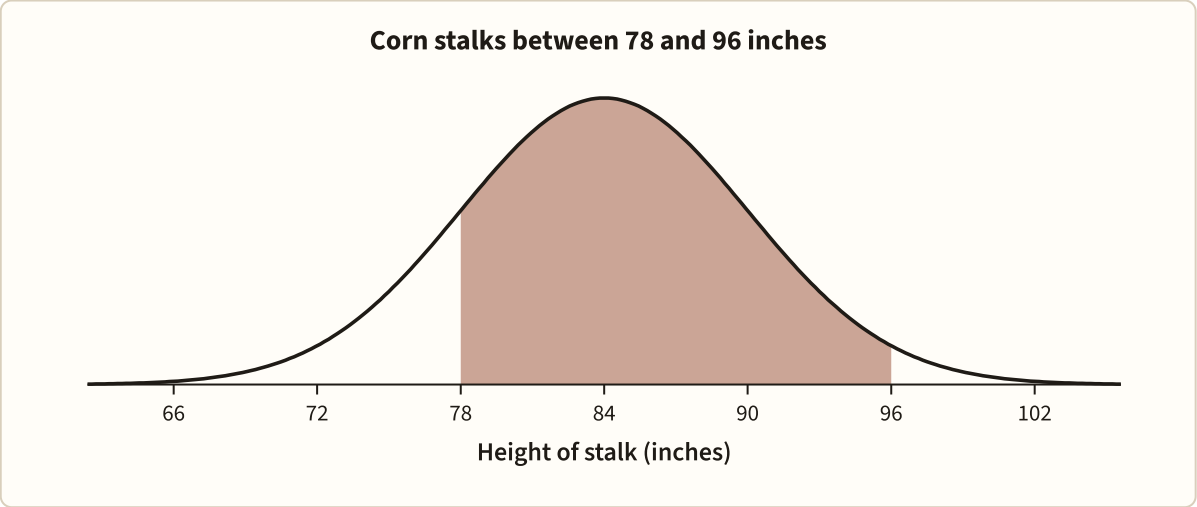
Example 1. The weights of eggs from the Miller family's flock are approximately $N(58, 4)$ grams. (a) Between what two weights do the middle 95% of eggs fall? (b) What percent of eggs weigh more than 62 grams?



1. Mark the curve: $\mu = 58$; $\mu \pm 1\sigma$ = 54 and 62; $\mu \pm 2\sigma$ = 50 and 66; $\mu \pm 3\sigma$ = 46 and 70.
2. (a) The middle 95% lies within 2σ of the mean: $58 - 2(4) = 50$ and $58 + 2(4) = 66$.
3. (b) 62 grams is $\mu + 1\sigma$. About 68% of eggs are within 1σ, so $100\% - 68\% = 32\%$ are outside.
4. By symmetry, half of that 32% is in each tail: $\dfrac{32\%}{2} = 16\%$ above 62 grams.
5. Conclusion: about 95% of the eggs weigh between 50 and 66 grams, and about 16% weigh more than 62 grams.

Answer: (a) 50 to 66 grams; (b) about 16%

Example 2. The heights of mature corn stalks in a field are approximately $N(84, 6)$ inches. (a) About what percent of stalks are between 78 and 96 inches tall? (b) At about what percentile is a stalk 72 inches tall?



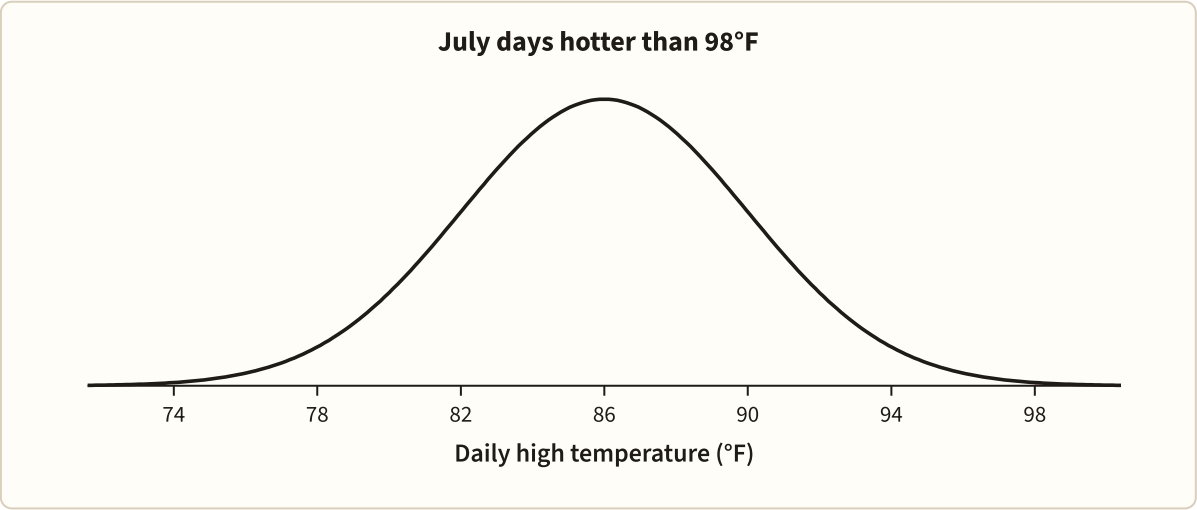
1. Mark the curve: $\mu = 84$; $\mu \pm 1\sigma$ = 78 and 90; $\mu \pm 2\sigma$ = 72 and 96.
2. (a) 78 is $\mu - 1\sigma$ and 96 is $\mu + 2\sigma$. Split the region at the mean.
3. From 78 to 84 (one σ below the mean): half of 68%, which is 34%.
4. From 84 to 96 (two σ above the mean): half of 95%, which is 47.5%.
5. Total: $34\% + 47.5\% = 81.5\%$.
6. (b) 72 is $\mu - 2\sigma$. Outside 2σ there is 5%, so $\frac{5\%}{2} = 2.5\%$ lies below 72.
7. Conclusion: about 81.5% of the stalks are between 78 and 96 inches tall, and a 72-inch stalk is at about the 2.5th percentile: only about 2.5% of the stalks are shorter.

Answer: (a) about 81.5%; (b) about the 2.5th percentile

PRACTICE

1. EASY Board lengths from a sawmill are approximately $N(96.0, 0.1)$ inches. Between what two lengths do the middle 68% of boards fall?
2. EASY Loaves from a bakery weigh approximately $N(680, 10)$ grams. About what percent of loaves weigh between 660 and 700 grams?
3. EASY Where are the inflection points of the normal curve $N(70, 3)$?
4. EASY Two normal curves have the same mean, but the first has $\sigma = 2$ and the second has $\sigma = 5$. Describe how the curves look different, and explain why the first must be taller.
5. MEDIUM The Miller family's eggs weigh approximately $N(58, 4)$ grams. About what percent of eggs weigh less than 50 grams?

6. MEDIUM Daily high temperatures in July in a town are approximately $N(86, 4)$ °F. About what percent of July days are hotter than 98°F?



7. MEDIUM Calf birth weights on a ranch are approximately $N(80, 8)$ pounds. About what percent of calves weigh between 88 and 96 pounds?

8. MEDIUM A farm stand sells 400 eggs from a flock whose weights are approximately $N(58, 4)$ grams. About how many of the eggs weigh more than 66 grams?

9. MEDIUM An orchard sells apples in bags. The weights of the bags are approximately normal, and the middle 68% of bags weigh between 41 and 49 ounces. Find μ and σ.

10. HARD Drives by the members of a golf club are approximately $N(250, 15)$ yards. The longest 2.5% of drives are longer than what distance?

11. HARD Weights of lambs at weaning are approximately $N(40, 5)$ pounds. At about what percentile is a lamb that weighs 35 pounds? Explain how you got it.

12. HARD Many astronomers each measure the brightness of the same star. Their measurements are approximately normal, centered close to the true brightness. Explain why measurement errors like these tend to follow a normal curve.

13. CHALLENGE Corn in Field A is approximately $N(84, 6)$ inches tall and corn in Field B is approximately $N(90, 3)$ inches. (a) About what percent of Field A's stalks are taller than 96 inches? (b) About what percent of Field B's stalks are taller than 96 inches? (c) Field B has the higher mean. Explain why it does not have more stalks over 96 inches.

COMMON MISTAKES

- Using 68% for one side only. 68% lies within 1σ on BOTH sides together; one side holds 34%.
- Forgetting to halve the tails. 5% lies outside $\mu \pm 2\sigma$, but only 2.5% lies above $\mu + 2\sigma$.
- Applying the 68–95–99.7 rule to data that are clearly skewed. The rule is a property of normal curves only; look at the shape first.
- Adding the wrong pieces. Draw the curve, mark $\mu \pm 1\sigma, 2\sigma, 3\sigma$, and shade before you add.

ANSWERS TO PRACTICE

1. 96.0 ± 0.1: from 95.9 to 96.1 inches.
2. 660 and 700 are $\mu \pm 2\sigma$, so about 95%.
3. One standard deviation on each side of the mean: at 67 and 73.
4. Both are centered at the same place, but the curve with $\sigma = 5$ is wider and more spread out, and the curve with $\sigma = 2$ is narrower. The total area under each curve must be 1, so the narrower curve has to be taller to hold the same area.
5. 50 is $\mu - 2\sigma$; $\dfrac{100\% - 95\%}{2} = 2.5\%$.
6. 98 is $\mu + 3\sigma$; $\dfrac{100\% - 99.7\%}{2} = 0.15\%$.
7. 88 is $\mu + 1\sigma$ and 96 is $\mu + 2\sigma$; the area between is $47.5\% - 34\% = 13.5\%$.
8. 66 is $\mu + 2\sigma$, so about 2.5% are heavier: $0.025 \times 400 = 10$ eggs.
9. The middle 68% is $\mu \pm \sigma$, so μ is the midpoint, $\dfrac{41 + 49}{2} = 45$ ounces, and $\sigma = 49 - 45 = 4$ ounces.
10. The top 2.5% lies beyond $\mu + 2\sigma = 250 + 2(15) = 280$ yards.
11. 35 is $\mu - 1\sigma$. Outside 1σ is 32%, half of it below: 16%. The lamb is at about the 16th percentile; about 16% of lambs weigh less.
12. Each measurement's error is the total of many small, independent effects (shimmer in the air, the instrument, the observer's eye and timing). Some push the reading up and some down, so they usually partly cancel and most errors are small; it is rare for nearly all of them to push the same way, so large errors are rare. Totals of many small independent effects pile up in a symmetric bell shape.
13. (a) $96 = 84 + 2(6)$, so about 2.5%. (b) $96 = 90 + 2(3)$, so also about 2.5%. (c) Field A is more spread out. Its larger σ makes its right tail reach farther, which makes up for its lower center; in both fields, 96 inches is exactly two standard deviations above the mean.

LESSON 3.4 · 2 SESSIONS OF 45 MINUTES

Finding Normal Proportions

OBJECTIVES

- Explain why standardizing turns any normal distribution into the standard normal distribution N(0, 1).
- Use the standard normal table or technology to find the proportion of values to the left of, to the right of, or between given values.
- Draw and shade a normal curve for every problem, and state the answer as a proportion in context.
- Explain why a single exact value has probability 0 for a continuous distribution.

VIDEO

WATCH

Finding Areas Using the Standard Normal Table (for tables that give the area to left of z)

jbstatistics · 6:15

When to watch: Before Example 1: reading Table A (area to the left of z).

https://www.youtube.com/watch?v=-UljIcq_rfc

BACKUP

Z Scores and Normal Distributions (Example Problems)

Mario's Math Tutoring · 6:17

When to watch: Extra practice: standardizing a value and finding a normal proportion.

https://www.youtube.com/watch?v=rkdbHQPUA1Y

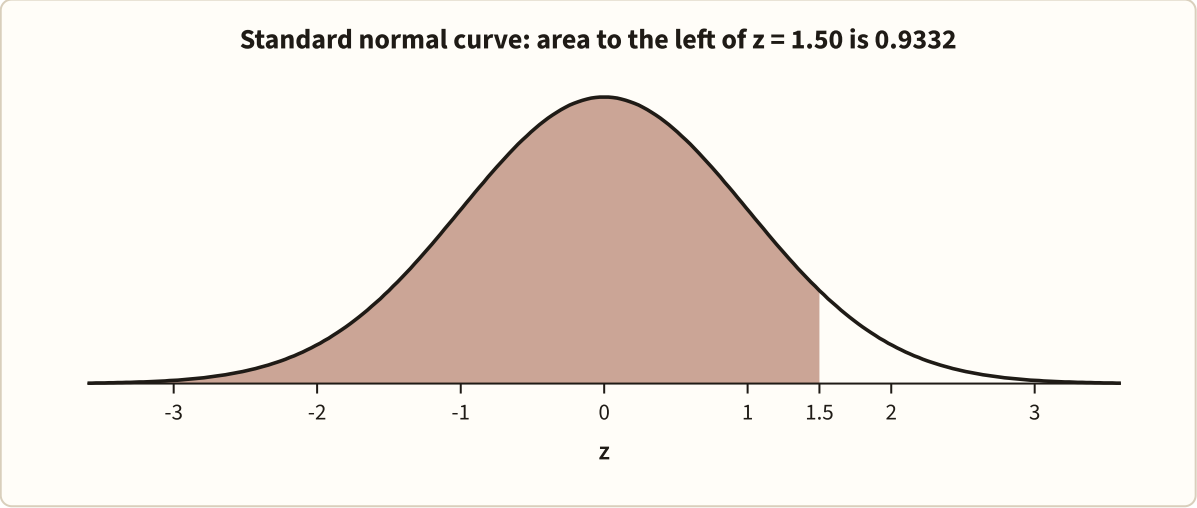
KEY TERMS

Standard normal distribution	The normal distribution with mean 0 and standard deviation 1, written $N(0, 1)$. Its values are z-scores, and the letter Z is used for it.
Standard normal table	A table that gives the area under the standard normal curve to the LEFT of each z-value, from about $z = -3.49$ to $z = 3.49$.
Cumulative proportion	The proportion of values less than a given value: the area under the density curve to its left.
Continuous distribution	A distribution whose values can be any number in an interval (like weights or lengths), described by a density curve. Probabilities are areas under the curve.

CORE IDEAS

The 68–95–99.7 rule answers questions only at whole numbers of standard deviations. To find the proportion of eggs heavier than 64 grams when the weights are $N(58, 4)$, we need the area beyond 1.50 standard deviations, and the rule cannot give it. We need one table that works for every normal distribution.

Standardizing makes one table enough. If x comes from $N(\mu, \sigma)$, its z-score $z = \dfrac{x - \mu}{\sigma}$ comes from $N(0, 1)$. Why? Subtracting μ moves the center to 0, dividing by σ makes the standard deviation 1 (Lesson 3.1), and a change of scale like this does not change the shape (Lesson 2.6), so the curve is still normal. The area to the left of 64 on the egg curve is exactly the area to the left of $z = 1.50$ on the standard normal curve, because both measure the same thing: the proportion of values less than 1.50 standard deviations above the mean.



Reading the table. The standard normal table gives the area to the LEFT of z. Find the row for the ones digit and tenths digit of z, then the column for the hundredths digit. For $z = 1.53$, go to row 1.5 and column .03: the area is 0.9370, so 93.70% of a normal distribution lies below 1.53 standard deviations above the mean. For $z = 1.50$ the area is 0.9332.

Part of the standard normal table (area to the left of z)

z	.00	.01	.02	.03	.04
1.4	0.9192	0.9207	0.9222	0.9236	0.9251
1.5	0.9332	0.9345	0.9357	0.9370	0.9382
1.6	0.9452	0.9463	0.9474	0.9484	0.9495

Right and between. The table gives only areas to the left, but the total area is 1, so the area to the RIGHT of z is $1 - (\text{area to the left})$. The area BETWEEN two values is (area to the left of the larger) minus (area to the left of the smaller). This is why a sketch is essential: it tells you which of these to do, and whether your answer should be more or less than 0.5.

Technology. A graphing calculator's normalcdf(lower, upper, μ, σ), or a free online normal calculator, finds the area directly without standardizing. It carries more decimal places than the table, so its answer may differ from a table answer in the fourth decimal place. In this course the answers give the technology value, sometimes followed by the table value.

A routine for every problem. (1) State the question with the variable: "find the proportion with $X > 64$." (2) Draw the normal curve, mark the mean and the value, and shade the area. (3) Standardize. (4) Find the area with the table or technology, and answer in context.

A single value has proportion 0. For a continuous variable, the proportion of eggs weighing EXACTLY 64 grams (64.000... with endless zeros) is the area above one point, a line with no width, which is 0. So $P(X < 64)$ and $P(X \leq 64)$ are the same number, and you do not need to worry about whether an endpoint is included.

KEY FORMULAS

$$z = \frac{x - \mu}{\sigma}$$

$$\text{area to the right of } z = 1 - (\text{area to the left of } z)$$

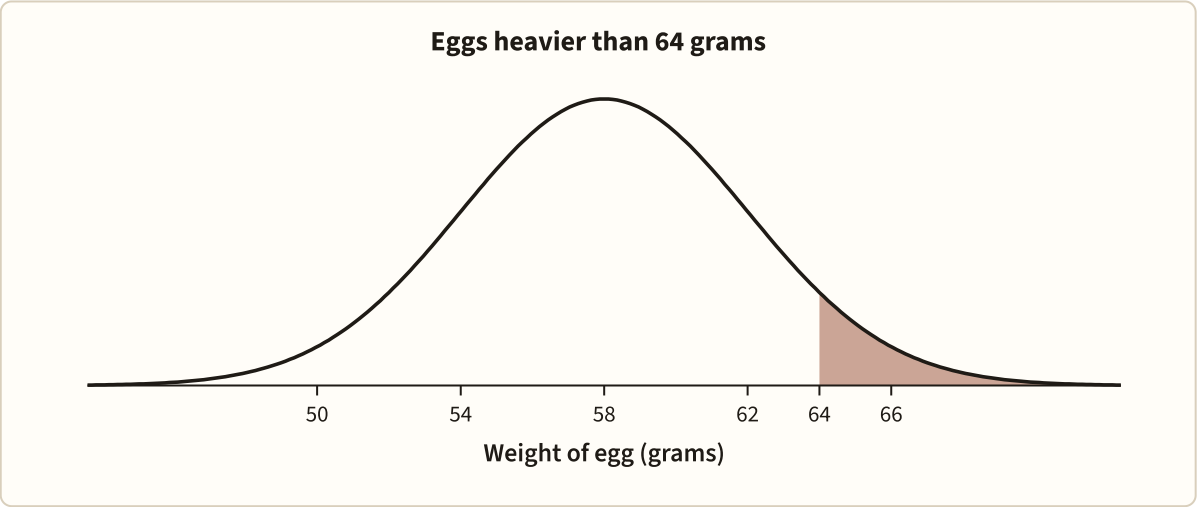
$$\text{area between } z_1 \text{ and } z_2 = (\text{area left of } z_2) - (\text{area left of } z_1)$$

$$P(X = a) = 0 \text{ for a continuous distribution, so } P(X < a) = P(X \leq a)$$

Practice habit. Always draw and shade the normal curve. Mark the mean and the value, shade the region the question asks for, and check that your final answer matches the picture: a small tail should get a small answer.

WORKED EXAMPLES

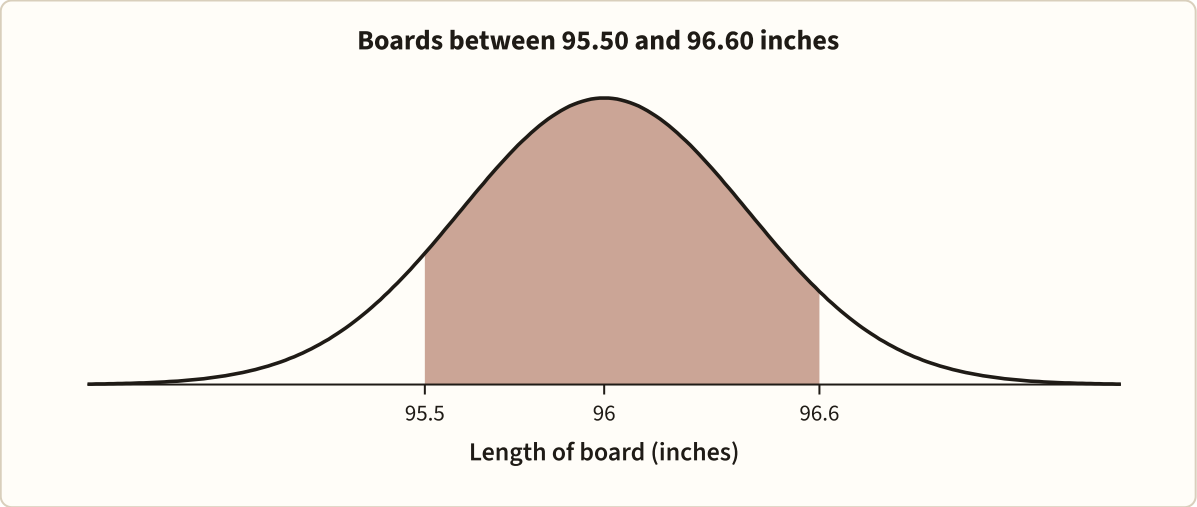
Example 1. The eggs from the Miller family's flock have weights that are approximately $N(58, 4)$ grams. What proportion of the eggs weigh more than 64 grams?



1. State: find the proportion with $X > 64$, where X is the weight of an egg.
2. Draw: mark 58 in the middle and 64 to its right, and shade the right tail beyond 64. The answer should be small, well under 0.5.
3. Standardize: $z = \dfrac{x - \mu}{\sigma} = \dfrac{64 - 58}{4} = 1.50$.
4. Table: the area to the left of $z = 1.50$ is 0.9332.
5. Area to the right: $1 - 0.9332 = 0.0668$. (Technology: normalcdf(64, 1E99, 58, 4) ≈ 0.0668.)
6. Conclusion: about 0.0668 of the eggs, roughly 7 eggs in every 100, weigh more than 64 grams.

Answer: $z = 1.50$; area to the right $= 1 - 0.9332 = 0.0668$

Example 2. A sawmill cuts boards whose lengths are approximately $N(96.00, 0.40)$ inches. A builder accepts boards from 95.50 to 96.60 inches long. What proportion of the boards are acceptable?



1. State: find the proportion with $95.50 < X < 96.60$.
2. Draw: shade the region between 95.50 (left of the mean) and 96.60 (right of the mean). The answer should be well over 0.5.
3. Standardize the lower end: $z = \dfrac{95.50 - 96.00}{0.40} = \dfrac{-0.50}{0.40} = -1.25$.
4. Standardize the upper end: $z = \dfrac{96.60 - 96.00}{0.40} = \dfrac{0.60}{0.40} = 1.50$.
5. Table: area left of 1.50 is 0.9332; area left of -1.25 is 0.1056.
6. Between: $0.9332 - 0.1056 = 0.8276$. Technology, normalcdf(95.5, 96.6, 96, 0.4), gives 0.8275; the table's rounded entries make the last digit differ.
7. Conclusion: about 0.8275 of the boards, nearly 83%, are acceptable to the builder; about 17% are too short or too long.

Answer: 0.8275 (table: $0.9332 - 0.1056 = 0.8276$)

PRACTICE

1. EASY Use the standard normal table or technology to find the proportion of the standard normal distribution with $Z < 1.20$.
2. EASY Find the proportion of the standard normal distribution with $Z > -0.75$.
3. EASY Find the proportion of the standard normal distribution with $-1.00 < Z < 1.50$.
4. EASY Loaf weights are approximately $N(680, 10)$ grams. Explain why the proportion of loaves weighing exactly 680 grams is 0, and why $P(X < 690)$ equals $P(X \leq 690)$.

5. MEDIUM Calf birth weights on a ranch are approximately $N(80, 8)$ pounds. What proportion of calves weigh less than 70 pounds?

6. MEDIUM Daily high temperatures in July are approximately $N(86, 4)$ °F. What proportion of July days have a high above 93°F?

7. MEDIUM Drives at a golf club are approximately $N(240, 20)$ yards. What proportion of drives travel between 230 and 260 yards?

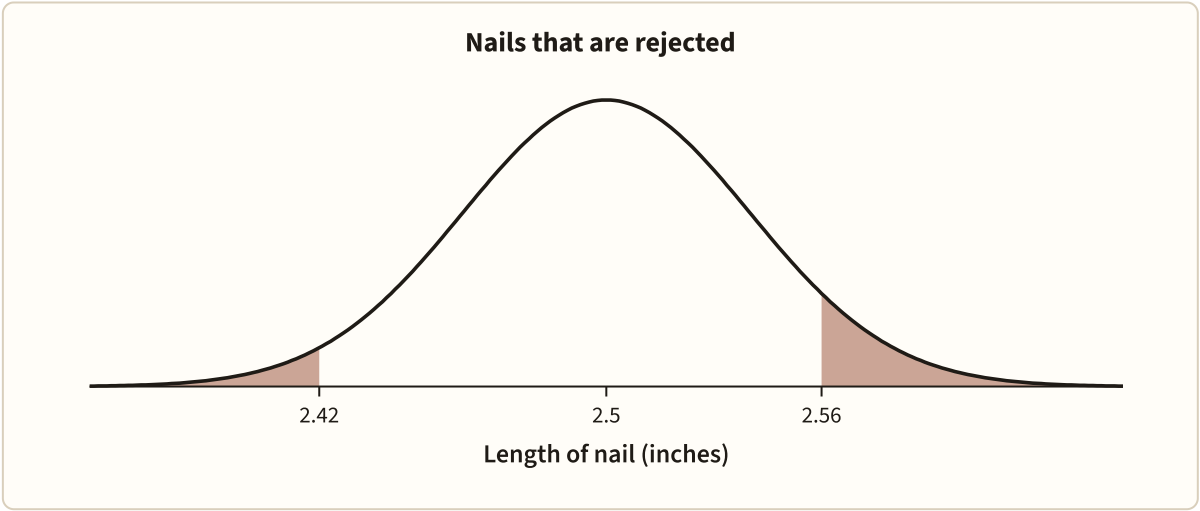
8. MEDIUM For eggs that weigh $N(58, 4)$ grams, a student found $P(X > 64) = 0.9332$. (a) Explain, without a table, how you know this must be wrong. (b) Give the correct value.

9. MEDIUM Apple diameters in an orchard are approximately $N(2.7, 0.4)$ inches. Apples wider than 3.5 inches are sold as "extra large." (a) What proportion of apples are extra large? (b) About how many of 1,000 apples would you expect to be extra large? Round to the nearest apple.

10. MEDIUM Yearly rainfall at a farm is approximately $N(38, 6)$ inches. What proportion of years have between 29 and 35 inches of rain?

11. HARD A bakery's loaves weigh approximately $N(680, 10)$ grams, but the label says 670 grams. (a) What proportion of loaves are underweight (less than 670 grams)? (b) The baker raises the mean to 690 grams with the same standard deviation. What proportion are underweight now?

12. HARD A machine makes nails whose lengths are approximately $N(2.50, 0.04)$ inches. Nails shorter than 2.42 inches or longer than 2.56 inches are rejected. What proportion of nails are rejected?



13. HARD Corn stalks in a field are approximately $N(84, 6)$ inches tall. Samuel's stalk is 90 inches tall. At what percentile is it, to the nearest whole percent? Interpret the answer in context.

14. CHALLENGE The Miller family's eggs weigh approximately $N(58, 4)$ grams. (a) What proportion of eggs weigh between 56 and 63 grams? (b) Out of 30 dozen eggs (360 eggs), about how many would you expect in this range? Round to the nearest egg.

COMMON MISTAKES

- Giving the area to the LEFT when the question asks for the area to the right. The table always gives the left area; for "more than" subtract from 1. A sketch shows which you need.
- Subtracting in the z-score the wrong way, or dividing by the wrong number. Always compute $\frac{x-\mu}{\sigma}$: value minus mean, divided by the standard deviation.
- For an area between two values, subtracting the z-scores instead of the areas. Find both left areas first, then subtract the areas.
- Rounding z too early or reading the wrong row. Round z to two decimals only for the table, and use the row (ones and tenths) and column (hundredths) carefully.

ANSWERS TO PRACTICE

1. 0.8849
2. Area to the left of -0.75 is 0.2266; $1 - 0.2266 = 0.7734$.
3. $0.9332 - 0.1587 = 0.7745$.
4. Weight is continuous, so proportions are areas under the curve. The area above a single value is a line with no width, so its area is 0. Including or leaving out the single value 690 adds an area of 0, so the two proportions are the same.
5. $z = \frac{70-80}{8} = -1.25$; area to the left $= 0.1056$.
6. $z = \frac{93-86}{4} = 1.75$; area to the right $= 1 - 0.9599 = 0.0401$.
7. $z = \frac{230-240}{20} = -0.50$ and $z = \frac{260-240}{20} = 1.00$; $0.8413 - 0.3085 = 0.5328$.
8. (a) 64 is above the mean of 58, so less than half of the area lies to its right; the answer must be less than 0.5. The student gave the area to the LEFT. (b) $1 - 0.9332 = 0.0668$.
9. (a) $z = \frac{3.5-2.7}{0.4} = 2.00$; $1 - 0.9772 = 0.0228$. (b) $0.0228 \times 1000 \approx 23$ apples.
10. $z = \frac{29-38}{6} = -1.50$ and $z = \frac{35-38}{6} = -0.50$; $0.3085 - 0.0668 = 0.2417$.
11. (a) $z = \frac{670-680}{10} = -1.00$; proportion 0.1587. (b) $z = \frac{670-690}{10} = -2.00$; proportion 0.0228. Moving the mean 10 grams cut the underweight loaves from about 16% to about 2%.
12. $z = \frac{2.42-2.50}{0.04} = -2.00$, left tail 0.0228; $z = \frac{2.56-2.50}{0.04} = 1.50$, right tail $1 - 0.9332 = 0.0668$; total $0.0228 + 0.0668 = 0.0896$.
13. $z = \frac{90-84}{6} = 1.00$; area to the left $= 0.8413$, so about the 84th percentile: Samuel's stalk is taller than about 84% of the stalks in the field.
14. (a) $z = \frac{56-58}{4} = -0.50$ and $z = \frac{63-58}{4} = 1.25$; $0.8944 - 0.3085 = 0.5859$ from the table (technology 0.5858). (b) $0.5858 \times 360 \approx 211$ eggs.

LESSON 3.5 · 2 SESSIONS OF 45 MINUTES

Finding Values from Normal Proportions

OBJECTIVES

- Work backward from a proportion to a z-score using the standard normal table or technology (inverse normal).
- Find percentiles and cutoff values of a normal distribution with $x = \mu + z\sigma$, and interpret them in context.
- Find the mean or the standard deviation of a normal distribution when the other one and a percentile are known.
- Explain why a given percentile is always the same number of standard deviations from the mean in every normal distribution.

VIDEO

WATCH

Finding Percentiles Using the Standard Normal Table (for tables that give the area to left of z)

jbstatistics · 7:32

When to watch: Before Example 1: working backward from an area to a z-score with Table A.

https://www.youtube.com/watch?v=9KOJtiHAavE

BACKUP

z-score Calculations & Percentiles in a Normal Distribution

ProfRobBob · 13:39

When to watch: Extra practice: z-scores and percentiles in normal problems.

https://www.youtube.com/watch?v=_86q-hn_3DQ

KEY TERMS

Inverse normal calculation	Starting from an area (a proportion) and finding the value that cuts off that area: the reverse of Lesson 3.4.
Percentile of a normal distribution	The value with the given percent of the distribution below it. The 90th percentile has area 0.90 to its left.
Cutoff value	A value that separates a group, such as the height a corn stalk must exceed to be in the tallest 10%.
Unstandardizing	Turning a z-score back into the original units with $x = \mu + z\sigma$: start at the mean and step z standard deviations.

CORE IDEAS

In Lesson 3.4 you started with a value and found an area. Often the question runs the other way. A seed company wants to save seed from the tallest 10% of its corn: how tall must a stalk be? A builder wants to know the length below which the shortest 5% of boards fall. Now you know the AREA and need the VALUE.

The steps, reversed. (1) Draw the curve and shade the area you are given. (2) Find the area to the LEFT of the unknown value, since the table and technology work with left areas: the tallest 10% means an area of 0.90 to the left. (3) Find the z-score with that left area. (4) Unstandardize: solving $z = \frac{x - \mu}{\sigma}$ for x gives $x = \mu + z\sigma$, which says "start at the mean and step z standard deviations" (to the left if z is negative).

Using the table backward. Look INSIDE the body of the table for the area closest to the one you want, then read z from the row and column. For a left area of 0.90, the closest entry is 0.8997, at $z = 1.28$. For a left area of 0.95, the entries 0.9495 ($z = 1.64$) and 0.9505 ($z = 1.65$) are equally close, so we use the halfway value $z = 1.645$.

Reading the standard normal table backward

z	Area to the left of z
1.27	0.8980
1.28	0.8997
1.29	0.9015
1.64	0.9495
1.65	0.9505

Using technology. A graphing calculator's invNorm(area to the left, μ, σ), or an online inverse normal calculator, gives the value directly. It finds z more precisely (1.2816 instead of 1.28), so answers can differ slightly from table answers; round as the problem asks and the two usually agree.

Why a percentile is always the same number of standard deviations from the mean. All normal curves are the same curve measured in standard deviations. So the 90th percentile of EVERY normal distribution is $\mu + 1.282\sigma$, whether it describes corn, eggs or golf drives; the 95th percentile is always $\mu + 1.645\sigma$. A small table of these z-values is worth knowing well, because you will meet them again when you build confidence intervals in Unit 10.

z-values for common percentiles (technology, 3 decimals)

Area to the left	Percentile	z
0.75	75th	0.674
0.8	80th	0.842
0.85	85th	1.036
0.9	90th	1.282
0.95	95th	1.645
0.975	97.5th	1.960
0.99	99th	2.326

Finding μ or σ. The equation $z = \dfrac{x - \mu}{\sigma}$ connects four numbers. If you know three, you can solve for the fourth. A beekeeper whose filling machine has $\sigma = 0.2$ ounce can find the mean fill that keeps only 2% of jars under 16 ounces; a sawmill with mean length 96.00 inches can find how small σ must be so that only 5% of boards are shorter than 95.50 inches. Find z from the area first, then solve the equation with Algebra I.

KEY FORMULAS

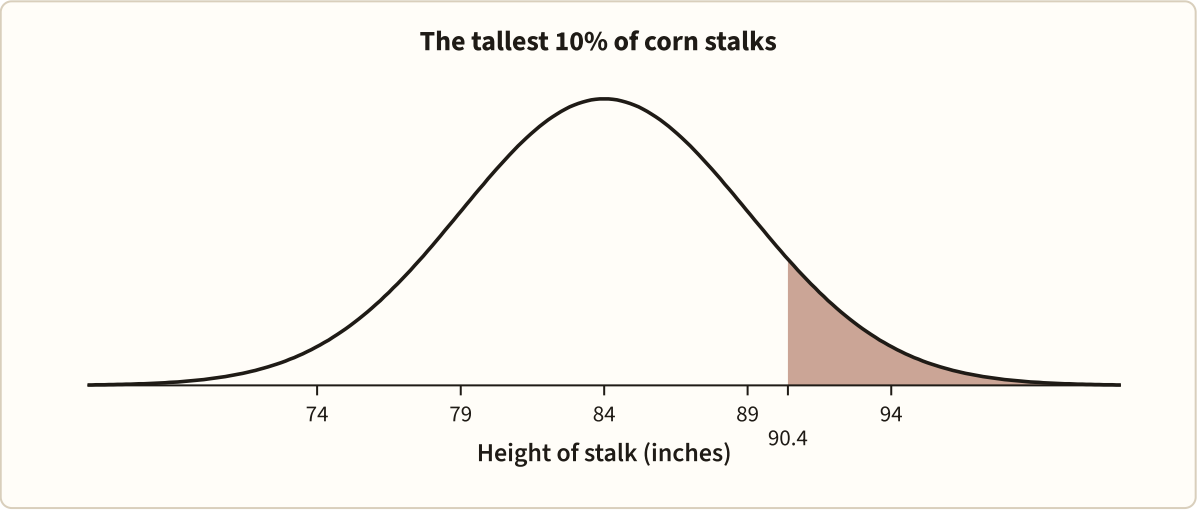
$$x = \mu + z\sigma$$

$$\mu = x - z\sigma \qquad \sigma = \frac{x - \mu}{z}$$

top p proportion: use left area $1 - p$

Practice habit. Before looking up anything, ask: "What is the area to the LEFT of the value I want?" Then know these by heart: left area 0.90 → z = 1.282, 0.95 → z = 1.645, 0.975 → z = 1.960, 0.99 → z = 2.326.

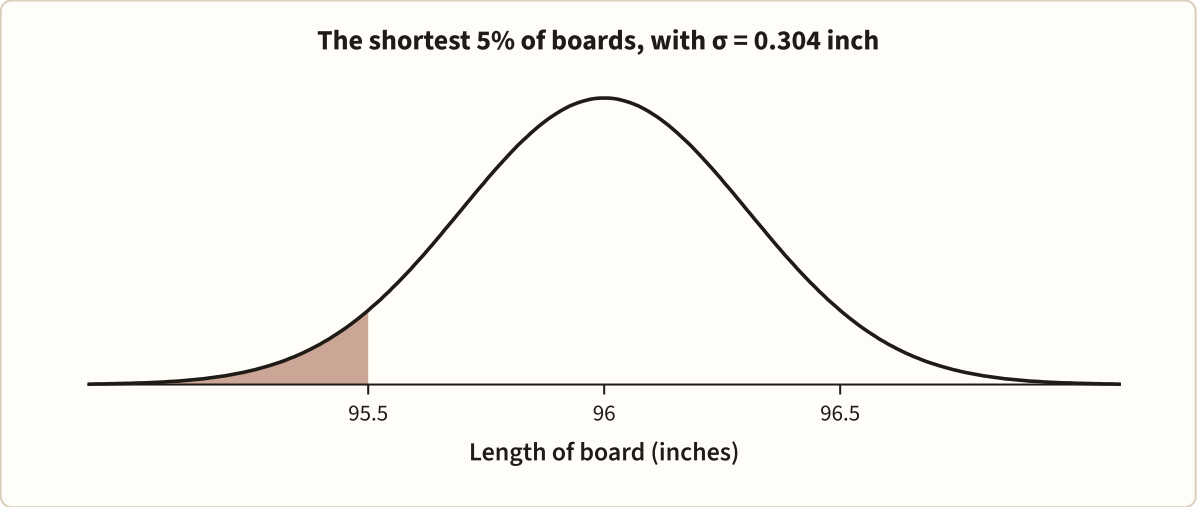
Example 1. The heights of corn stalks in a seed field are approximately $N(84, 5)$ inches. The farmer will save seed only from the tallest 10% of stalks. How tall must a stalk be to make the cut? Round to the nearest tenth of an inch.



1. Draw: shade the right tail with area 0.10. The cutoff is above the mean.
2. Left area: the cutoff has $1 - 0.10 = 0.90$ of the stalks below it.
3. Table: the area closest to 0.90 is 0.8997, at $z = 1.28$. Technology: invNorm(0.90) gives $z \approx 1.2816$.
4. Unstandardize: $x = \mu + z\sigma = 84 + 1.2816 \times 5 = 84 + 6.408 = 90.408$.
5. Check with the table value: $84 + 1.28 \times 5 = 90.4$. Both round to 90.4.
6. Conclusion: a stalk must be at least about 90.4 inches tall to be among the tallest 10% in the field.

Answer: $x = 84 + 1.2816(5) \approx 90.4$ inches

Example 2. A sawmill cuts boards with mean length 96.00 inches. The lengths are approximately normal. The mill wants only 5% of its boards to be shorter than 95.50 inches. How small must the standard deviation be? Round to 3 decimal places.



1. Draw: 95.50 is below the mean, with a left area of 0.05.
2. Find z with left area 0.05: the table entries 0.0505 ($z = -1.64$) and 0.0495 ($z = -1.65$) are equally close, so $z = -1.645$ (technology: -1.6449).
3. Write the z equation with σ unknown: $-1.645 = \dfrac{95.50 - 96.00}{\sigma} = \dfrac{-0.50}{\sigma}$.
4. Solve: $\sigma = \dfrac{-0.50}{-1.645} \approx 0.30395$.
5. Round: $\sigma \approx 0.304$ inch.
6. Check: $z = \dfrac{95.50 - 96.00}{0.304} \approx -1.64$, with a left area of about 0.05.
7. Conclusion: the mill must hold its standard deviation to about 0.304 inch or less; then no more than 5% of its boards will be shorter than 95.50 inches.

Answer: $\sigma = \dfrac{0.50}{1.645} \approx 0.304$ inch

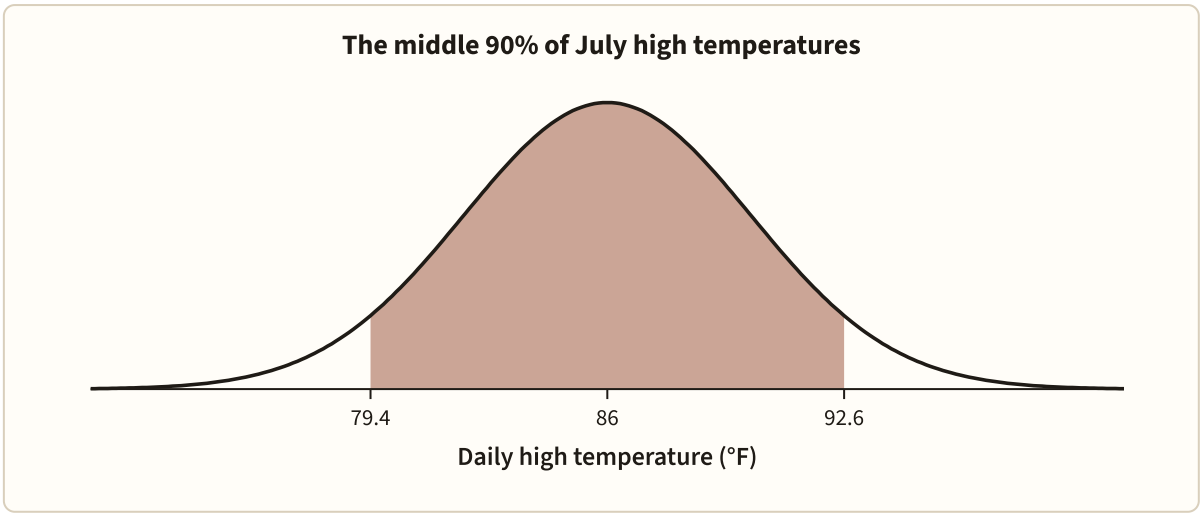
PRACTICE

1. EASY Find the z-score that has area 0.80 to its left under the standard normal curve. Round to 2 decimal places.
2. EASY Find the z-score that has area 0.05 to its RIGHT under the standard normal curve. Round to 3 decimal places.
3. EASY Eggs from a flock weigh approximately $N(58, 4)$ grams. Find the 75th percentile of egg weight. Round to the nearest tenth of a gram.
4. EASY To find the cutoff for the TOP 10% of a normal distribution, Grace looks up an area of 0.90 in the table, not 0.10. Explain why.

5. MEDIUM Calf birth weights are approximately $N(80, 8)$ pounds. The rancher watches the lightest 20% of calves more closely. Below what weight is a calf in the lightest 20%? Round to the nearest tenth of a pound.

6. MEDIUM Golf drives at a club are approximately $N(240, 20)$ yards. How long must a drive be to be among the longest 2.5%? Round to the nearest yard.

7. MEDIUM July high temperatures are approximately $N(86, 4)$ °F. Between what two temperatures are the middle 90% of July highs? Round to the nearest tenth of a degree.



8. MEDIUM A bakery's loaves weigh approximately $N(680, 10)$ grams. The heaviest 1% are set aside for display. What is the cutoff weight? Round to the nearest tenth of a gram.

9. MEDIUM Explain why the 90th percentile of every normal distribution is about $\mu + 1.28\sigma$, whatever the values of μ and σ.

10. HARD A beekeeper's filling machine puts honey in jars with amounts that are approximately normal with $\sigma = 0.2$ ounce. At what mean should the machine be set so that only 2% of jars contain less than 16 ounces? Round to 2 decimal places.

11. HARD Lamb weights at weaning are approximately normal with mean 40 pounds, and 95% of lambs weigh less than 48 pounds. Find the standard deviation. Round to 2 decimal places.

12. HARD Winter snowfall in a mountain town is approximately $N(64, 12)$ inches. The town calls a winter "severe" if its snowfall is in the top 15%. (a) Find the cutoff, to the nearest inch. (b) Last winter brought 80 inches. Was it severe? Answer yes or no.

13. CHALLENGE Finishing times in a 5-km church fun run are approximately normal. 5% of runners finish in under 20 minutes and 5% take longer than 32 minutes. Find the mean and the standard deviation. Round the standard deviation to 2 decimal places.

14. CHALLENGE Apple weights in an orchard are approximately normal. 20% of the apples weigh less than 150 grams and 10% weigh more than 210 grams. Find μ and σ, each to the nearest gram.

COMMON MISTAKES

- Looking up the right-tail area as if it were a left area (using 0.10 for the top 10%). Always convert to the area to the LEFT of the value first.
- Forgetting the sign of z for values below the mean. A cutoff for the lowest 20% has a negative z, so $x = \mu + z\sigma$ comes out less than μ.
- Stopping at z. The question asks for a weight or a length; unstandardize with $x = \mu + z\sigma$ and give units.
- Using the z-score's area as the answer. In these problems the area is GIVEN; the answer is a value on the measurement scale.

ANSWERS TO PRACTICE

1. $z \approx 0.84$ (technology 0.8416; table closest 0.7995 at 0.84).
2. Left area $1 - 0.05 = 0.95$, so $z = 1.645$.
3. $z \approx 0.6745$; $x = 58 + 0.6745(4) \approx 60.7$ grams.
4. The table gives the area to the LEFT of z. A value with 10% of the distribution above it has the other 90% below it, so its left area is $1 - 0.10 = 0.90$.
5. Left area 0.20: $z \approx -0.8416$; $x = 80 - 0.8416(8) \approx 73.3$ pounds.
6. Left area 0.975: $z = 1.960$; $x = 240 + 1.96(20) = 279.2 \approx 279$ yards.
7. 5% in each tail, so $z = \pm 1.645$; $86 \pm 1.645(4) = 86 \pm 6.58$: from 79.4°F to 92.6°F.
8. Left area 0.99: $z \approx 2.326$; $x = 680 + 2.326(10) \approx 703.3$ grams.
9. Standardizing turns every normal distribution into the same standard normal curve. On that curve the point with area 0.90 to its left is always $z \approx 1.28$, so in the original units the 90th percentile is always 1.28 standard deviations above the mean.
10. Left area 0.02: $z \approx -2.054$ (table -2.05). $\mu = x - z\sigma = 16 - (-2.054)(0.2) = 16 + 0.411 \approx 16.41$ ounces.
11. Left area 0.95: $z = 1.645$. $\sigma = \dfrac{48 - 40}{1.645} \approx 4.86$ pounds.
12. (a) Left area 0.85: $z \approx 1.036$; $x = 64 + 1.036(12) \approx 76$ inches. (b) Yes: 80 inches is above the cutoff of about 76 inches.
13. By symmetry $\mu = \dfrac{20 + 32}{2} = 26$ minutes. 32 has left area 0.95, so $z = 1.645$: $\sigma = \dfrac{32 - 26}{1.645} \approx 3.65$ minutes.
14. Left area 0.20: $z = -0.842$, so $\mu - 0.842\sigma = 150$. Left area 0.90: $z = 1.282$, so $\mu + 1.282\sigma = 210$. Subtract: $2.124\sigma = 60$, $\sigma \approx 28.25$; then $\mu = 150 + 0.842(28.25) \approx 173.8$. So $\mu \approx 174$ grams and $\sigma \approx 28$ grams.

LESSON 3.6 · 2 SESSIONS OF 45 MINUTES

Assessing Normality

OBJECTIVES

- Explain why you must check that data are roughly normal before using normal calculations.
- Judge normality from the shape of a dotplot, stemplot or histogram and by comparing the mean and median.
- Check a data set against the 68–95–99.7 rule by counting values within 1, 2 and 3 standard deviations of the mean.
- Interpret a normal probability plot: a roughly straight plot means roughly normal; a curved plot means skewed.

VIDEO

WATCH

Normal Quantile-Quantile Plots

jbstatistics · 12:08

When to watch: Before Example 1: reading a normal probability (quantile) plot.

https://www.youtube.com/watch?v=X9_ISJ0YpGw

BACKUP

Normal Probability Plots & the TI-84

ProfRobBob · 8:28

When to watch: After Example 1: making a normal probability plot on a calculator.

https://www.youtube.com/watch?v=-kaB03IqpRw

> **Note:** Both videos plot data values against normal scores; the axes may be swapped from the course's plots, so read straight versus curved the same way.

KEY TERMS

Roughly normal	Close enough to a normal shape that normal calculations give useful answers: single-peaked, roughly symmetric, with no outliers and tails about as long as a normal curve's.
Normal probability plot	A graph that plots each data value against the z-score it would be expected to have if the data came from a normal distribution. A roughly straight line of points means the data are roughly normal.

Expected z-score	The z-score a value would have, given its position in order, if the data were exactly normal. For 10 values, the smallest is placed at the 5th percentile ($z \approx -1.64$), the next at the 15th ($z \approx -1.04$), and so on up to the 95th.
Assessing normality	Using graphs and the 68–95–99.7 rule to decide whether a normal model fits a data set well enough to use.

CORE IDEAS

Every calculation in Lessons 3.4 and 3.5 assumed the distribution was normal. If it is not, the answers can be badly wrong. A normal model for the rain from summer storms, for example, would predict that some storms drop a negative amount of rain! So before using a normal model, ask: do the data look normal? You will ask this question again in Units 10–12, where many methods require roughly normal data.

Method 1: Plot the data. A dotplot, stemplot or histogram of roughly normal data has one peak, is roughly symmetric, has tails that thin out on both sides, and has no outliers. Strong skewness, two peaks, gaps or outliers are warnings. A quick numerical check helps: in a symmetric distribution the mean and median are close; when the mean is far above the median, the data are skewed right. The split stemplot of 20 corn stalk heights below has a single peak in the low 80s and thins out evenly on both sides, and its mean (82.0 inches) equals its median (82 inches).

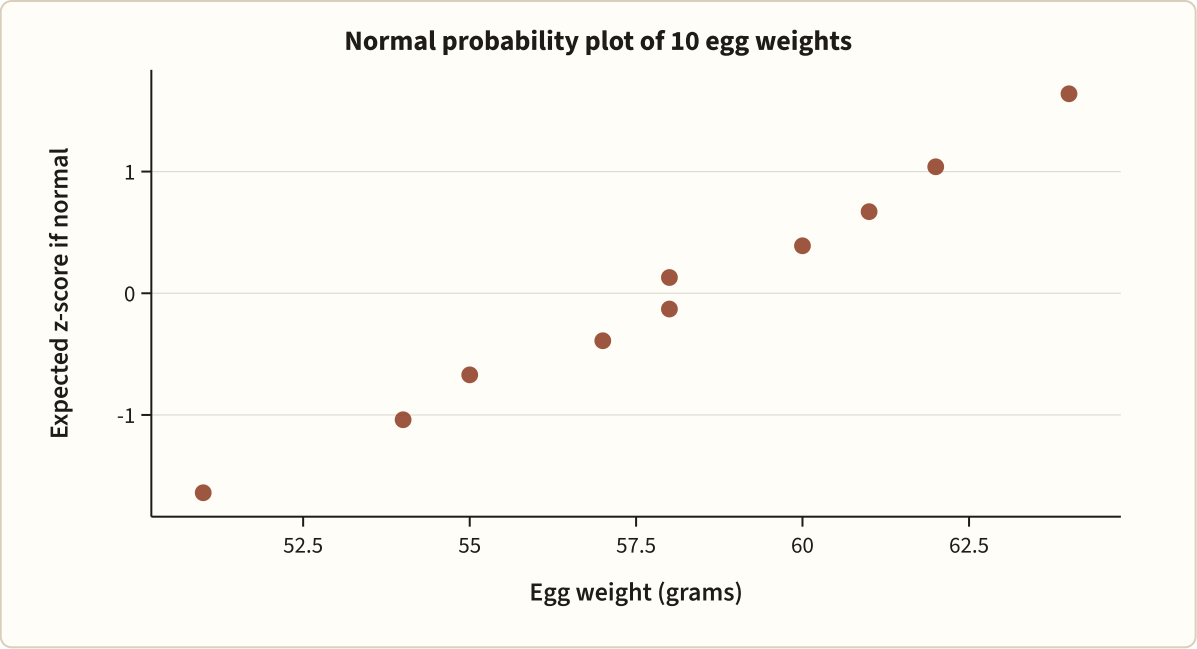
Heights of 20 corn stalks

7	0 4
7	6 7 8 9
8	0 0 1 2 2 3 4
8	5 5 6 8 9
9	0 1
9	

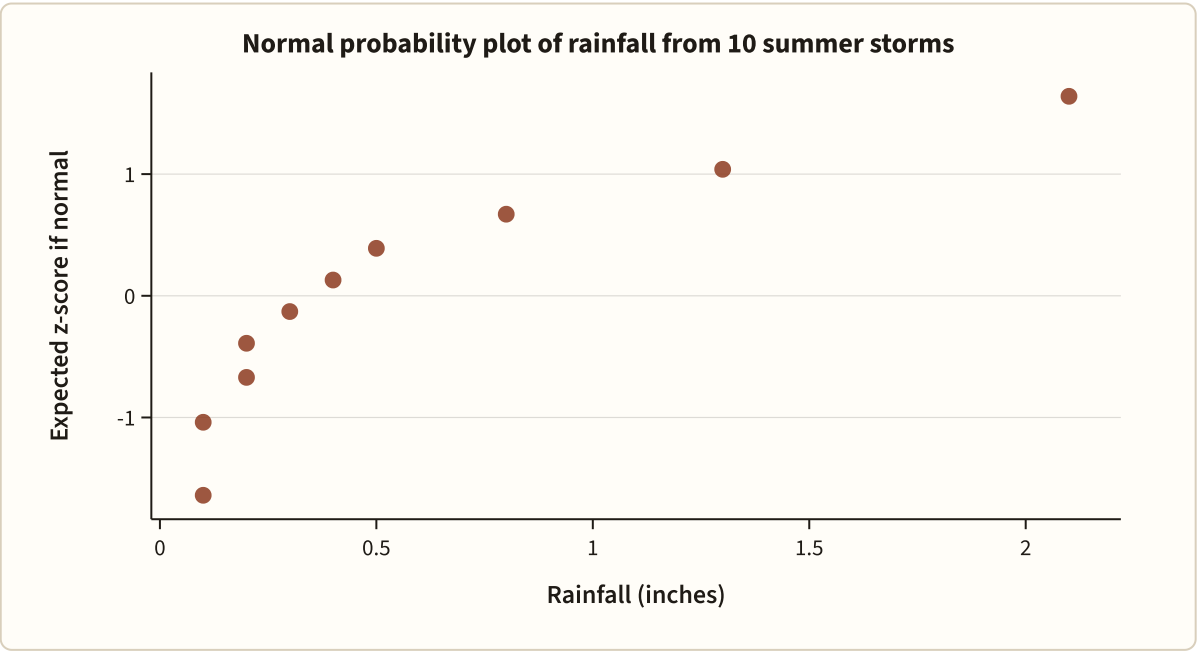
Key: 7 | 4 means 74 inches

Method 2: Check the 68–95–99.7 rule. Compute $\bar{x}$ and s, then count the values in $\bar{x} \pm s$, $\bar{x} \pm 2s$ and $\bar{x} \pm 3s$. If the percents are near 68%, 95% and 99.7%, a normal model fits well. For the corn, $\bar{x} = 82.0$ and $s \approx 5.49$ inches, and the percents are 65%, 95% and 100%: very close. Small data sets will never match exactly; with 20 values, 65% instead of 68% is less than one stalk's difference.

Method 3: The normal probability plot. Put the data in order and give each value the z-score it would have if the data were perfectly normal (technology does this). Then plot each data value against its expected z-score. Why should normal data make a straight line? If the data are normal, each value is about $x = \mu + z\sigma$, and that is the equation of a straight line. So the more nearly straight the plot, the more nearly normal the data. The plot of 10 egg weights below is close to a straight line.



When the data are skewed, the plot bends. In right-skewed data the small values are crowded together and the large values are stretched far out, much farther than a normal curve would place them. So the plot climbs steeply at the left, where many values share a small range, and flattens at the right, where a few values are spread over a long distance. The plot of rainfall from 10 summer storms bends this way. Left-skewed data bend the opposite way (flat at the left, steep at the right), and an outlier shows up as a single point far from the line at one end.



No real data set is exactly normal, so the question is always "close enough?" Look at all three checks together and use judgment. Mild bumps in a small data set are normal (in both senses); strong skewness or outliers mean you should not use the normal table for these data, and you should describe them with the median and IQR instead (Lesson 2.5).

KEY FORMULAS

$$\text{normal data: } x \approx \mu + z\sigma \;\Rightarrow\; \text{a straight-line normal probability plot}$$

$$\text{check: about } 68\%,\ 95\%,\ 99.7\% \text{ of values in } \bar{x} \pm s,\ \bar{x} \pm 2s,\ \bar{x} \pm 3s$$

Practice habit. Always plot the data first. Before any normal calculation, look at a graph of the data and ask "is this roughly symmetric and single-peaked, with no outliers?" If not, a normal model does not fit.

WORKED EXAMPLES

Example 1. The heights of 20 corn stalks (inches) are shown in the stemplot: 70, 74, 76, 77, 78, 79, 80, 80, 81, 82, 82, 83, 84, 85, 85, 86, 88, 89, 90, 91. Their mean is 82.0 inches and their standard deviation is 5.49 inches. What percent of the heights lie within 1, 2 and 3 standard deviations of the mean? Do the data follow the 68–95–99.7 rule?

Heights of 20 corn stalks

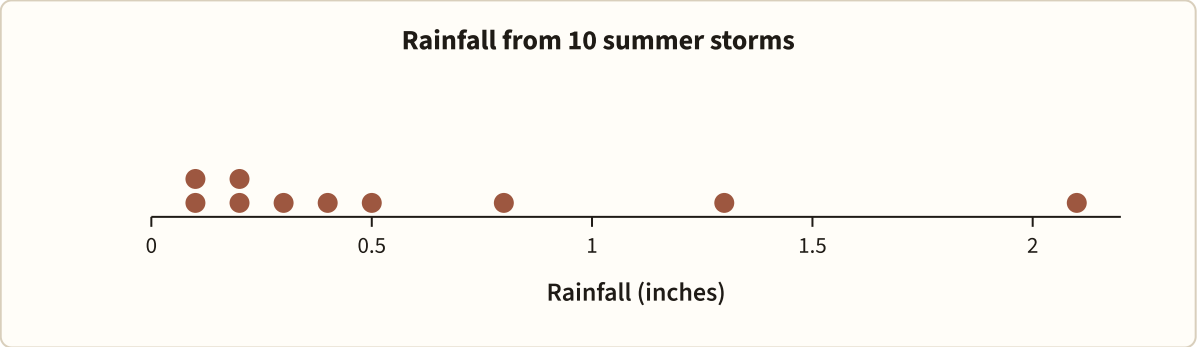
7	0 4
7	6 7 8 9
8	0 0 1 2 2 3 4
8	5 5 6 8 9
9	0 1
9	

Key: 7 | 4 means 74 inches

1. Within 1 standard deviation: 82.0 ± 5.49 gives 76.51 to 87.49.
2. Count the heights from 77 through 86: 77, 78, 79, 80, 80, 81, 82, 82, 83, 84, 85, 85, 86. That is 13 of 20, or 65%.
3. Within 2 standard deviations: $82.0 \pm 2(5.49) = 82.0 \pm 10.98$ gives 71.02 to 92.98. Every height except 70 is inside: 19 of 20, or 95%.
4. Within 3 standard deviations: $82.0 \pm 3(5.49) = 82.0 \pm 16.47$ gives 65.53 to 98.47. All 20 heights, or 100%.
5. Compare with 68%, 95% and 99.7%: 65%, 95% and 100% are very close. The stemplot is single-peaked and roughly symmetric with no outliers.
6. Conclusion: the corn heights are roughly normal, so a normal model is reasonable for these stalks.

Answer: 65%, 95% and 100%; these are close to 68%, 95% and 99.7%, so the heights are roughly normal.

Example 2. A farmer recorded the rainfall from 10 summer storms, in inches: 0.1, 0.1, 0.2, 0.2, 0.3, 0.4, 0.5, 0.8, 1.3, 2.1. The mean is 0.60 inch and the standard deviation is 0.64 inch. Using the dotplot, the normal probability plot, the mean and median, and the 68–95–99.7 rule, decide whether it is reasonable to use normal calculations for these data. Answer yes or no.

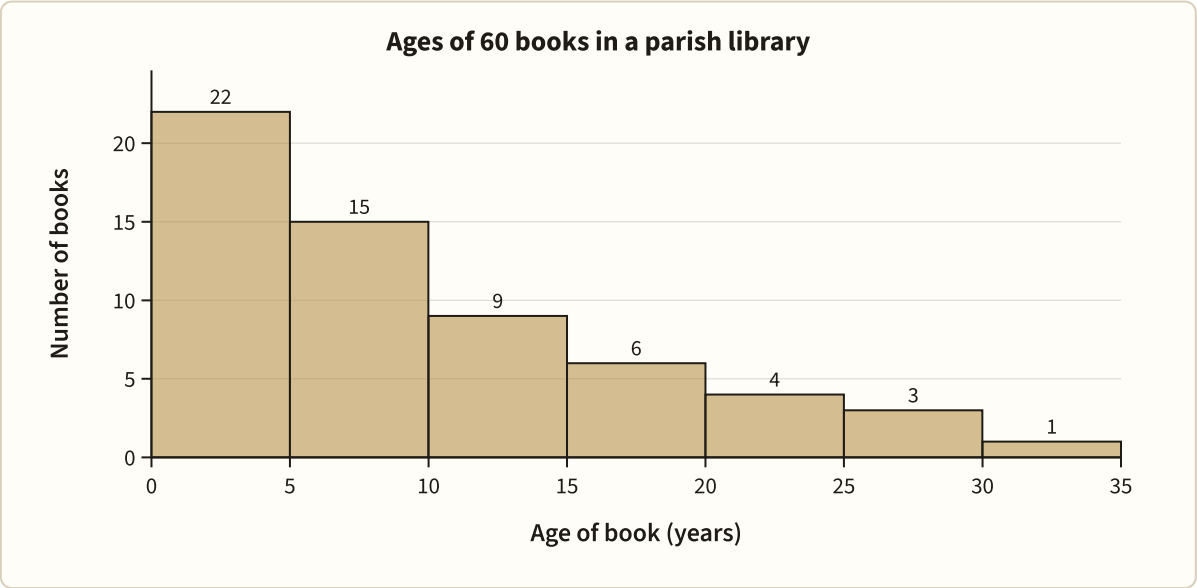


1. Shape: the dotplot shows the values crowded near 0 with a long tail toward 2.1 inches.
2. Mean versus median: the median is $\frac{0.3 + 0.4}{2} = 0.35$ inch, well below the mean of 0.60 inch, so the large storms pull the mean up: skewed right.
3. 68–95–99.7 check: 0.60 ± 0.64 runs from -0.04 to 1.24; 8 of the 10 storms (80%) are inside, far more than 68%.
4. Impossible values: $\bar{x} - s = -0.04$ is already below 0. A normal model, $z = \frac{0 - 0.60}{0.64} \approx -0.94$, would say about 17% of storms drop a negative amount of rain, which cannot happen.
5. Normal probability plot (shown in the lesson above for these same storms): the points climb steeply at the left and flatten at the right, a clear curve, which is the pattern of right skewness.
6. Conclusion: no. Storm rainfall here is strongly skewed right, so normal calculations would give misleading answers; describe these data with the median and IQR instead.

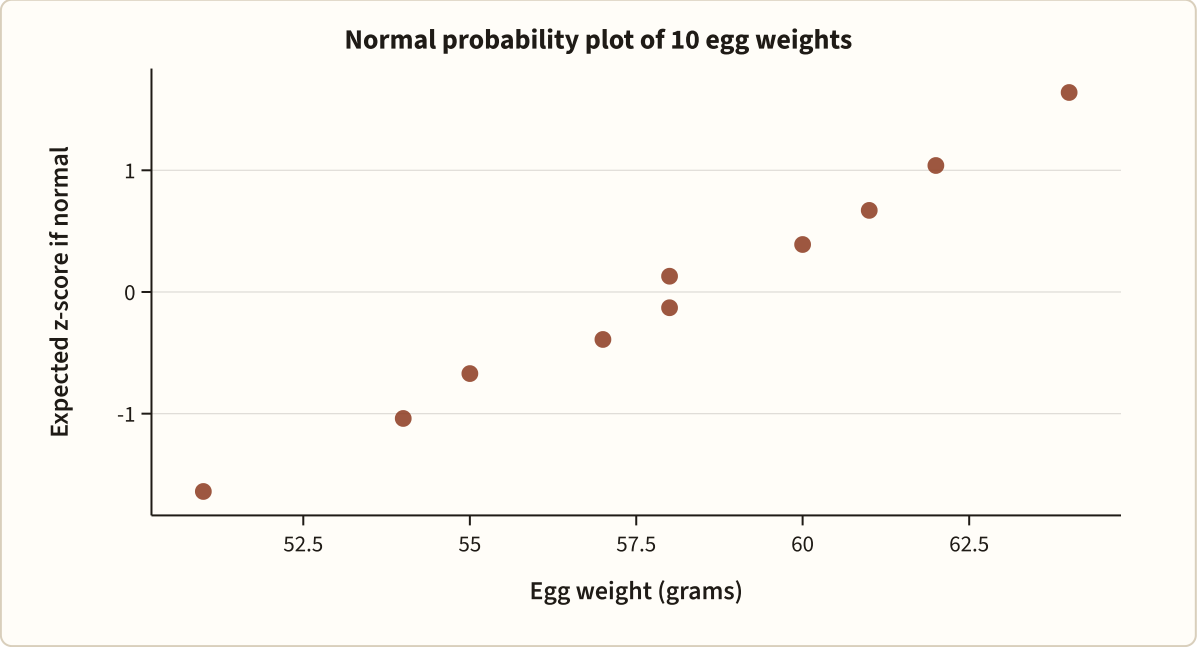
Answer: No. The data are strongly skewed right (median 0.35 < mean 0.60, 80% within one standard deviation, a curved normal probability plot), so a normal model does not fit.

PRACTICE

1. **EASY** The histogram shows the ages of 60 books in a parish library. Describe the shape. Would a normal model fit these data?



2. **EASY** Look at the normal probability plot of 10 egg weights. Are the egg weights roughly normal? Answer yes or no and give the reason.



3. **EASY** In a data set of 50 daily wind readings, 34 values lie within one standard deviation of the mean. What percent is that, and how does it compare with a normal distribution?

4. **EASY** Explain why you should check whether data are roughly normal before using the standard normal table to answer questions about them.

5. **MEDIUM** Use the stemplot of 20 corn stalk heights. Find the median, and compare it with the mean of 82.0 inches. What does the comparison suggest about the shape?

Heights of 20 corn stalks

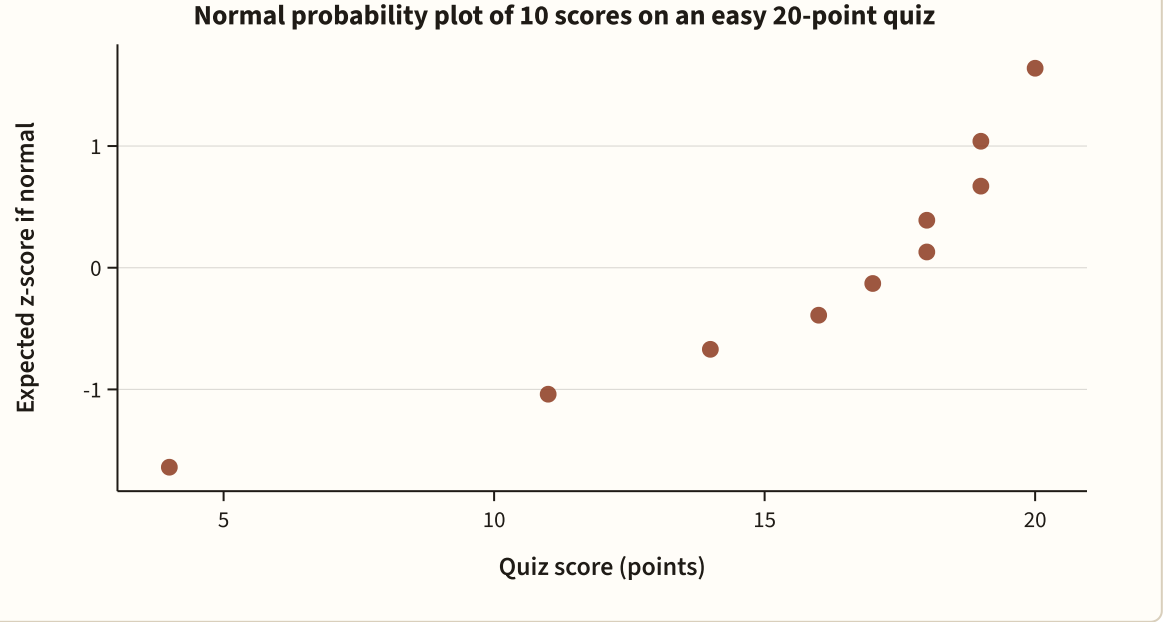
7	0 4
7	6 7 8 9
8	0 0 1 2 2 3 4
8	5 5 6 8 9
9	0 1
9	

Key: 7 | 4 means 74 inches

6. **MEDIUM** The rainfall (inches) from 20 summer storms was 0.1, 0.1, 0.2, 0.2, 0.2, 0.3, 0.3, 0.3, 0.4, 0.4, 0.5, 0.5, 0.6, 0.7, 0.8, 1.0, 1.2, 1.5, 1.9, 2.6, with $\bar{x} = 0.69$ and $s = 0.66$ inch. (a) What percent of the storms lie within one standard deviation of the mean? (b) Do the data follow the 68–95–99.7 rule well? Answer yes or no.

7. **MEDIUM** Twelve choir members practiced at home last week for a mean of 45 minutes, with standard deviation 30 minutes. (a) If the times were $N(45, 30)$, what proportion would be less than 0 minutes? (b) Explain why the times cannot be roughly normal.

8. **MEDIUM** The normal probability plot shows 10 scores on an easy 20-point quiz. Describe the shape of the distribution.



9. **MEDIUM** A rancher weighed 40 calves: $\bar{x} = 82$ pounds and $s = 7$ pounds. 27 calves were within 1 standard deviation of the mean, 38 within 2, and all 40 within 3. (a) Give the three percents. (b) Is a normal model reasonable? Answer yes or no.

10. **HARD** Ten eggs weigh 51, 54, 55, 57, 58, 58, 60, 61, 62 and 64 grams. (a) Find the mean. (b) Find the standard deviation, to one decimal place. (c) What percent of the eggs are within one standard deviation of the mean?

11. **HARD** A weather station recorded daily wind speeds for a year: mean 12.4 mph, median 9.0 mph, minimum 2 mph and maximum 41 mph. Is a normal model reasonable for these wind speeds? Answer yes or no and explain.

12. **HARD** For the 20 corn heights ($\bar{x} = 82.0$, $s = 5.49$ inches): (a) What percent are below $\bar{x} - s$? (b) What percent are above $\bar{x} + s$? (c) A normal distribution puts 16% in each of these tails. Comment.

Heights of 20 corn stalks

7	0 4
7	6 7 8 9
8	0 0 1 2 2 3 4
8	5 5 6 8 9
9	0 1
9	

Key: 7 | 4 means 74 inches

13. **CHALLENGE** A large flock's lamb weights are roughly normal, with median 40 pounds and third quartile 44 pounds. (a) In a normal distribution Q3 is the 75th percentile, at $z = 0.674$. Estimate σ to the nearest whole pound. (b) Using $N(40, \sigma)$ with your rounded σ, estimate the proportion of lambs heavier than 52 pounds.

COMMON MISTAKES

- Skipping the check and using the normal table on skewed data. Always plot the data first, and look for skewness and outliers.
- Expecting the percents to be exactly 68, 95 and 99.7. Real data, especially small samples, only come close; judge "close enough" together with the graph.
- Reading a normal probability plot as a scatterplot of two variables. It asks only one question: is it straight? A curve means skewness; a lone point far from the line means an outlier.
- Missing the quick signs of skewness: a mean much larger (or smaller) than the median, or a smallest possible value (such as 0) that lies less than two standard deviations below the mean.

ANSWERS TO PRACTICE

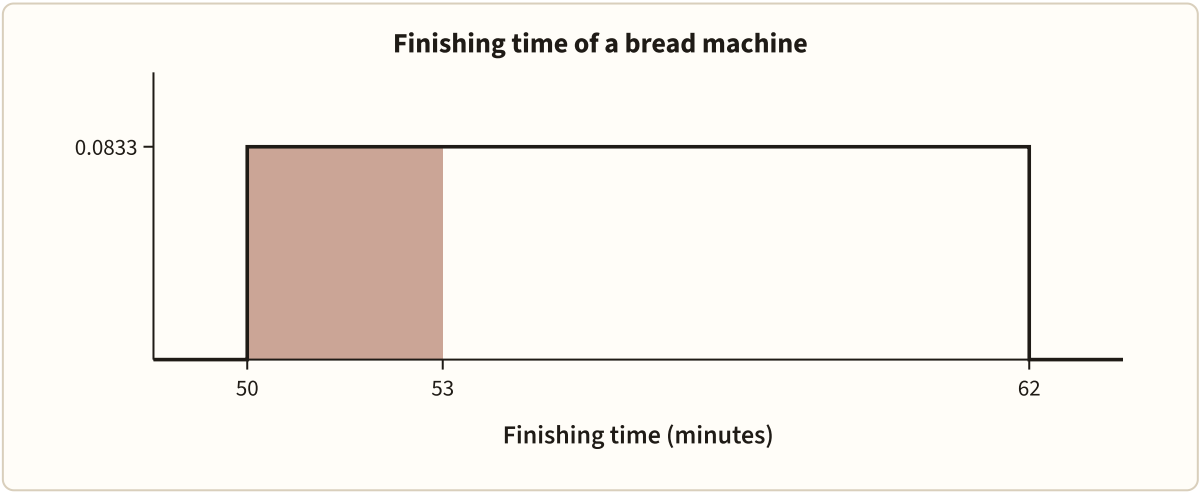
1. Skewed right: most books are less than 10 years old and a few are much older. A normal model would not fit.
2. Yes. The points lie close to a straight line, with no curve and no point far off the line.
3. $\frac{34}{50} = 68\%$, exactly what the 68–95–99.7 rule predicts for a normal distribution.

4. The table gives areas under a normal curve. If the data are skewed or have outliers, the true proportions can be very different from the normal areas, so the answers would be wrong. A normal model should be used only when it fits.
5. The 10th and 11th heights are both 82, so the median is 82 inches. It equals the mean, which fits a symmetric distribution.
6. (a) 0.69 ± 0.66 runs from 0.03 to 1.35; 17 of 20 storms are inside: 85%. (b) No: 85% is far more than 68%, and $\bar{x} - 2s$ is negative. The data are skewed right.
7. (a) $z = \dfrac{0 - 45}{30} = -1.50$; proportion 0.0668. (b) No one can practice a negative number of minutes, yet a normal model puts almost 7% of times below 0. The minimum possible value, 0, is only 1.5 standard deviations below the mean, so the times must be skewed right.
8. Skewed left. Most scores are crowded near the top (16 to 20), so the plot is steep at the right, while the low score of 4 sits far to the left of the others and the plot is flat there.
9. (a) $\dfrac{27}{40} = 67.5\%$, $\dfrac{38}{40} = 95\%$, $\dfrac{40}{40} = 100\%$. (b) Yes; these are very close to 68%, 95% and 99.7%.
10. (a) $\bar{x} = \dfrac{580}{10} = 58.0$ grams. (b) Squared deviations 49, 16, 9, 1, 0, 0, 4, 9, 16, 36 add to 140; $s = \sqrt{\dfrac{140}{9}} \approx 3.9$ grams. (c) 58.0 ± 3.9 runs from 54.1 to 61.9: 55, 57, 58, 58, 60, 61, so 6 of 10 = 60%. With only 10 eggs this is close enough to 68%; the normal plot of these eggs is straight.
11. No. The mean is much larger than the median, and the maximum is 28.6 mph above the mean while the minimum is only 10.4 mph below it. Both signs point to a long right tail: the data are skewed right.
12. (a) Below 76.51: 70, 74, 76, so 3 of 20 = 15%. (b) Above 87.49: 88, 89, 90, 91, so 4 of 20 = 20%. (c) Both are close to 16%; with only 20 stalks one stalk is 5%, so the difference is less than one stalk in each tail. The data are roughly normal.
13. (a) $0.674 = \dfrac{44 - 40}{\sigma}$, so $\sigma = \dfrac{4}{0.674} \approx 5.93 \approx 6$ pounds. (b) $z = \dfrac{52 - 40}{6} = 2.00$; proportion $1 - 0.9772 = 0.0228$.

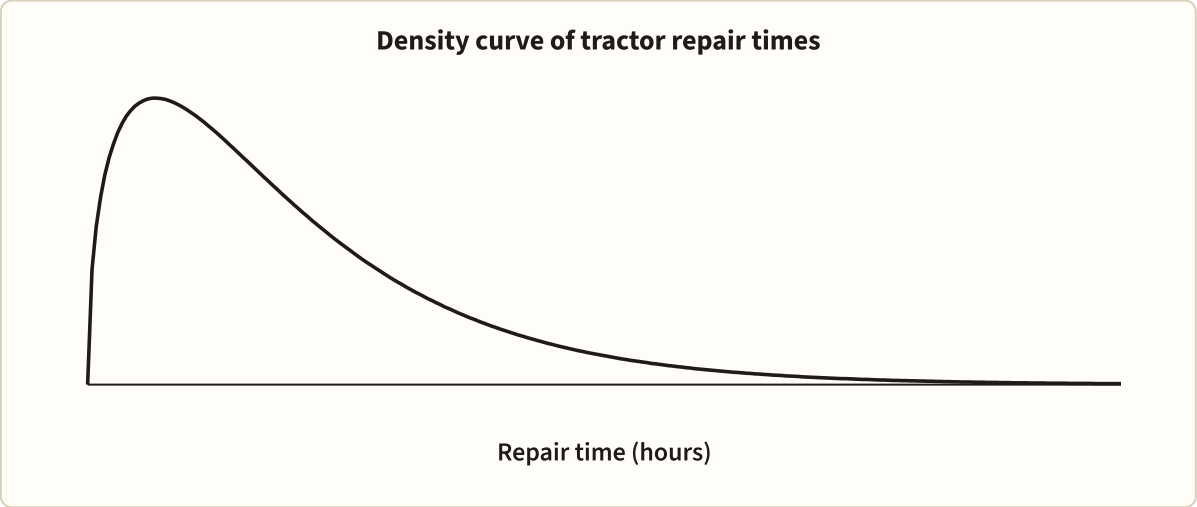
Unit 3 test

You have 55 minutes. You may use a scientific or graphing calculator and the standard normal table at the back of the course. Show your work: write the formula, the numbers you substitute and the result, and sketch and shade a normal curve for every normal calculation. Keep full precision until the last step. Unless a problem says otherwise, give proportions to 4 decimal places and z-scores to 2 decimal places; answers from the table may differ from technology in the last digit. State answers in context with units.

1. **6 PTS** The weights (in pounds) of 15 pumpkins entered at a county fair were 8, 11, 12, 14, 15, 15, 17, 18, 20, 21, 23, 24, 26, 29, 35. What percentile is the 21-pound pumpkin?
2. **6 PTS** Hannah's beehive produced 72 pounds of honey; hives in her region average 60 pounds with standard deviation 8 pounds. Peter's tomato plant gave 18 pounds of tomatoes; plants in his garden average 14 pounds with standard deviation 2.5 pounds. Find both z-scores. Whose harvest was more remarkable compared with its own group?
3. **6 PTS** During a meteor shower, an observer's hourly meteor counts have mean 48 and standard deviation 8. One cloudy hour had a z-score of -2.25. How many meteors were counted that hour?
4. **7 PTS** A bread machine finishes a loaf at a time that is uniformly distributed between 50 and 62 minutes after it starts. (a) What is the height of the density curve? Round to 4 decimal places. (b) What proportion of loaves finish in under 53 minutes?



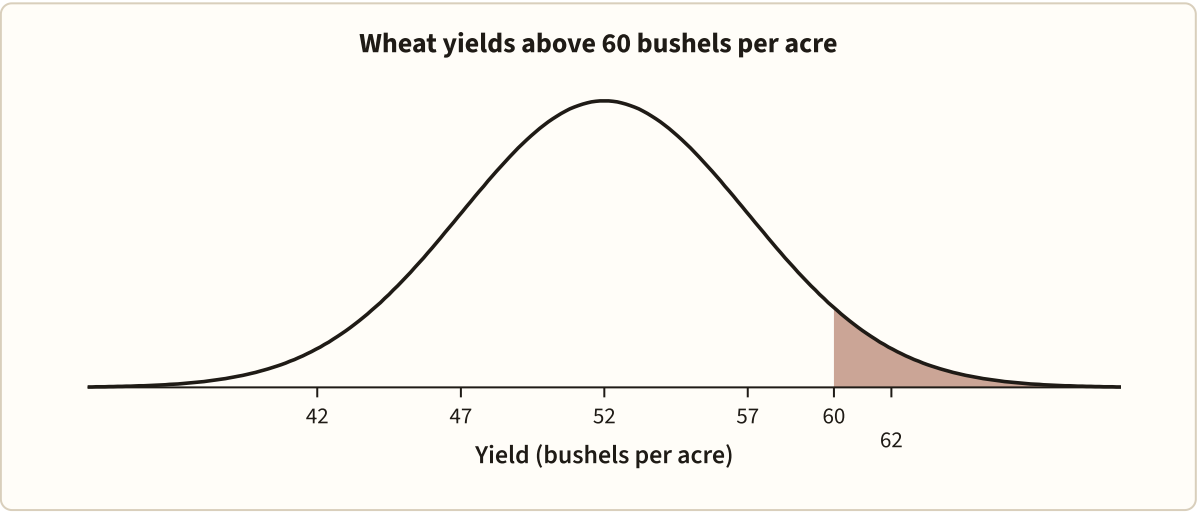
5. **6 PTS** The density curve shows the times needed to repair a farm's tractors. Which is larger, the mean repair time or the median repair time? Explain why, using the ideas of a balance point and equal areas.



6. **7 PTS** The weights of watermelons from a farm are approximately $N(20, 3)$ pounds. (a) Between what two weights do the middle 99.7% of the melons fall? (b) About what percent of melons weigh more than 26 pounds?

7. **6 PTS** The length of pregnancy for the ewes in a flock is approximately $N(147, 2)$ days. Use the 68–95–99.7 rule to find about what percent of pregnancies last between 149 and 153 days.

8. **8 PTS** Yields of wheat on a county's test plots are approximately $N(52, 5)$ bushels per acre. What proportion of plots yield more than 60 bushels per acre?



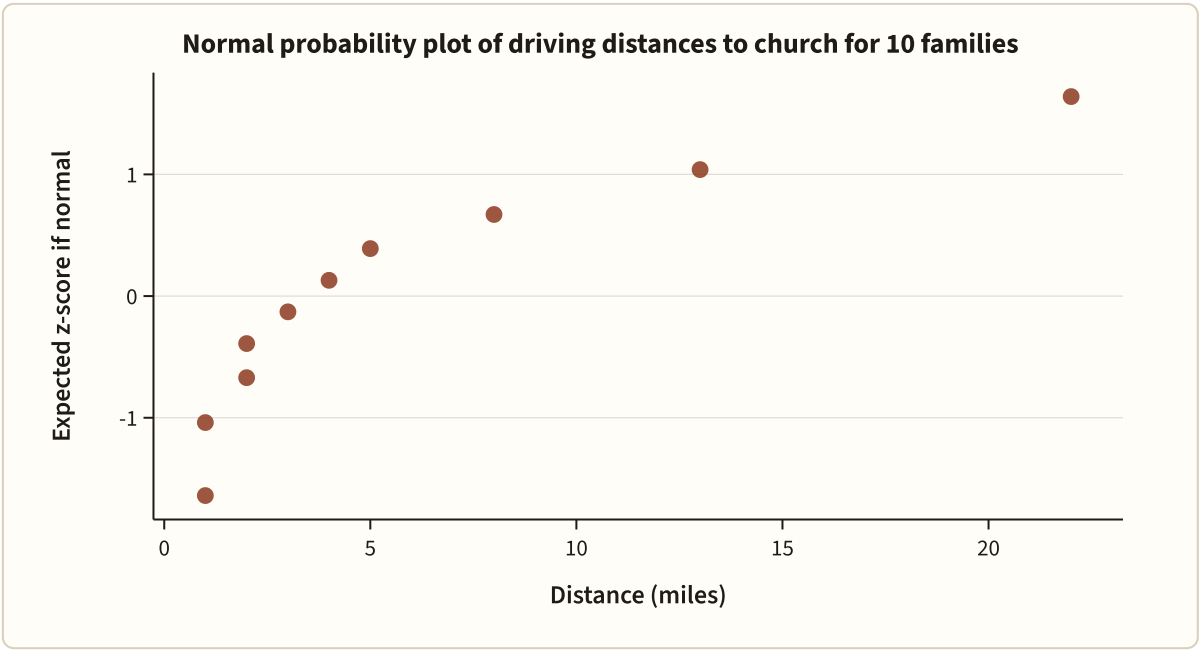
9. **8 PTS** A fencing crew buys posts whose lengths are approximately $N(72.0, 0.5)$ inches. Posts from 71.2 to 72.6 inches long are usable. What proportion of the posts are usable?

10. **8 PTS** Weaning weights of a ranch's Angus calves are approximately $N(550, 40)$ pounds. Calves in the heaviest 5% earn a premium price. What is the minimum weight for the premium, to the nearest pound?

11. **8 PTS** A mill fills 25-pound sacks of flour. The fill amounts are approximately normal with standard deviation 0.15 pound. At what mean should the mill set its machine so that only 1% of sacks contain less than 25 pounds? Round to 2 decimal places.

12. **7 PTS** A 4-H member weighed 25 rabbits. Of the 25 weights, 18 were within 1 standard deviation of the mean, 24 within 2, and all 25 within 3. (a) Give the three percents. (b) Is a normal model reasonable for these weights? Answer yes or no.

13. **6 PTS** The normal probability plot shows the distances 10 families drive to church each Sunday. (a) Describe the shape of the distribution. (b) Is it reasonable to use normal calculations for these distances? Answer yes or no.



14. **5 PTS** Explain, using the 68–95–99.7 rule, why a value one standard deviation above the mean of a normal distribution is at about the 84th percentile.

15. **6 PTS** Peach weights in a county's orchards are approximately normal. 10% of the peaches weigh less than 120 grams and 5% weigh more than 200 grams. Find μ and σ, each to the nearest gram.

UNIT 3 TEST: ANSWER KEY

1. 9 pumpkins weigh less than 21 pounds; $\frac{9}{15} = 60\%$, so the 60th percentile.
2. Honey: $z = \frac{72 - 60}{8} = 1.50$. Tomatoes: $z = \frac{18 - 14}{2.5} = 1.60$. Peter's tomato plant was slightly more remarkable: 1.60 standard deviations above its mean, versus 1.50.
3. $x = 48 + (-2.25)(8) = 48 - 18 = 30$ meteors.
4. (a) $\frac{1}{62 - 50} = \frac{1}{12} \approx 0.0833$. (b) $(53 - 50) \times \frac{1}{12} = \frac{3}{12} = 0.25$.
5. The mean is larger. The curve is skewed right, so a little area lies far out in the long right tail. Area far from the center has great leverage, so it pulls the balance point (the mean) to the right, while the median only splits the area in half and is not pulled as far.
6. (a) $20 \pm 3(3)$: from 11 to 29 pounds. (b) 26 is $\mu + 2\sigma$, so about $\frac{100\% - 95\%}{2} = 2.5\%$.
7. 149 is $\mu + 1\sigma$ and 153 is $\mu + 3\sigma$. From the mean to $+3\sigma$ is 49.85%; from the mean to $+1\sigma$ is 34%. Between: $49.85\% - 34\% = 15.85\%$.
8. $z = \frac{60 - 52}{5} = 1.60$; area to the right $= 1 - 0.9452 = 0.0548$. About 5.5% of plots yield more than 60 bushels per acre.
9. $z = \frac{71.2 - 72.0}{0.5} = -1.60$, left area 0.0548; $z = \frac{72.6 - 72.0}{0.5} = 1.20$, left area 0.8849. Between: $0.8849 - 0.0548 = 0.8301$.
10. Left area 0.95: $z = 1.645$. $x = 550 + 1.645(40) = 550 + 65.8 \approx 616$ pounds.
11. Left area 0.01: $z \approx -2.326$. $\mu = x - z\sigma = 25 - (-2.326)(0.15) = 25 + 0.349 \approx 25.35$ pounds.
12. (a) $\frac{18}{25} = 72\%$, $\frac{24}{25} = 96\%$, $\frac{25}{25} = 100\%$. (b) Yes; these are close to 68%, 95% and 99.7%.
13. (a) Skewed right: the plot is steep at the left, where many families live close by, and flattens at the right, where a few live far away. (b) No; the plot is clearly curved.
14. Half of the distribution (50%) lies below the mean. Between the mean and one standard deviation above it lies half of 68%, which is 34%. So about $50\% + 34\% = 84\%$ of the values are less than $\mu + \sigma$: it is at about the 84th percentile.
15. Left area 0.10: $z = -1.282$, so $\mu - 1.282\sigma = 120$. Left area 0.95: $z = 1.645$, so $\mu + 1.645\sigma = 200$. Subtract: $2.927\sigma = 80$, $\sigma \approx 27.33$; $\mu = 120 + 1.282(27.33) \approx 155.0$. So $\mu \approx 155$ grams and $\sigma \approx 27$ grams.

UNIT 4

Scatterplots, Correlation and Regression

Units 1–3 studied one variable at a time. This unit studies relationships between two quantitative variables measured on the same individuals: the age of a tree and its yield, the rainfall of a season and its corn harvest, a calf's age and its weight. Students learn to picture a relationship with a scatterplot, to measure its strength with the correlation r, and to summarize it with the least-squares regression line for prediction.

The emphasis is on reasoning, not only calculation. Students see why r is built from products of z-scores (Unit 3), why the regression line passes through the point of means and is chosen by least squares, what r^2 and the residuals tell us about how well a line fits, and why even a strong correlation does not prove cause. The unit builds directly on the mean and standard deviation (Unit 2) and z-scores (Unit 3), and the question of cause and effect leads straight into Unit 5, where students learn how randomized experiments can show it.

PREREQUISITES

- Compute a mean and a sample standard deviation s, by hand for small data sets and with a calculator (Lessons 2.1 and 2.4).
- Compute and interpret a z-score (Lesson 3.1).
- Describe a distribution's shape, center, spread and outliers in context (Lesson 1.5).
- Plot points on the coordinate plane and interpret the slope and y-intercept of y = mx + b (Algebra I).
- Use a calculator's two-variable statistics mode to enter paired data.

MASTERY CHECKLIST

☐ I can identify explanatory and response variables and make a scatterplot with labeled axes.
☐ I can describe a scatterplot by direction, form, strength and unusual points, in context.
☐ I can explain why the products of z-scores make r positive for an uphill pattern and negative for a downhill pattern.
☐ I can compute r by hand for a small data set and with a calculator, and interpret it.
☐ I can state the properties of r and explain why r near 0 does not mean "no relationship."
☐ I can find the least-squares regression line from summary statistics or raw data and explain why it passes through $(\bar{x}, \bar{y})$.
☐ I can interpret the slope and y-intercept in context and recognize the danger of

extrapolation.

☐ I can explain regression toward the mean using $\hat{z}_y = r z_x$.

☐ I can compute residuals, read a residual plot, and decide whether a line is a suitable model.

☐ I can interpret s and r^2 and read regression output from a calculator or computer.

☐ I can recognize outliers and influential points and check influence by refitting without the point.

☐ I can explain why association does not imply causation and name possible lurking variables.

LESSON 4.1 · 2 SESSIONS OF 45 MINUTES

Scatterplots and Association

OBJECTIVES

- Identify the explanatory and response variables in a relationship between two quantitative variables.
- Make a scatterplot with the explanatory variable on the horizontal axis and labeled axes.
- Describe a scatterplot by its direction, form, strength and unusual points, in context.
- Explain why an association does not prove cause and effect, and name a possible lurking variable.

VIDEO

WATCH

Scatter Plots (1 of 2: Introduction to Bivariate Data)

Eddie Woo · 8:06

When to watch: Before Example 1: explanatory and response variables and bivariate data.

https://www.youtube.com/watch?v=VdDvFRg6gzA

BACKUP

Scatter Plots (2 of 2: Visual introduction)

Eddie Woo · 8:33

When to watch: Before Example 2: describing direction, form and strength from a scatterplot.

https://www.youtube.com/watch?v=8W62pHb5SIc

> **Note:** Eddie Woo says “independent” and “dependent” variable for explanatory and response.

KEY TERMS

Explanatory variable	The variable we think helps explain or predict changes in another variable; it goes on the horizontal (x) axis.
Response variable	The variable whose changes we want to understand or predict; it goes on the vertical (y) axis.
Scatterplot	A graph of two quantitative variables measured on the same individuals; each individual is one point (x, y).
Positive association	Larger values of one variable tend to go with larger values of the other, so the points slope uphill from left to right.
Negative association	Larger values of one variable tend to go with smaller values of the other, so the points slope downhill from left to right.

Lurking variable A variable that is not among the variables studied but affects both of them, so it can create an association between them or hide one.

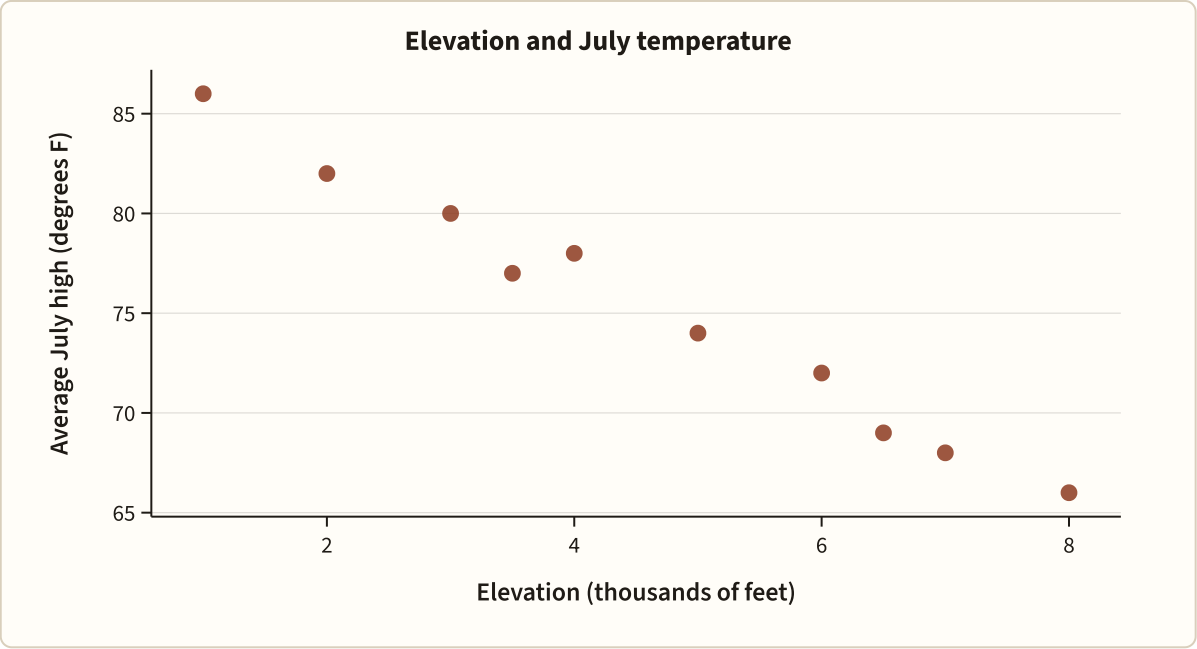
CORE IDEAS

In Units 1–3 you studied one variable at a time. Many of the most useful questions, though, are about two quantitative variables measured on the same individuals. Do older apple trees give more fruit? Is it cooler at higher elevations? Do warmer days sell more lemonade? When we think one variable helps explain or predict the other, we call it the **explanatory variable** (x) and the other the **response variable** (y). Sometimes neither variable explains the other (a person's arm span and height); then either one may go on the x-axis.

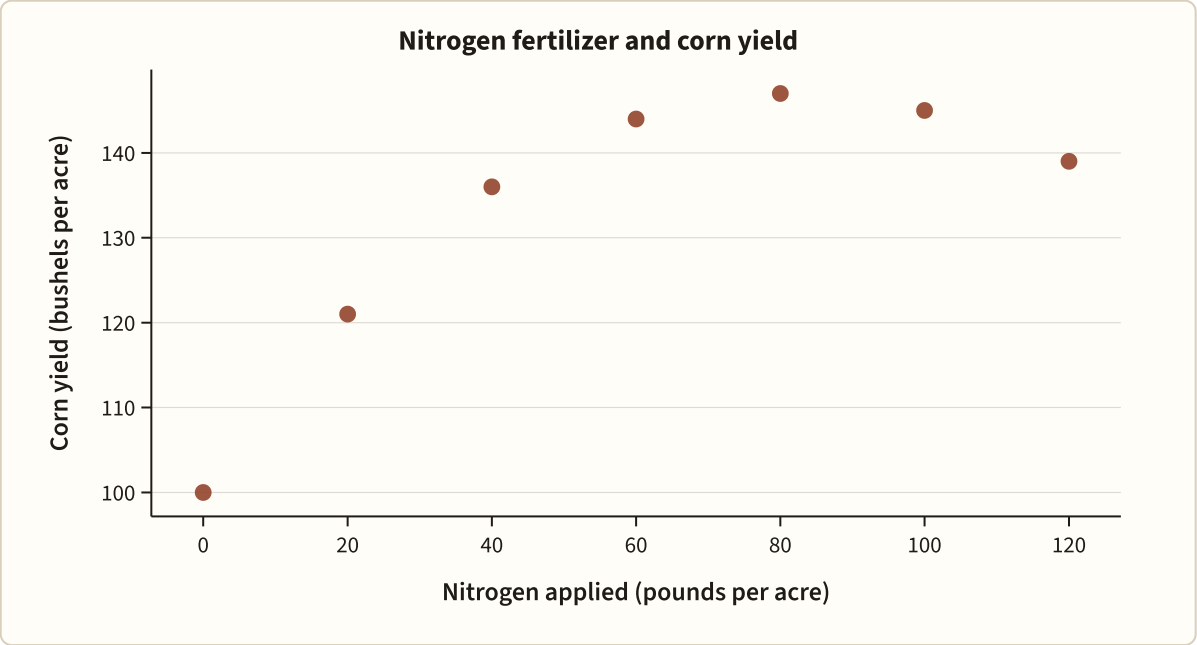
A **scatterplot** shows the relationship. Each individual becomes one point: its x-value tells how far across, its y-value how far up. Put the explanatory variable on the horizontal axis, choose scales that fit the data (the axes need not start at 0), and label both axes with the variable and its units. Below, each point is one tree in the Miller family's orchard.

To describe a scatterplot, look for four things and always say them in context. **Direction**: do the points rise from left to right (positive association) or fall (negative association)? **Form**: do they follow a straight line (linear), a curve, or form clusters? **Strength**: how closely do the points follow that form? Tightly packed points show a strong relationship; a loose cloud shows a weak one. **Unusual points**: are there outliers that fall far from the overall pattern? For the orchard: *there is a strong, positive, linear association between a tree's age and its yield: older trees tend to produce more bushels, about 1.2 more bushels for each extra year, with no outliers.*

Here is a negative association. Each point is a weather station in the mountains; as elevation rises, the average July high temperature falls in a fairly straight, strong pattern. This makes sense: air cools as it rises and expands.



Not every relationship is a straight line. Below, corn yield rises quickly as more nitrogen fertilizer is added, then levels off and even drops when too much is used. The form is **curved**. Later in this unit you will fit straight lines to data, so it matters to notice a curve first: a straight line would describe this relationship badly.



An association, however strong, does not by itself show that one variable **causes** the other. A **lurking variable** may drive both. Among elementary-school children, shoe size and reading level are positively associated, but big feet do not make good readers: age is the lurking variable, and older children have both bigger feet and more reading practice. Whenever you see an association in data that were simply observed, ask: *what else could explain this?* In Unit 5 you will see how a well-designed experiment can show cause and effect.

Practice habit. Always plot the data first: make a scatterplot before you compute anything, and describe it with direction, form, strength and unusual points, in context.

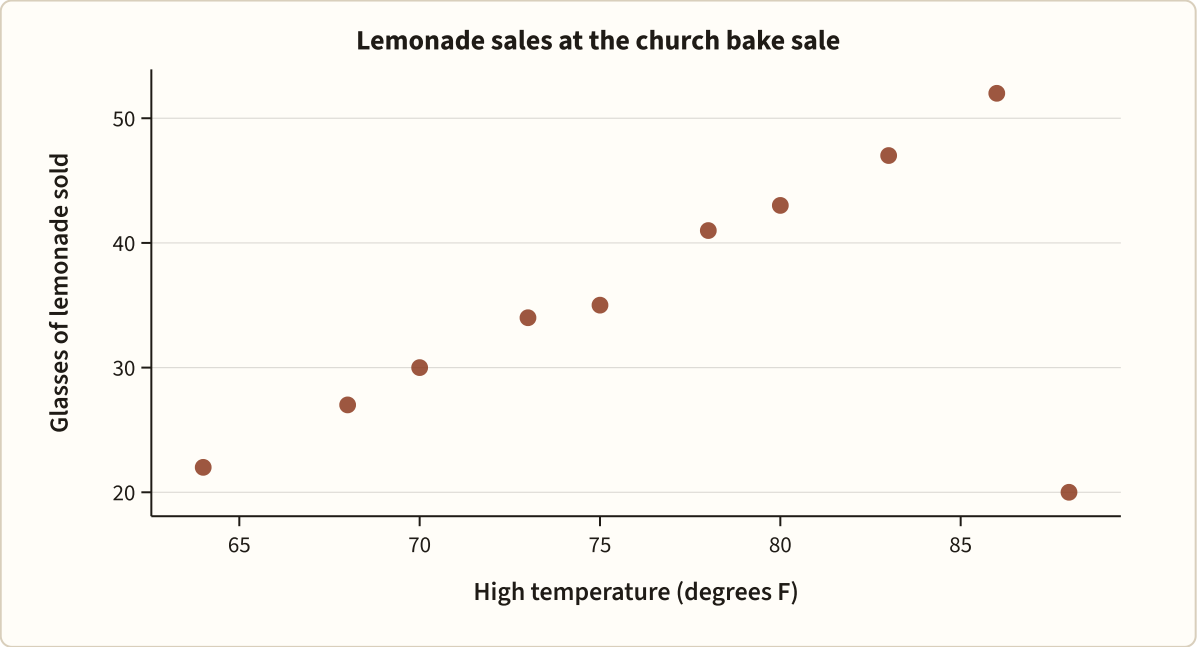
WORKED EXAMPLES

Example 1. For each pair of variables, name the explanatory variable and the response variable, or say that either could play either role. (a) The age of a calf in days and its weight in pounds. (b) The hours of sunlight a sunflower receives each day and its height after eight weeks. (c) A person's arm span and height.

1. Ask of each pair: does one variable help explain or predict the other?
2. (a) A calf gains weight as it grows older; age helps explain weight, not the other way around. Explanatory: age. Response: weight.
3. (b) Sunlight feeds the plant's growth; height cannot change the sunlight. Explanatory: hours of sunlight. Response: height.
4. (c) Neither variable causes or explains the other; both reflect how large a person is. We could predict either one from the other, so either may be explanatory.

Answer: (a) Explanatory: age; response: weight. (b) Explanatory: hours of sunlight; response: height. (c) Either variable may be explanatory; we simply study how the two vary together.

Example 2. On ten Saturdays the youth group at St. Andrew's sold lemonade at the church bake sale. They recorded the high temperature (°F) and the number of glasses sold. Describe the relationship shown in the scatterplot, in context.



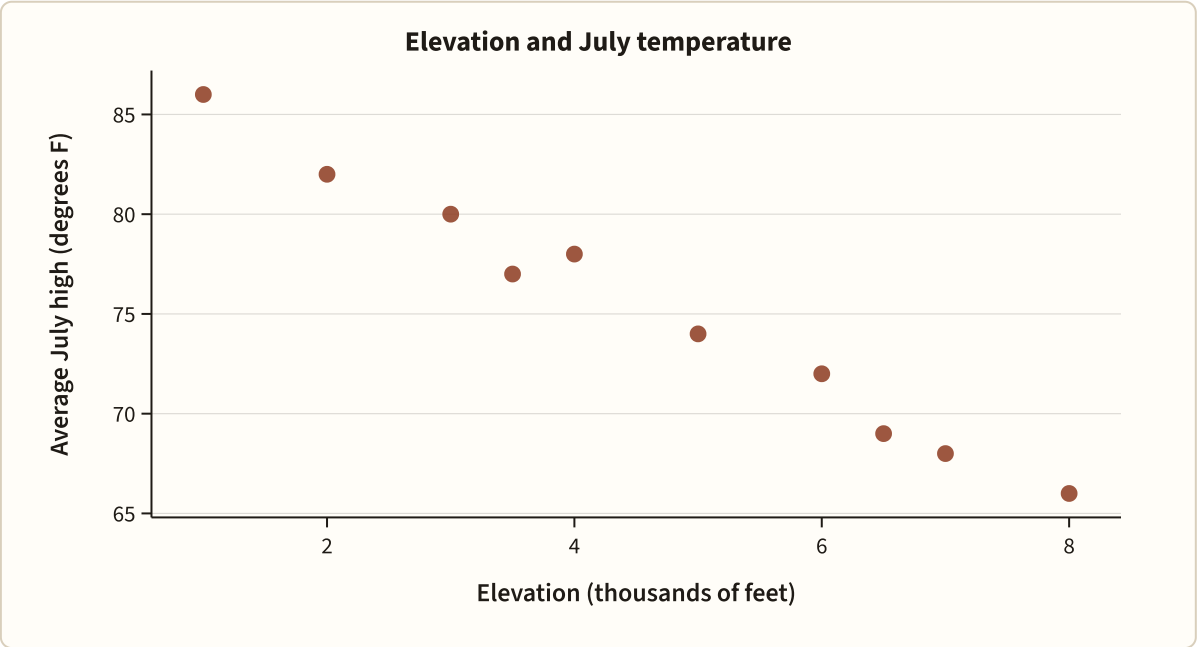
1. Variables: explanatory x = high temperature (°F); response y = glasses of lemonade sold.
2. Direction: apart from one point, the points rise from left to right, so the association is positive: warmer days tend to bring more sales.
3. Form: the nine points from 64°F to 86°F lie close to a straight line, so the form is linear.
4. Strength: those nine points hug the line closely, so the relationship is strong.
5. Unusual points: the day at 88°F with only 20 glasses sold falls far below the pattern (the pattern suggests well over 50). It is an outlier; perhaps an afternoon thunderstorm kept people indoors, and the group should check its records.
6. Conclusion in context: there is a strong, positive, linear association between the day's high temperature and the number of glasses of lemonade sold, with one clear outlier at 88°F and 20 glasses.

Answer: There is a strong, positive, linear association between high temperature and glasses of lemonade sold: hotter Saturdays tend to bring more sales. One day (88°F, 20 glasses) is a clear outlier far below the pattern.

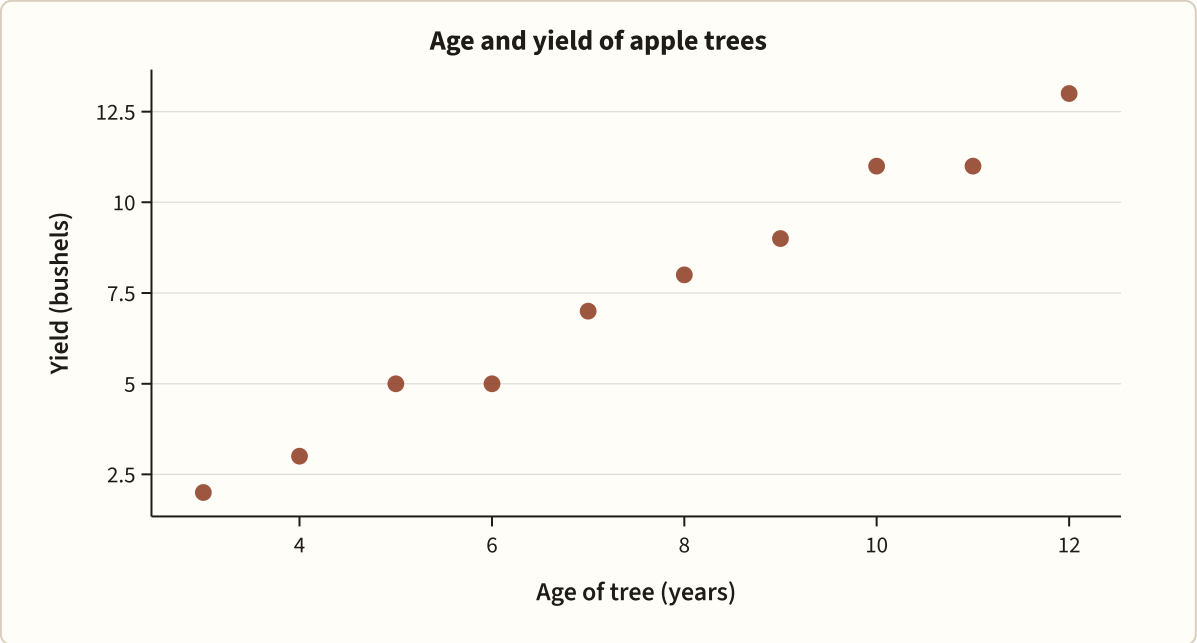
PRACTICE

1. EASY A farmer records the inches of rain that fall each June and the tons of hay per acre from that summer's first cutting. Which variable is explanatory and which is the response?

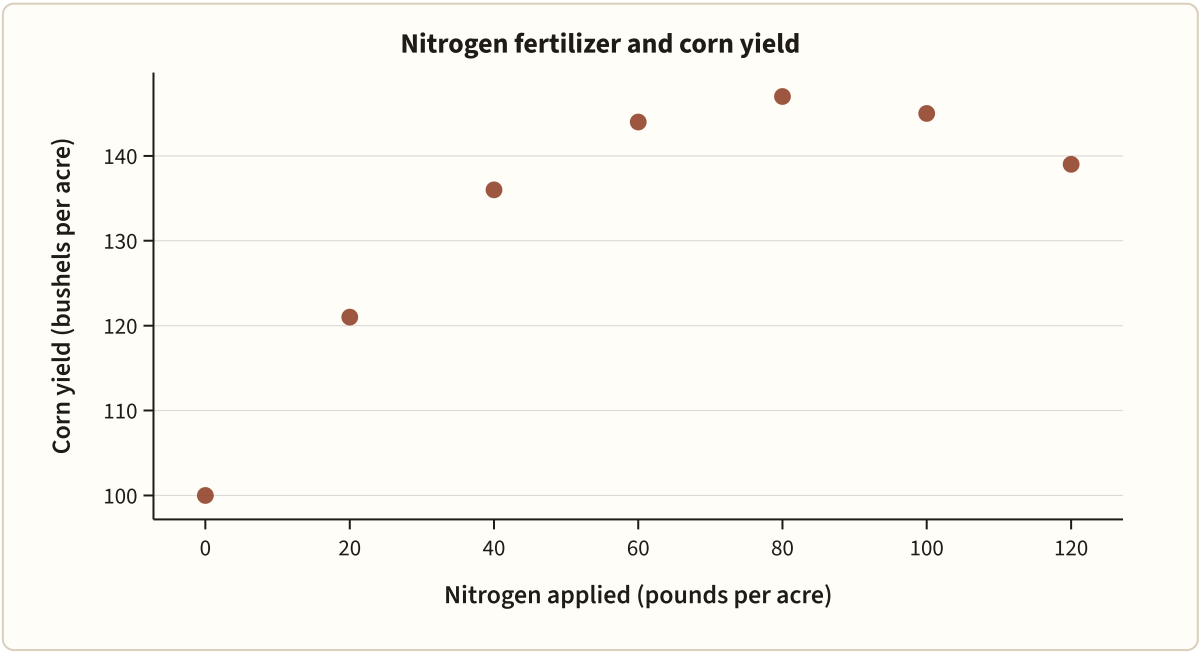
2. **EASY** The scatterplot shows the elevation and the average July high temperature at ten weather stations. Is the association positive or negative?



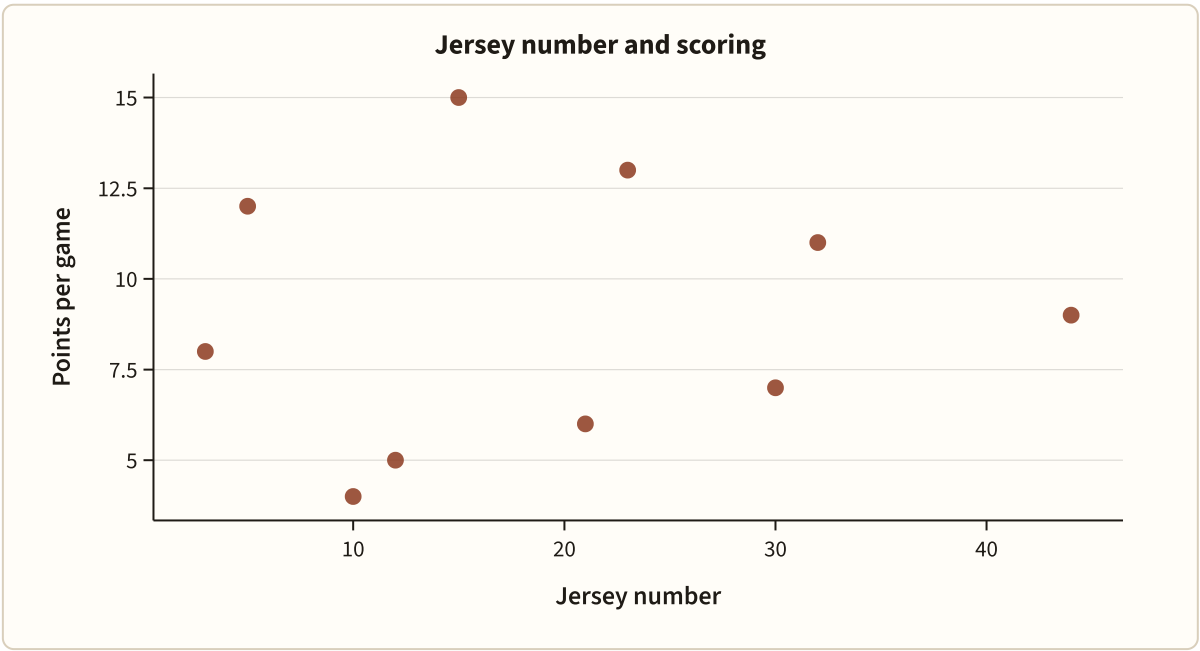
3. **EASY** The scatterplot shows the age and yield of ten apple trees. Is the association positive or negative?



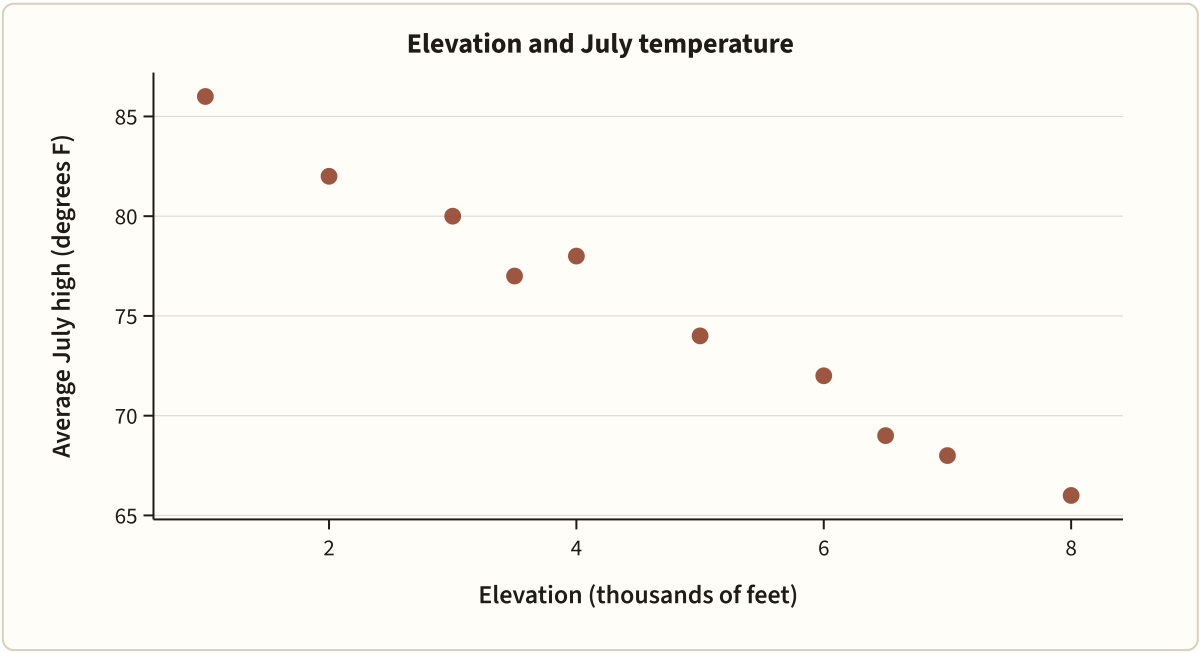
4. **EASY** The scatterplot shows corn yield for seven plots that received different amounts of nitrogen fertilizer. Is the form of the relationship linear? Answer yes or no.



5. **MEDIUM** A coach plotted each basketball player's jersey number against the player's points per game. Is there a positive association, a negative association, or none?



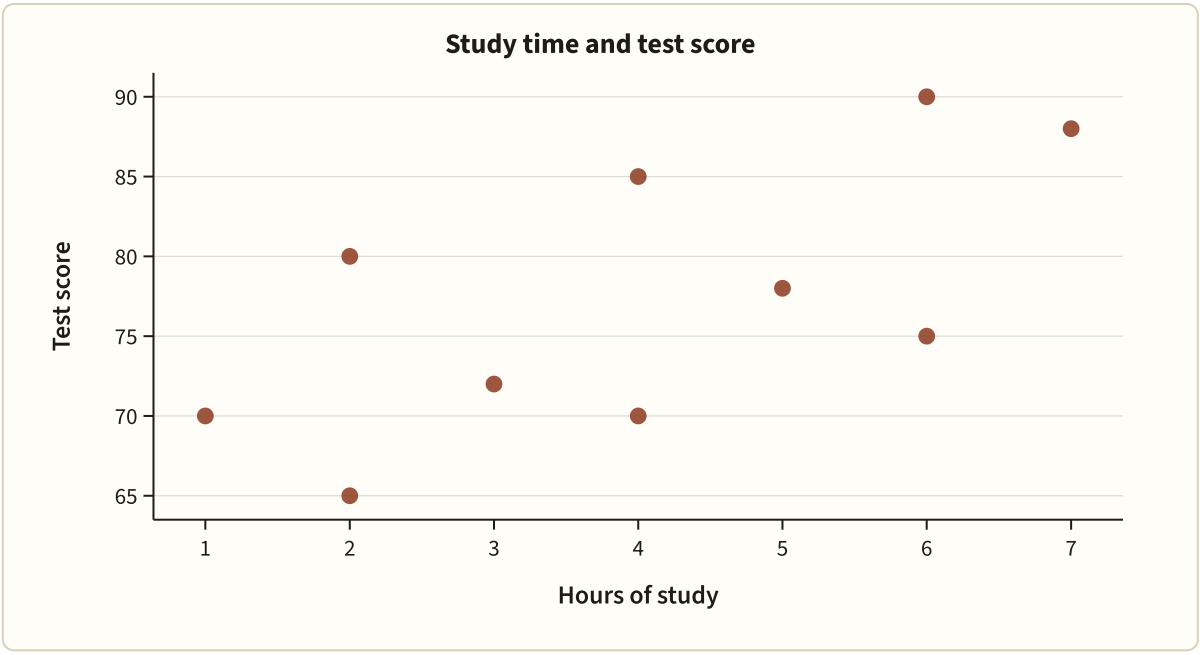
6. MEDIUM Write a complete description (direction, form, strength, unusual points) of the relationship between elevation and July temperature, in context.



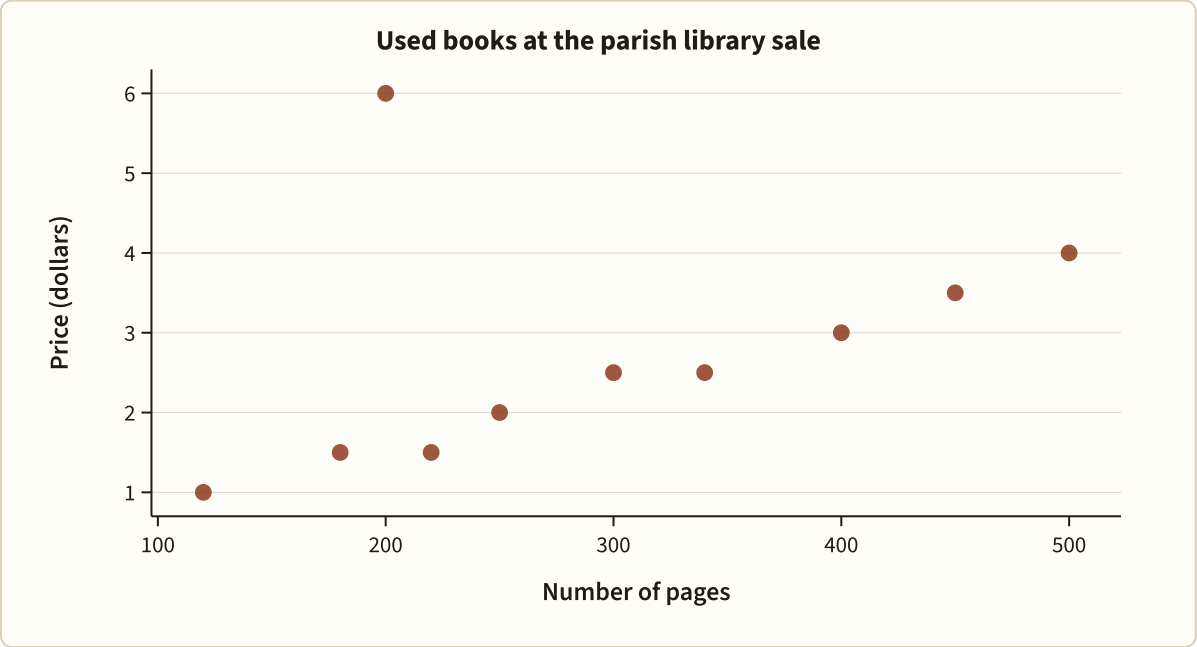
7. MEDIUM On eight summer evenings, Grace recorded the air temperature (°F) and the number of cricket chirps heard in one minute. (a) Which variable is explanatory? (b) Crickets chirp faster when they are warmer. What direction of association do you expect?

8. MEDIUM A survey of the children at a summer reading program found that children with larger shoe sizes tend to read at a higher level. Explain why buying bigger shoes would not improve a child's reading, and name a lurking variable.

9. MEDIUM Compare the apple-tree scatterplot (age and yield) with the scatterplot below of hours of study and test score for ten students. Which shows the stronger association? Explain.

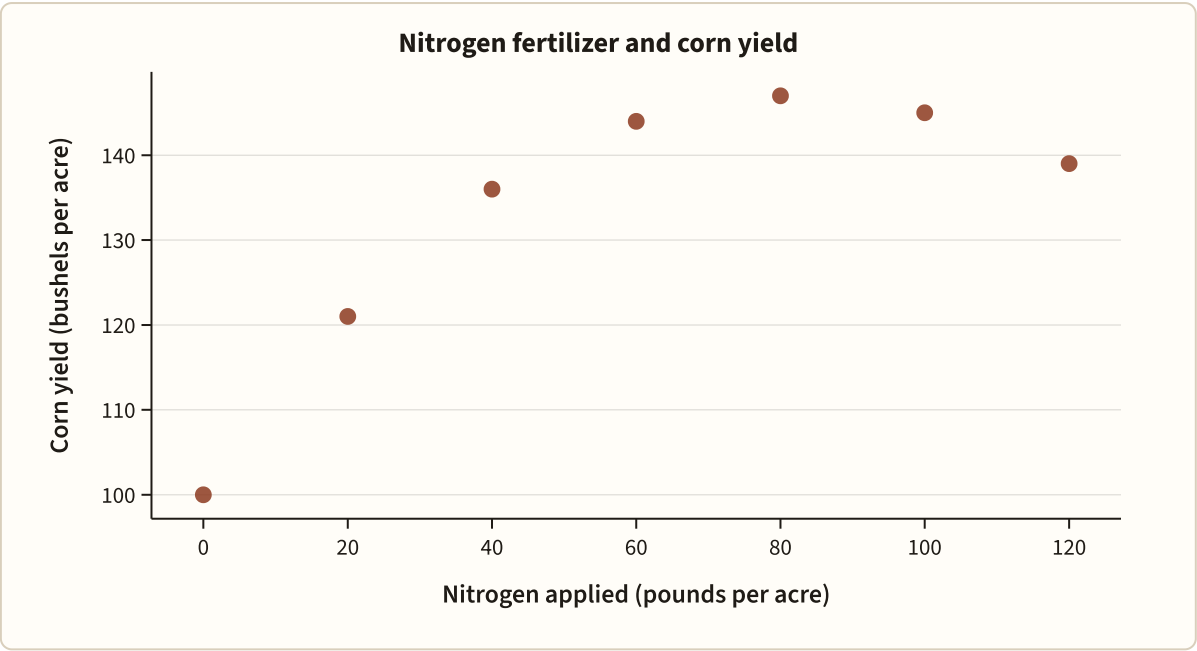


10. **HARD** At the parish library sale, used books were priced by size. The scatterplot shows the number of pages and the price of ten books. One book does not fit the pattern. Give its number of pages and its price in dollars.



11. **HARD** An agricultural survey of farms in one county found that farms using more tons of fertilizer harvest more total tons of grain. Samuel concludes that fertilizer is the reason those farms harvest more. Name a lurking variable that could explain much of this association, and explain how it affects both variables.

12. **HARD** Describe the relationship between nitrogen applied and corn yield (direction, form, strength, unusual points) in context, and explain why a farmer should not use a straight line to summarize it.



13. **CHALLENGE** Ruth made a scatterplot of a tree's age (x) and its yield (y), and then made a second one with the axes swapped: yield on the horizontal axis and age on the vertical axis. Does the direction of the association change? Answer yes or no, and explain why.

COMMON MISTAKES

- Putting the response variable on the horizontal axis. Ask "which variable helps explain the other?" and put that one on the x-axis.
- Describing a scatterplot without context, such as "positive, linear, strong." Name the variables and say what the pattern means: "older trees tend to yield more bushels."
- Calling every pattern linear. Look for curves, clusters and outliers before describing the form, and say "curved" when the points bend.
- Concluding that one variable causes the other because they are strongly associated. Always ask whether a lurking variable could drive both.

ANSWERS TO PRACTICE

1. Explanatory: inches of June rain. Response: tons of hay per acre, because rain helps the hay grow.
2. Negative: stations at higher elevations tend to have lower July temperatures.
3. Positive: older trees tend to yield more bushels.
4. No. Yield rises quickly at first, levels off around 80 pounds of nitrogen, then falls a little: the form is curved.
5. None. The points form a scattered cloud with no uphill or downhill trend; a jersey number tells us nothing about scoring, as we would expect.
6. There is a strong, negative, linear association between elevation and average July high temperature: higher stations tend to be cooler, by roughly 3 degrees for each additional thousand feet, and there are no outliers.
7. (a) Temperature is explanatory; the chirp rate responds to it. (b) Positive: warmer evenings should go with more chirps per minute.
8. Shoe size does not cause reading skill. The lurking variable is age: older children have bigger feet and have also had more years of reading practice, so age drives both variables.
9. The apple-tree plot shows the stronger association. Its points lie close to a straight line, while the study-time points are much more scattered about their uphill trend (for example, two students who studied 4 hours scored 70 and 85).
10. The book with 200 pages priced at $6.00 is an outlier; similar books sold for about $1.50. Perhaps it is a rare or nicely bound book.

11. Farm size (the number of acres) is a lurking variable. A larger farm uses more total fertilizer simply because it has more land to cover, and it harvests more total grain for the same reason. Fertilizer may help too, but these data cannot separate its effect from the effect of size.
12. Yield rises steeply as nitrogen increases from 0 to about 60 pounds per acre, levels off near 145 to 147 bushels around 80 pounds, and then falls slightly at 100 and 120 pounds. The form is clearly curved and the points follow the curve closely (a strong relationship), with no outliers. A straight line would predict that more nitrogen always means more corn, which these data contradict: beyond about 80 pounds, extra fertilizer is wasted or even harmful.
13. No. Swapping the axes reflects the plot across a diagonal line, but the pairing of values stays the same: large yields still go with large ages, so the points still rise from left to right and the association is still positive. Strength and form (linear) are unchanged as well.

LESSON 4.2 · 2 SESSIONS OF 45 MINUTES

The Correlation Coefficient

OBJECTIVES

- Explain why the average product of z-scores is positive for an uphill pattern and negative for a downhill one.
- Compute the correlation r for a small data set by hand with a table of z-scores, and with a calculator.
- Interpret r in context, giving its direction and strength.
- State the properties of r: it lies between −1 and 1, has no units, is unchanged by changes of units or by swapping x and y, measures only linear strength, and is not resistant.

VIDEO

WATCH

What is the Correlation Coefficient r ?

Mario's Math Tutoring · 7:14

When to watch: Before Example 1: what r measures and what it does not.

https://www.youtube.com/watch?v=z37oW9fTqAY

BACKUP

Pearson's Correlation Coefficient (1 of 3: Unpacking the formula)

Eddie Woo · 9:17

When to watch: After Example 2: where the formula for r comes from.

https://www.youtube.com/watch?v=-KdrlIRRbVA

KEY TERMS

Correlation (r)	A number between −1 and 1 that measures the direction and strength of the LINEAR relationship between two quantitative variables.
Standardized value (z-score)	How many standard deviations a value lies above (positive) or below (negative) its mean: $z = \dfrac{x - \bar{x}}{s_x}$.
Perfect correlation	r = 1 or r = −1: every point lies exactly on a straight line with positive or negative slope.
Resistant	Not strongly affected by a few extreme values. The median is resistant; r, like the mean, is not.

CORE IDEAS

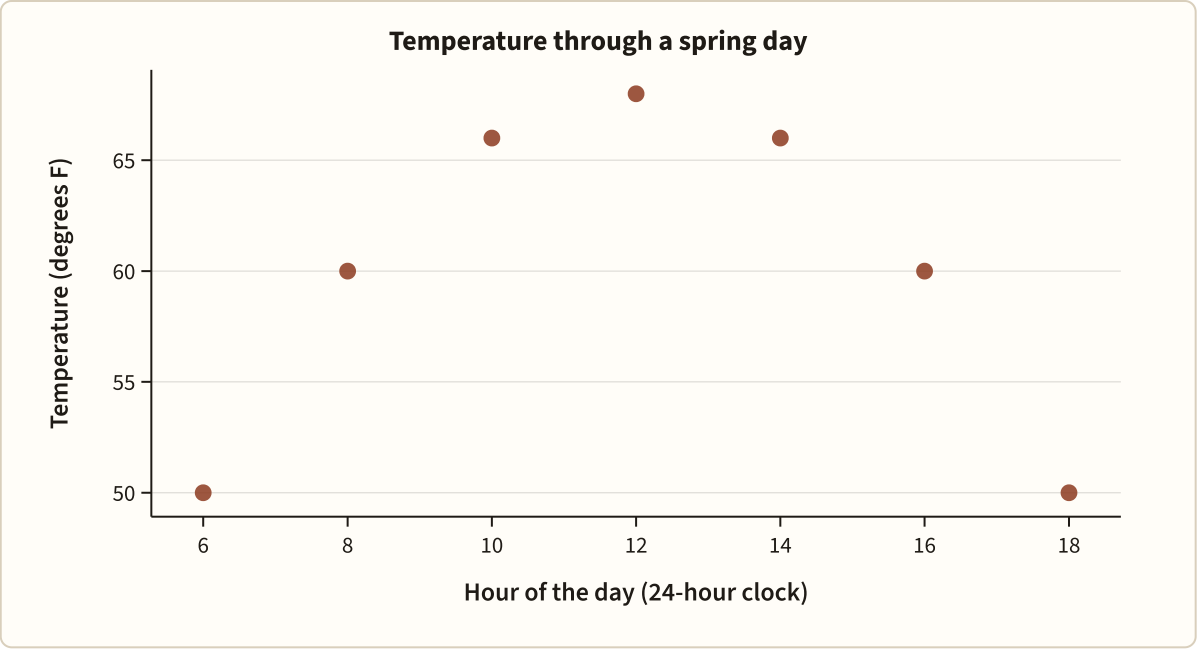
In Lesson 4.1 you judged strength by eye, but eyes are easily fooled: stretching an axis can make the same points look tighter or looser. We want a single number that measures how closely the points follow a straight line, whatever the units. That number is the **correlation** r.

The idea is to compare each point with the center of the data, the point $(\bar{x}, \bar{y})$. Standardize each variable: z_x says how many standard deviations a point's x lies above or below $\bar{x}$, and z_y does the same for y. Now multiply. A point up and to the right of the center has $z_x > 0$ and $z_y > 0$, so its product is positive; a point down and to the left has both negative, and the product is again positive. Points in the other two corners (upper left, lower right) give negative products.

In this plot of practice hours and free throws made, the center is $(\bar{x}, \bar{y}) = (5, 12)$. Every point is either below and left of it, like (2, 9), or above and right of it, like (6, 14). All the products are positive, so their average is positive: an uphill pattern. A downhill pattern puts the points in the upper-left and lower-right corners, giving mostly negative products and a negative r. A shapeless cloud has points in all four corners, so the products cancel and r is near 0.

So r is the *average* product of z-scores, where (just as for s) we divide by $n - 1$: $r = \dfrac{1}{n-1}\sum z_x z_y$. Using z-scores instead of raw deviations removes the units and puts both variables on the same scale, which is why r has no units and why converting inches to centimeters or °F to °C does not change it. Swapping x and y swaps the two z's in each product, which changes nothing, so the correlation of y with x equals the correlation of x with y.

Facts about r. It is always between −1 and 1. Values near 1 or −1 mean the points lie close to a straight line; r = 1 or −1 only when they lie exactly on one. The sign gives the direction. As a rough guide, |r| above 0.8 is strong, 0.5 to 0.8 is moderate, and below 0.5 is weak, but always look at the plot too.



Caution: r measures only **linear** strength. The temperatures above (taken every two hours on a spring day) follow a smooth, strong curve, yet r = 0 exactly: the morning points (rising) and the afternoon points (falling) give products that cancel. A correlation near 0 means "no straight-line relationship," not "no relationship." Also, r is **not resistant**: because each point's product counts fully, one outlier far from the pattern can raise or lower r sharply. Plot the data before trusting r.

KEY FORMULAS

$$r = \frac{1}{n-1}\sum z_x z_y, \quad \text{where } z_x = \frac{x-\bar{x}}{s_x} \text{ and } z_y = \frac{y-\bar{y}}{s_y}$$

$$-1 \le r \le 1$$

Practice habit. Plot before you trust r. A correlation describes only straight-line patterns, so look at the scatterplot for curves and outliers before you report r.

WORKED EXAMPLES

Example 1. A farmer finds that the correlation between the inches of rain during the growing season and the corn yield in bushels per acre on the farm's fields over 15 years is r = 0.82. (a) Interpret this value. (b) What would r be if rain were measured in centimeters instead of inches? (c) What would r be if yield were put on the x-axis and rainfall on the y-axis?

1. (a) The sign is positive, so wetter seasons tend to go with higher yields. Since 0.82 is close to 1, the points lie fairly close to a straight line: a strong positive linear relationship.
2. (b) Changing units multiplies every rainfall by the same positive number (2.54). Each z-score stays the same, because the value, the mean and the standard deviation are all multiplied by 2.54. So r = 0.82.
3. (c) Swapping x and y only swaps the two factors in each product $z_x z_y$, so the sum and r are unchanged: r = 0.82.
4. Conclusion: there is a strong, positive, linear relationship between growing-season rainfall and corn yield, and the number 0.82 does not depend on units or on which variable is on which axis.

Answer: (a) A strong, positive, linear relationship: seasons with more rain tend to have higher corn yields. (b) r = 0.82. (c) r = 0.82.

Example 2. Five players on a church-league basketball team recorded their hours of practice per week x and the number of free throws they made out of 20 attempts y: (2, 9), (4, 11), (6, 13), (6, 14), (7, 13). Compute r by hand using z-scores, and interpret it.

Computing r for the free-throw data

Hours x	Made y	z_x	z_y	z_x times z_y
2	9	-1.5	-1.5	2.25
4	11	-0.5	-0.5	0.25
6	13	0.5	0.5	0.25
6	14	0.5	1	0.5
7	13	1	0.5	0.5
Sum				3.75

1. Means: $\bar{x} = \dfrac{2+4+6+6+7}{5} = \dfrac{25}{5} = 5$ hours and $\bar{y} = \dfrac{9+11+13+14+13}{5} = \dfrac{60}{5} = 12$ free throws.
2. Deviations of x: −3, −1, 1, 1, 2. Squares: 9, 1, 1, 1, 4, sum 16. $s_x = \sqrt{\dfrac{16}{4}} = \sqrt{4} = 2$ hours.
3. Deviations of y: −3, −1, 1, 2, 1. Squares: 9, 1, 1, 4, 1, sum 16. $s_y = \sqrt{\dfrac{16}{4}} = 2$ free throws.
4. z-scores: divide each deviation by 2. z_x: −1.5, −0.5, 0.5, 0.5, 1.0. z_y: −1.5, −0.5, 0.5, 1.0, 0.5.
5. Products $z_x z_y$: 2.25, 0.25, 0.25, 0.5, 0.5. All are positive, because every point is lower-left or upper-right of (5, 12). Sum = 3.75.
6. $r = \dfrac{1}{n-1}\sum z_x z_y = \dfrac{3.75}{4} = 0.9375$.
7. Check with a calculator's two-variable statistics: r = 0.9375.
8. Conclusion: r ≈ 0.94 shows a strong, positive, linear relationship: players who practice more hours tend to make more of their free throws.

Answer: $r = \dfrac{3.75}{4} = 0.9375$: a strong, positive, linear relationship between practice hours and free throws made.

PRACTICE

1. EASY For used pickup trucks of one model, the correlation between age (years) and resale value (dollars) is r = −0.85. What is the direction of the association? Also describe its strength.

2. **EASY** Which correlation shows the stronger linear relationship, r = 0.62 or r = −0.81? Give the value.

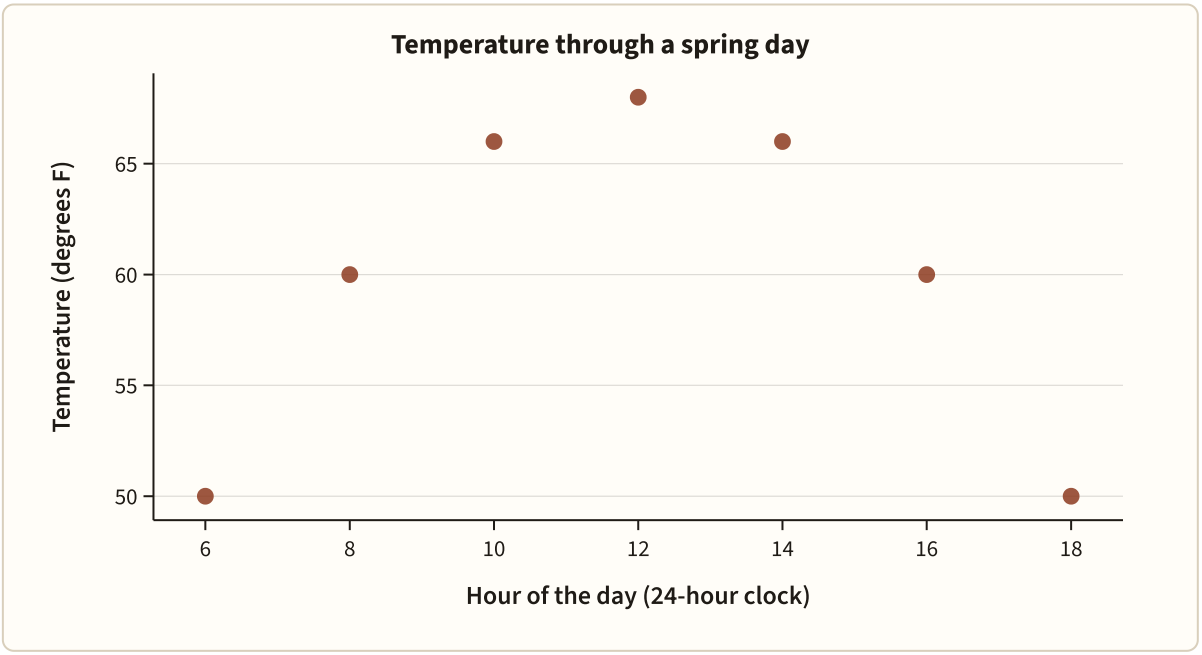
3. **EASY** For 20 trout caught in a mountain stream, the correlation between length in inches and weight in ounces is r = 0.93. Henry converts the lengths to centimeters and the weights to grams. What is the new correlation?

4. **EASY** One point in a scatterplot has $z_x = 1.2$ and $z_y = -0.5$. Find its product $z_x z_y$. Does this point pull r up or down?

5. **MEDIUM** For 8 weeks, a beekeeper recorded the average daily temperature and the pounds of honey gained by a hive. The sum of the products of z-scores is $\sum z_x z_y = 6.3$. Find r.

6. **MEDIUM** Five hens in a flock had ages (years) and weekly egg counts of (1, 6), (1, 5), (5, 5), (5, 3), (8, 1). Here $\bar{x} = 4$, $s_x = 3$, $\bar{y} = 4$ and $s_y = 2$. Compute r by hand with z-scores.

7. **MEDIUM** On a spring day the temperature (°F) was recorded every two hours: at hours 6, 8, 10, 12, 14, 16 and 18 it was 50, 60, 66, 68, 66, 60 and 50. (a) Find r. (Hint: $\bar{x} = 12$ and $\bar{y} = 60$; look at the products of the deviations.) (b) Does r = 0 mean there is no relationship between time of day and temperature? Explain.



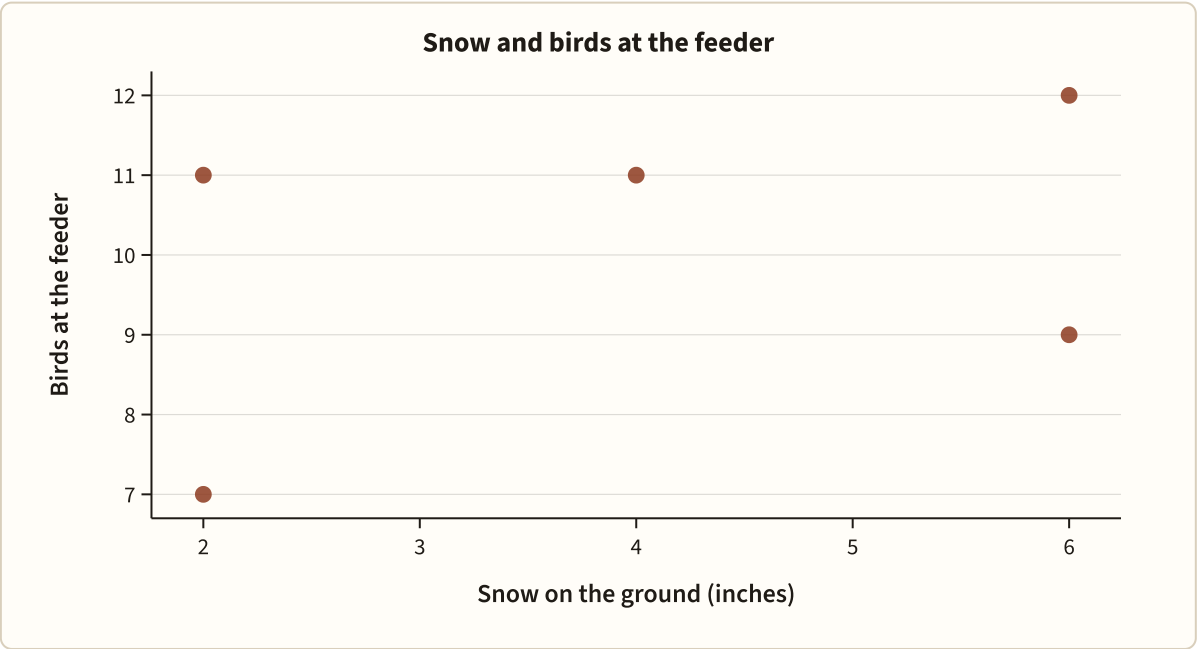
8. **MEDIUM** In Problem 6, r = −0.875 for predicting eggs from age. What is the correlation if a farmer instead puts eggs per week on the x-axis and age on the y-axis?

9. **MEDIUM** For 7 days at a sawmill, the sum of the products of deviations is $\sum(x-\bar{x})(y-\bar{y}) = 48$, with $s_x = 2$ and $s_y = 5$. Since $z_x z_y = \dfrac{(x-\bar{x})(y-\bar{y})}{s_x s_y}$, find r.

10. **HARD** Explain why the correlation between the weight of a calf (pounds) and its age (days) has no units, and why it would be the same with weight in kilograms.

11. **HARD** Ruth enters a set of garden data and reports r = 1.24 between hours of sunshine and inches of plant growth. Henry reports r = −0.3 for the same data and says the relationship is strong because 0.3 is a large number. Explain what is wrong with each report.

12. **HARD** On five winter mornings Samuel recorded the inches of snow on the ground x and the number of birds at the family's feeder y: (2, 7), (2, 11), (4, 11), (6, 9), (6, 12). Compute r by hand and describe the relationship. (Show $\bar{x}$, $\bar{y}$, s_x and s_y.)



13. **CHALLENGE** Suppose every point lies exactly on a line $y = 3 + 2x$. Explain why $z_y = z_x$ for every point, and use the fact that $\sum z_x^2 = n - 1$ (because $s_x^2 = \dfrac{\sum (x - \bar{x})^2}{n - 1}$) to show that r = 1.

COMMON MISTAKES

- Thinking a negative r means a weak relationship. The sign gives direction; strength comes from how close r is to 1 or −1, so r = −0.9 is strong.
- Reading r near 0 as "no relationship." r measures only straight-line strength; check the scatterplot for a curve.
- Dividing the sum of products by n instead of n − 1. Use n − 1, just as for the standard deviation s.
- Trusting r without checking for outliers. One unusual point can change r a great deal, so plot the data first.

ANSWERS TO PRACTICE

1. Negative: older trucks tend to sell for less. Since −0.85 is close to −1, the linear relationship is strong.
2. r = −0.81, because strength depends on how close r is to 1 or −1, and |−0.81| = 0.81 is larger than 0.62. The sign shows only direction.
3. r = 0.93. Changing units does not change the z-scores, so r is unchanged.
4. $z_x z_y = (1.2)(-0.5) = -0.6$. The point lies to the right of $\bar{x}$ and below $\bar{y}$ (the lower-right corner), so its negative product pulls r down.

5. $r = \dfrac{6.3}{8-1} = \dfrac{6.3}{7} = 0.9$: a strong positive linear relationship.
6. z_x: −1, −1, 0.3333, 0.3333, 1.3333; z_y: 1, 0.5, 0.5, −0.5, −1.5. Products: −1, −0.5, 0.1667, −0.1667, −2; sum = −3.5. $r = \dfrac{-3.5}{4} = -0.875$: a strong negative linear relationship; older hens lay fewer eggs.
7. (a) Deviations of x: −6, −4, −2, 0, 2, 4, 6; of y: −10, 0, 6, 8, 6, 0, −10. Products: 60, 0, −12, 0, 12, 0, −60, which add to 0, so r = 0. (b) No. Temperature follows a strong, smooth curve, rising until noon and falling after. r = 0 only says there is no straight-line relationship.
8. r = −0.875. Swapping x and y only swaps the factors in each product $z_x z_y$.
9. $\sum z_x z_y = \dfrac{48}{2 \times 5} = 4.8$, so $r = \dfrac{4.8}{7-1} = 0.8$.
10. r is built from z-scores. Each z-score is a deviation divided by a standard deviation in the same units (pounds divided by pounds, days divided by days), so the units cancel. Changing pounds to kilograms multiplies each weight, the mean weight and the standard deviation by the same number, so every z-score, and therefore r, stays the same.
11. Ruth's value is impossible, because r is always between −1 and 1; there must be an entry or calculation error. Henry's r = −0.3 shows a weak negative linear relationship, not a strong one: strength depends on how close r is to −1 or 1, and −0.3 is much closer to 0.
12. $\bar{x} = 4$, $\bar{y} = 10$; deviations of x: −2, −2, 0, 2, 2 (squares sum to 16, $s_x = \sqrt{16/4} = 2$); deviations of y: −3, 1, 1, −1, 2 (squares sum to 16, $s_y = 2$). Products $z_x z_y$: 1.5, −0.5, 0, −0.5, 1; sum = 1.5. $r = \dfrac{1.5}{4} = 0.375$: a weak positive linear relationship; more snow goes with slightly more birds, but the pattern is loose.
13. If y = 3 + 2x for every point, then $\bar{y} = 3 + 2\bar{x}$, so $y - \bar{y} = 2(x - \bar{x})$, and $s_y = 2s_x$. Dividing, $z_y = \dfrac{2(x-\bar{x})}{2s_x} = z_x$. Then $r = \dfrac{1}{n-1}\sum z_x z_x = \dfrac{1}{n-1}\sum z_x^2$. Since $\sum z_x^2 = \dfrac{\sum (x-\bar{x})^2}{s_x^2} = n - 1$, we get $r = \dfrac{n-1}{n-1} = 1$.

LESSON 4.3 · 2 SESSIONS OF 45 MINUTES

The Least-Squares Regression Line

OBJECTIVES

- Explain the least-squares idea: the regression line makes the sum of squared residuals as small as possible.
- Compute the least-squares regression line $\hat{y} = a + bx$ from summary statistics with $b = r\dfrac{s_y}{s_x}$ and $a = \bar{y} - b\bar{x}$.
- Interpret the slope and (when it makes sense) the y-intercept in context, and use the line to make predictions.
- Explain the danger of extrapolation and describe regression toward the mean.

VIDEO

WATCH

Pearson's Correlation Coefficient (3 of 3: Least squares regression line)

Eddie Woo · 9:58

When to watch: Before Example 1: the least-squares line from r and the standard deviations.

https://www.youtube.com/watch?v=Y4s4k0lVOCs

BACKUP

Simple Linear Regression: The Least Squares Regression Line

jbstatistics · 7:24

When to watch: Review: the least-squares line and interpreting slope and intercept.

https://www.youtube.com/watch?v=coQAAN4eY5s

KEY TERMS

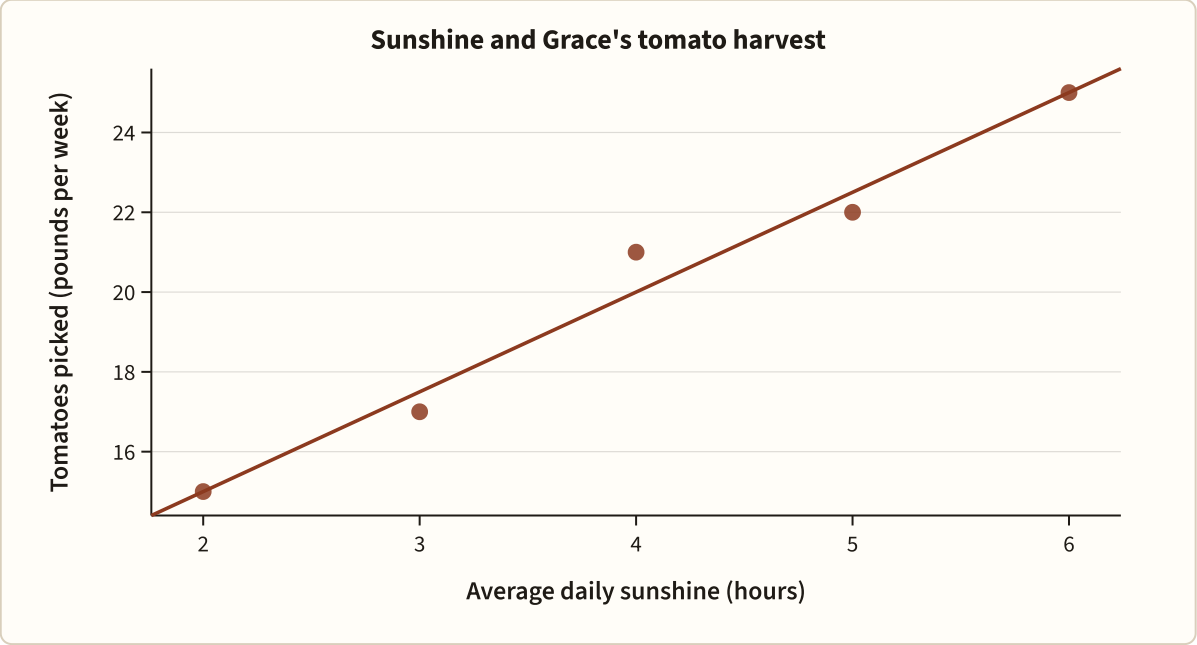
Regression line	A straight line $\hat{y} = a + bx$ that describes how a response variable y changes as an explanatory variable x changes; used to predict y from x.
Predicted value ($\hat{y}$)	The value of y that the line gives for a particular x. Read $\hat{y}$ as "y-hat"; the hat means "predicted."
Residual	The vertical miss of the line at a data point: residual $= y - \hat{y}$ (actual minus predicted).
Least-squares regression line	The one line that makes the sum of the squared residuals as small as possible.

Extrapolation	Using a regression line to predict for x-values far outside the range of the data; such predictions are often badly wrong.
Regression toward the mean	Because |r| is less than 1, the predicted y is fewer standard deviations from its mean than x is from its mean.

CORE IDEAS

When a scatterplot shows a linear pattern, a straight line can summarize it and let us predict. We write the line as $\hat{y} = a + bx$, where b is the slope and a the y-intercept. (This is y = mx + b from Algebra I with different letters; statisticians put the intercept first.) The hat on $\hat{y}$ reminds us that the line gives a *predicted* value, not the actual one.

Many lines could be drawn through a cloud of points. Which is best? At each data point the line misses by a **residual**, $y - \hat{y}$: the vertical distance from the point to the line (vertical, because we are predicting y). We want the residuals small overall. Adding them does not work: positive and negative misses cancel, and many poor lines have residuals that add to 0. Instead we **square** them. Squares are never negative, so misses cannot cancel, and squaring makes a big miss count much more than a small one, so the best line stays close to every point. The **least-squares regression line** is the one line that makes $\sum(y - \hat{y})^2$ as small as possible.



There are simple formulas for this line. Think in standard units first. If z_x tells how many standard deviations x is from $\bar{x}$, the best straight-line prediction of z_y turns out to be rz_x: multiply by the correlation. Converting back to the original units, $\dfrac{\hat{y} - \bar{y}}{s_y} = r\dfrac{x - \bar{x}}{s_x}$, so $\hat{y} = \bar{y} + r\dfrac{s_y}{s_x}(x - \bar{x})$. This line has slope $b = r\dfrac{s_y}{s_x}$ and intercept $a = \bar{y} - b\bar{x}$. Notice what happens when $x = \bar{x}$: the prediction is exactly $\bar{y}$. **The least-squares line always passes through the point** $(\bar{x}, \bar{y})$, the center of the data.

Interpreting the slope. b is the predicted change in y for each one-unit increase in x. For Grace's tomatoes above, $\hat{y} = 10 + 2.5x$ means: *for each additional hour of daily sunshine, the predicted weekly harvest increases by about 2.5 pounds.* Say "predicted" or "on average": the line describes a tendency, not a law that every week obeys. **Interpreting the intercept.** a is the predicted y when x = 0. Interpret it only when x = 0 makes sense and is near the data. A week with zero sunshine is far outside Grace's data (2 to 6 hours), so the intercept of 10 pounds is just the starting height of the line, not a real prediction.

Prediction and extrapolation. To predict, substitute x into the line: for 4.5 hours, $\hat{y} = 10 + 2.5(4.5) = 21.25$ pounds. Predictions are trustworthy only within the range of the x-values in the data. **Extrapolation**, predicting far outside that range, assumes the straight-line pattern continues where we have no evidence. Twelve hours of sunshine a day would predict 40 pounds, but heat, water and the size of the garden would surely change the pattern.

Regression toward the mean. Since $\hat{z}_y = r z_x$ and |r| < 1 (unless the points lie exactly on a line), the predicted y is always fewer standard deviations from its mean than x is. If r = 0.5 and a ballplayer's batting average this season is 2 standard deviations above the mean, the prediction for next season is only 1 standard deviation above. This is not a mysterious force; it reflects the fact that an extreme result is partly skill and partly good fortune, and good fortune does not usually repeat.

KEY FORMULAS

$$\hat{y} = a + bx$$

$$b = r\frac{s_y}{s_x}$$

$$a = \bar{y} - b\bar{x}$$

$$\text{residual} = y - \hat{y}$$

From history. The French mathematician Adrien-Marie Legendre published the method of least squares in 1805, in a work on finding the orbits of comets. Carl Friedrich Gauss had already used the method: in 1801 he used it to predict where the newly found dwarf planet Ceres would reappear after it was lost in the glare of the Sun, and astronomers found it close to the place he predicted.

Practice habit. Interpret every slope with its units and the word "predicted": "for each additional [unit of x], the predicted [y] increases (or decreases) by about b [units of y]."

WORKED EXAMPLES

Example 1. Over 15 seasons a farmer recorded growing-season rainfall x (inches) and corn yield y (bushels per acre). The summary statistics are in the table, and r = 0.8. (a) Find the slope of the least-squares regression line. (b) Find its intercept. (c) Interpret the slope. (d) Predict the yield for a season with 22 inches of rain.

Rainfall and corn yield, 15 seasons

Variable	Mean	Standard deviation
Rainfall x (inches)	20	4
Yield y (bushels per acre)	150	15

1. Slope: $b = r\dfrac{s_y}{s_x} = 0.8 \times \dfrac{15}{4} = 0.8 \times 3.75 = 3$.
2. Intercept: $a = \bar{y} - b\bar{x} = 150 - 3(20) = 150 - 60 = 90$.
3. Line: $\hat{y} = 90 + 3x$.
4. Slope in context: for each additional inch of growing-season rain, the predicted corn yield increases by about 3 bushels per acre.
5. Prediction: $\hat{y} = 90 + 3(22) = 90 + 66 = 156$ bushels per acre. Since 22 inches is half a standard deviation above the mean rainfall, this is a sensible prediction inside the range of the data.
6. The intercept (90 bushels with no rain at all) would be an extrapolation far outside the data and should not be interpreted.

Answer: (a) b = 3. (b) a = 90, so $\hat{y} = 90 + 3x$. (c) For each additional inch of rain, the predicted yield rises by about 3 bushels per acre. (d) $\hat{y} = 90 + 3(22) = 156$ bushels per acre.

Example 2. For five weeks Grace recorded the average daily hours of sunshine x and the pounds of tomatoes picked y: (2, 15), (3, 17), (4, 21), (5, 22), (6, 25). Find the least-squares regression line, interpret the slope, and predict the harvest for a week averaging 4.5 hours of sunshine.

1. Means: $\bar{x} = \dfrac{20}{5} = 4$ hours; $\bar{y} = \dfrac{15 + 17 + 21 + 22 + 25}{5} = \dfrac{100}{5} = 20$ pounds.
2. Deviations of x: −2, −1, 0, 1, 2; squares add to 10, so $s_x = \sqrt{\dfrac{10}{4}} = \sqrt{2.5} = 1.58114$.
3. Deviations of y: −5, −3, 1, 2, 5; squares 25, 9, 1, 4, 25 add to 64, so $s_y = \sqrt{\dfrac{64}{4}} = 4$.
4. Correlation (calculator, or z-scores as in Lesson 4.2): r = 0.98821.
5. Slope: $b = r\dfrac{s_y}{s_x} = 0.98821 \times \dfrac{4}{1.58114} = 2.5$ (keep full precision to get exactly 2.5).
6. Intercept: $a = \bar{y} - b\bar{x} = 20 - 2.5(4) = 10$.
7. Line: $\hat{y} = 10 + 2.5x$. Check: at $x = \bar{x} = 4$, $\hat{y} = 20 = \bar{y}$.
8. Slope in context: for each additional hour of daily sunshine, the predicted weekly harvest increases by about 2.5 pounds.
9. Prediction: $\hat{y} = 10 + 2.5(4.5) = 10 + 11.25 = 21.25$ pounds of tomatoes.

Answer: $\hat{y} = 10 + 2.5x$. Each additional hour of daily sunshine raises the predicted weekly harvest by about 2.5 pounds; for 4.5 hours, $\hat{y} = 21.25$ pounds.

PRACTICE

1. EASY For young pine trees on a tree farm, the line $\hat{y} = 1.5 + 0.8x$ predicts height y (feet) from age x (years). Predict the height of a 10-year-old pine.
2. EASY Interpret the slope of the pine-tree line $\hat{y} = 1.5 + 0.8x$ in context.
3. EASY For a set of data, r = 0.6, $s_x = 4$ and $s_y = 10$. Find the slope b of the least-squares line.
4. EASY A least-squares line has slope b = 1.5, and the data have $\bar{x} = 20$ and $\bar{y} = 50$. Find the intercept a.
5. MEDIUM For 40 golfers at a county tournament, the table gives summary statistics for clubhead speed x (mph) and drive distance y (yards), and r = 0.92. Find the least-squares regression line.

Clubhead speed and drive distance, 40 golfers

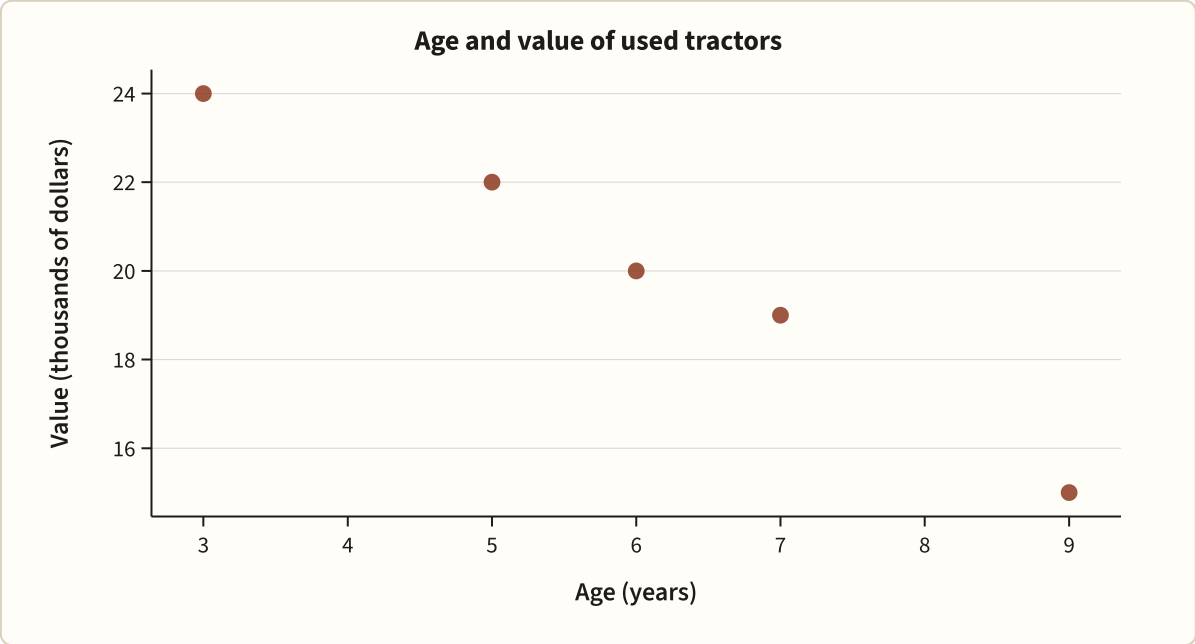
Variable	Mean	Standard deviation
Clubhead speed x (mph)	100	8
Drive distance y (yards)	250	20

6. MEDIUM Use the golf line $\hat{y} = 20 + 2.3x$ to predict the drive distance for a golfer with a clubhead speed of 108 mph.

7. MEDIUM The golf line has intercept 20. Should we interpret it as "a golfer who swings at 0 mph drives 20 yards"? Explain.

8. MEDIUM The pine-tree line $\hat{y} = 1.5 + 0.8x$ was found from trees aged 2 to 15 years. (a) What height does it predict for an 80-year-old pine? (b) Why should we not trust this prediction?

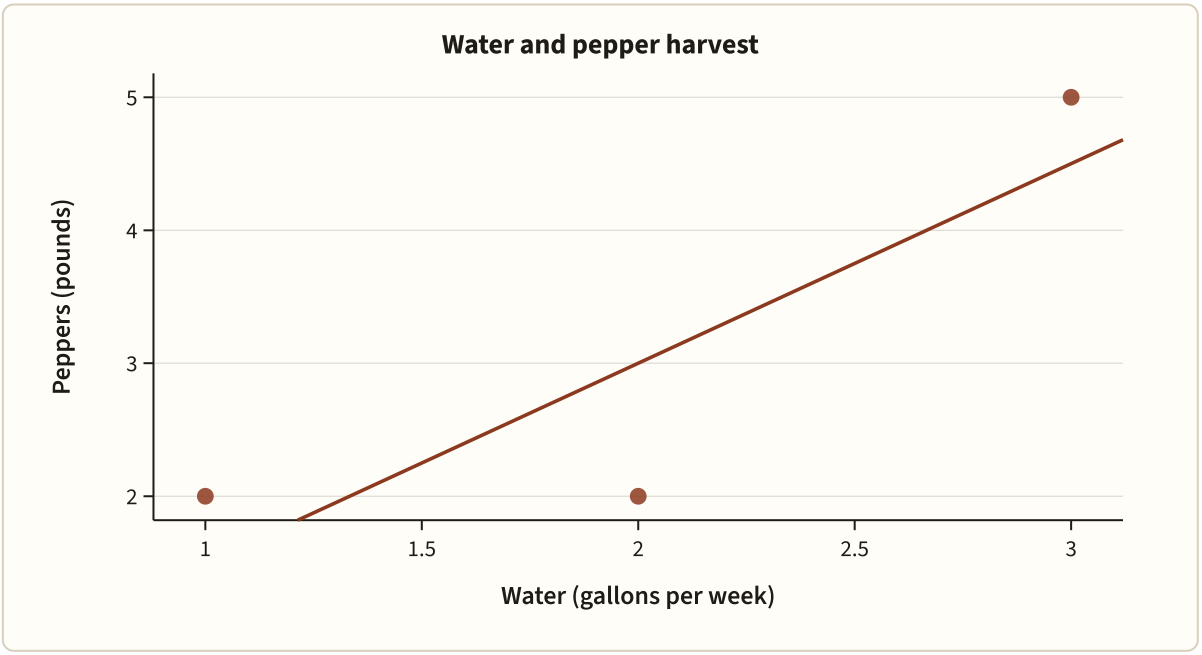
9. HARD A dealer recorded the ages (years) and values (thousands of dollars) of five used tractors of one model: (3, 24), (5, 22), (6, 20), (7, 19), (9, 15). Find the least-squares regression line. (You may use a calculator for r; round a and b to two decimal places.)



10. HARD For a set of data, r = 0.5. One individual's x-value is 2 standard deviations above $\bar{x}$. How many standard deviations above $\bar{y}$ is its predicted y-value?

11. HARD In a baseball league, batting averages in two seasons both have mean 0.260 and standard deviation 0.030, and the correlation between a player's first-season and second-season averages is r = 0.4. A player hit 0.320 in the first season. Use the regression line to predict the player's second-season average. Round to three decimal places.

12. HARD Henry gave three pepper plants 1, 2 and 3 gallons of water per week and harvested 2, 2 and 5 pounds of peppers. The least-squares line is $\hat{y} = 1.5x$. (a) Find the sum of the squared residuals for this line. (b) Find the sum of the squared residuals for the line $\hat{y} = 1 + x$. (c) Which is smaller, and why must it be?



13. CHALLENGE For trees in a woodlot, trunk diameter x has mean 10 inches and standard deviation 2 inches, height y has mean 30 feet and standard deviation 6 feet, and r = 0.5. (a) Find the slope of the least-squares line for predicting height from diameter. (b) Find the slope of the least-squares line for predicting diameter from height. Round to four decimal places. (c) Explain why the second line is not found by solving the first line for x.

COMMON MISTAKES

- Writing the line without the hat, as y = 10 + 2.5x. The line gives predicted values; write $\hat{y}$.
- Interpreting the slope as a certainty ("each hour of sun adds 2.5 pounds"). Say "the predicted harvest increases by about 2.5 pounds."
- Interpreting the intercept when x = 0 is impossible or far from the data. Check whether x = 0 makes sense before you interpret a.
- Rounding r, s_x or s_y early and getting a slightly wrong slope. Keep full precision until the final answer.

ANSWERS TO PRACTICE

1. $\hat{y} = 1.5 + 0.8(10) = 9.5$ feet.
2. For each additional year of age, the predicted height of a pine increases by about 0.8 foot.
3. $b = 0.6 \times \dfrac{10}{4} = 1.5$.
4. $a = 50 - 1.5(20) = 50 - 30 = 20$.

5. $b = 0.92 \times \dfrac{20}{8} = 2.3$; $a = 250 - 2.3(100) = 20$. $\hat{y} = 20 + 2.3x$.
6. $\hat{y} = 20 + 2.3(108) = 20 + 248.4 = 268.4$ yards.
7. No. A clubhead speed of 0 mph means no swing at all, which is impossible for a drive and far outside the data (speeds near 100 mph). The intercept only fixes the height of the line; it has no meaning in this context.
8. (a) $\hat{y} = 1.5 + 0.8(80) = 65.5$ feet. (b) This is extrapolation: 80 years is far outside the ages in the data. Trees do not keep growing at a steady rate for decades; growth slows as they mature, so the straight-line pattern probably does not continue.
9. $\bar{x} = 6$, $\bar{y} = 20$, $s_x = \sqrt{5} = 2.2361$, $s_y = \sqrt{11.5} = 3.3912$, r = −0.98907. $b = -0.98907 \times \dfrac{3.3912}{2.2361} = -1.50$; $a = 20 - (-1.5)(6) = 29.00$. $\hat{y} = 29 - 1.5x$: each additional year of age lowers the predicted value by about 1.5 thousand dollars.
10. $\hat{z}_y = r z_x = 0.5 \times 2 = 1$: the prediction is 1 standard deviation above $\bar{y}$, closer to the mean than x is (regression toward the mean).
11. $z_x = \dfrac{0.320 - 0.260}{0.030} = 2$; predicted $z_y = 0.4 \times 2 = 0.8$; $\hat{y} = 0.260 + 0.8(0.030) = 0.284$. The prediction regresses toward the mean.
12. (a) Predictions 1.5, 3, 4.5; residuals 0.5, −1, 0.5; squares 0.25 + 1 + 0.25 = 1.5. (b) Predictions 2, 3, 4; residuals 0, −1, 1; squares 0 + 1 + 1 = 2. (c) The least-squares line gives the smaller sum, 1.5, because by definition no other line has a smaller sum of squared residuals.
13. (a) $b = 0.5 \times \dfrac{6}{2} = 1.5$ feet per inch. (b) $b = 0.5 \times \dfrac{2}{6} = 0.1667$ inch per foot. (c) Solving $\hat{y} = 15 + 1.5x$ for x would give slope $\dfrac{1}{1.5} = 0.6667$, not 0.1667. Each least-squares line minimizes the squared misses in its own response variable (vertical misses for y, horizontal misses for x), and both lines regress toward the mean, so they are different lines unless r = 1 or −1.

LESSON 4.4 · 2 SESSIONS OF 45 MINUTES

Residuals and R^2

OBJECTIVES

- Compute residuals (actual minus predicted) and interpret their signs in context.
- Make and read a residual plot, and decide whether a straight line is a suitable model.
- Compute and interpret the standard deviation of the residuals s and the coefficient of determination r^2.
- Read regression output from a calculator or computer (constant, slope, s, r^2) and use it to predict.

VIDEO

WATCH

Simple Linear Regression: Checking Assumptions with Residual Plots

jbstatistics · 8:04

When to watch: Before Example 1: residual plots and what a pattern means.

https://www.youtube.com/watch?v=iMdtTCX2Q70

BACKUP

The Correlation Coefficient and Coefficient of Determination (Old, fast version)

jbstatistics · 6:16

When to watch: Before Example 2: r^2 as the fraction of variation explained.

https://www.youtube.com/watch?v=dCgIavyFWIo

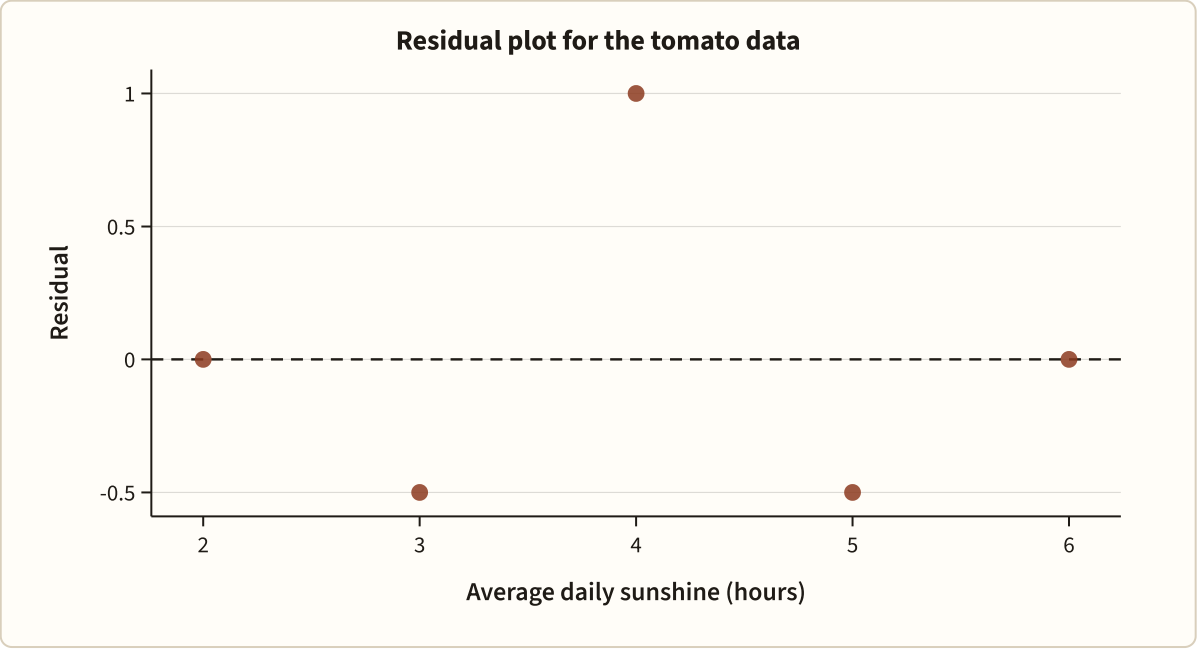
KEY TERMS

Residual	Actual y minus predicted y: $y - \hat{y}$. A positive residual means the point lies above the line (the line predicted too low); a negative residual means it lies below.
Residual plot	A scatterplot of the residuals (vertical axis) against x (horizontal axis), with a horizontal line at 0.
Standard deviation of the residuals (s)	The typical size of a prediction error: about how far actual y-values fall from the regression line.
Coefficient of determination (r^2)	The fraction of the variation in y that is accounted for by the linear relationship with x.

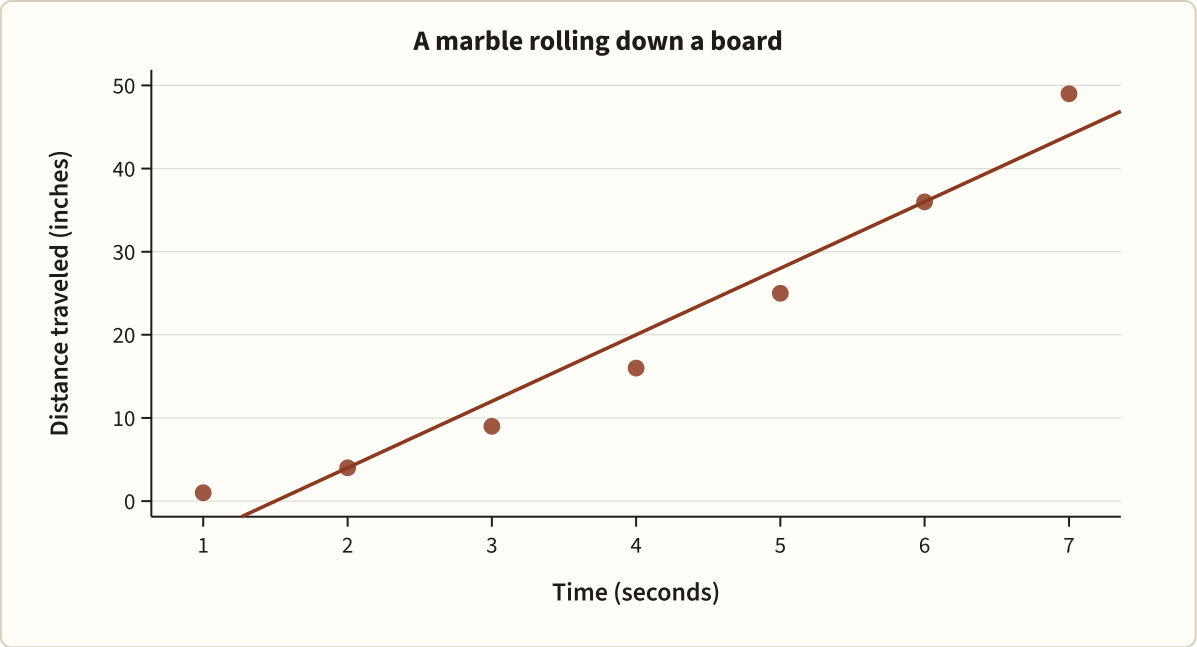
CORE IDEAS

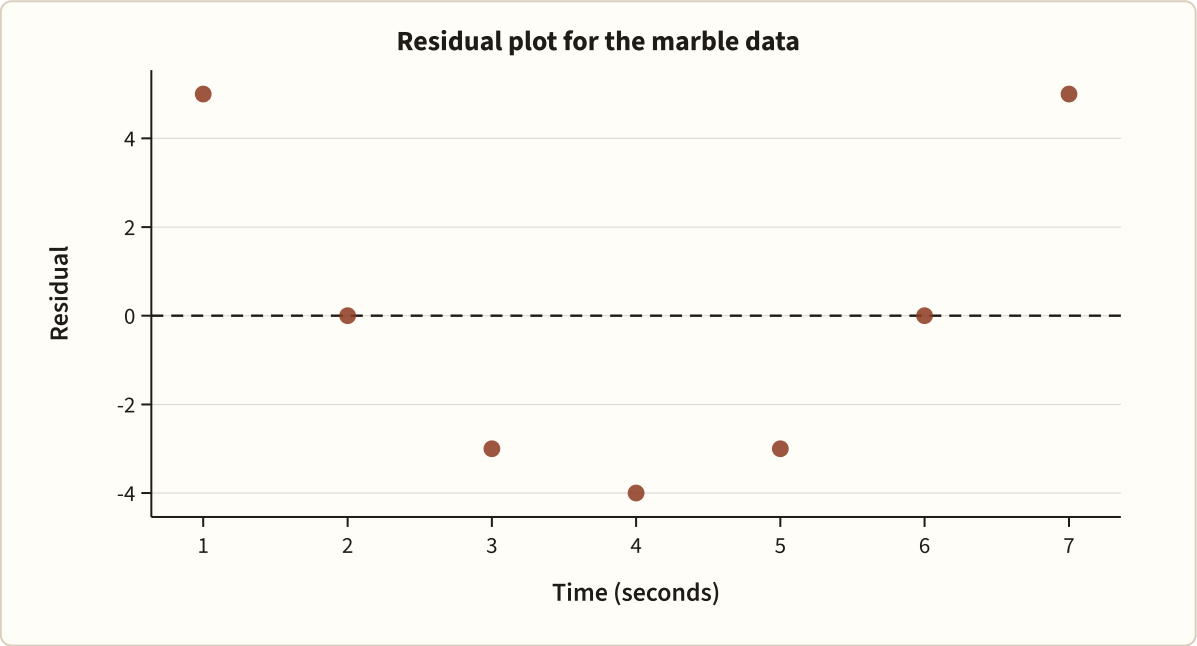
A regression line is only useful if it fits the data well. The **residuals** tell us. Each residual, $y - \hat{y}$, is how far the actual value is above (positive) or below (negative) the line. For the least-squares line the residuals always add to 0, because the line passes through the center of the data: the misses above and below balance.

To see the misses clearly, make a **residual plot**: plot each residual against its x-value, with a horizontal line at 0. This plot lays the regression line flat and magnifies the leftover scatter. Here is the residual plot for Grace's tomato data from Lesson 4.3 (line $\hat{y} = 10 + 2.5x$; residuals 0, −0.5, 1, −0.5, 0). A good residual plot looks like this: the points bounce above and below 0 with **no pattern**, and the spread stays about the same from left to right.



A pattern in the residuals means the line has missed something. Ruth timed a marble rolling down a long, gently sloping board and measured the distance it had traveled after 1, 2, ..., 7 seconds: 1, 4, 9, 16, 25, 36 and 49 inches. The correlation is r = 0.98, very close to 1, and the line $\hat{y} = -12 + 8x$ seems to fit well.





But the residuals (5, 0, −3, −4, −3, 0, 5) form a clear U: the line is too high in the middle and too low at both ends. The marble speeds up as it rolls, so the true relationship is curved, and a straight line is the wrong model even though r is large. **A curved pattern in the residual plot means a line is not appropriate.** A fan shape (residuals spreading out as x grows) warns that predictions are less precise for larger x.

How big are the misses? The **standard deviation of the residuals** is $s = \sqrt{\dfrac{\sum(y - \hat{y})^2}{n - 2}}$. It works like the standard deviation of Unit 2: square the misses, average them, take the square root. We divide by $n - 2$ because the line used two numbers, a and b, computed from the same data (just as s for one variable

divides by $n-1$ because $\bar{x}$ was computed from the data). For the tomatoes, $s = \sqrt{\dfrac{1.5}{3}} = 0.71$ pound: predictions from the line are typically off by about 0.7 pound.

Finally, how much does knowing x help? Without x, our best prediction for every week is $\bar{y} = 20$ pounds, and the total squared miss is $\sum (y - \bar{y})^2 = 64$. With the line, the total squared miss is only $\sum (y - \hat{y})^2 = 1.5$. The line removed $64 - 1.5 = 62.5$ of the 64, a fraction $\dfrac{62.5}{64} = 0.977$. This fraction equals the square of the correlation: $r^2 = 0.98821^2 = 0.977$. So r^2 **is the fraction of the variation in y accounted for by the linear relationship with x:** about 97.7% of the week-to-week variation in Grace's harvest is accounted for by the linear relationship with hours of sunshine. The rest is left over in the residuals.

KEY FORMULAS

$$\text{residual} = y - \hat{y}$$

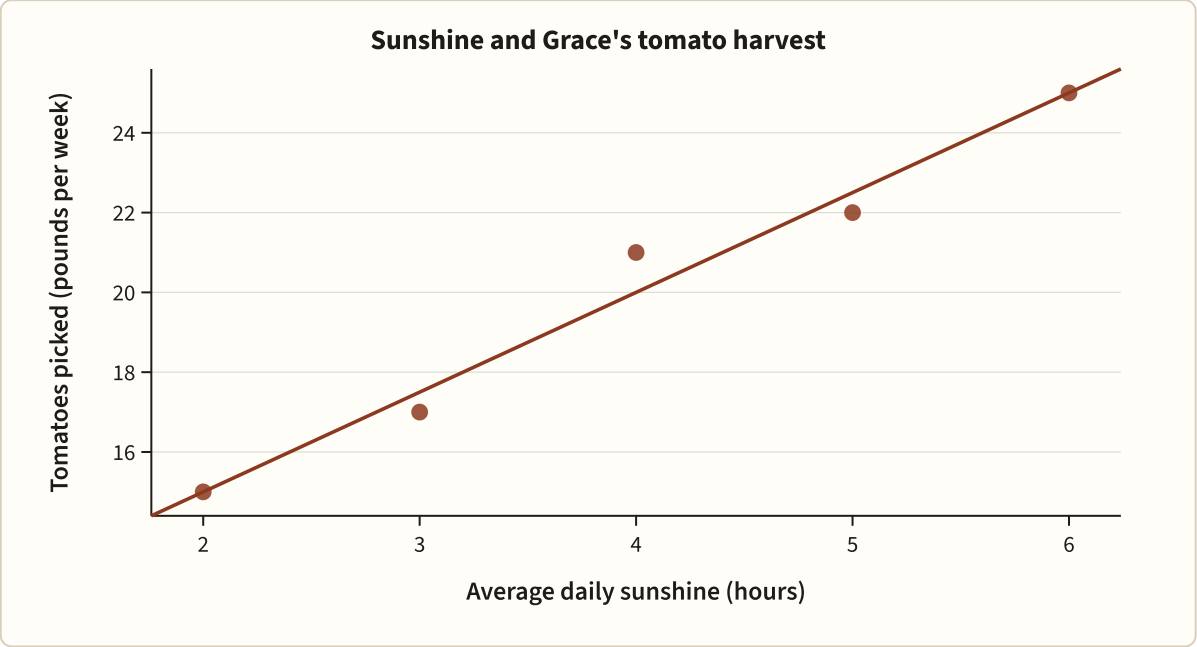
$$s = \sqrt{\frac{\sum (y - \hat{y})^2}{n - 2}}$$

$$r^2 = \frac{\sum (y - \bar{y})^2 - \sum (y - \hat{y})^2}{\sum (y - \bar{y})^2}$$

Practice habit. Always look at the residual plot before trusting a regression line: no pattern means the line fits; a curve means a straight line is the wrong model, however large r is.

WORKED EXAMPLES

Example 1. Grace's tomato data are (2, 15), (3, 17), (4, 21), (5, 22), (6, 25), with x = average daily hours of sunshine and y = pounds picked, and the least-squares line is $\hat{y} = 10 + 2.5x$. (a) Find the residual for the week with 4 hours of sunshine and interpret it. (b) Find the standard deviation of the residuals s. Round to two decimal places. (c) Given $\sum(y - \bar{y})^2 = 64$, find r^2 and interpret it. Round to three decimal places.



1. (a) Predicted: $\hat{y} = 10 + 2.5(4) = 20$ pounds. Residual: $y - \hat{y} = 21 - 20 = 1$ pound.
2. In context: that week Grace picked 1 pound more than the line predicted for 4 hours of sunshine.
3. (b) All predictions: 15, 17.5, 20, 22.5, 25. All residuals: 0, −0.5, 1, −0.5, 0 (they add to 0, as they must).
4. Squares: 0, 0.25, 1, 0.25, 0; sum $= 1.5$.
5. $s = \sqrt{\dfrac{1.5}{5-2}} = \sqrt{0.5} = 0.71$ pound: predictions from this line are typically off by about 0.7 pound.
6. (c) $r^2 = \dfrac{64 - 1.5}{64} = \dfrac{62.5}{64} = 0.977$.
7. In context: about 97.7% of the variation in weekly tomato harvest is accounted for by the linear relationship with hours of sunshine.

Answer: (a) Residual $= 21 - 20 = 1$ pound: Grace picked 1 pound more than predicted. (b) $s = \sqrt{1.5/3} = 0.71$ pound. (c) $r^2 = 62.5/64 = 0.977$: about 97.7% of the variation in harvest is accounted for by the linear relationship with sunshine.

Example 2. A dairy farmer weighed 15 calves aged 2 to 90 days and ran a regression of weight (pounds) on age (days). The output is shown; the slope is positive. (a) Predict the weight of a 60-day-old calf. (b) A 60-day-old calf weighs 200 pounds. Find its residual. (c) Find the correlation r. (d) Interpret s and r^2 in context.

Regression output: calf weight (pounds) on age (days)

Predictor	Coef
Constant	84
Age	1.75
S	12.5
R-Sq	81.0%

1. Read the output: the constant is a = 84.0 and the slope is b = 1.75, so $\hat{y} = 84.0 + 1.75x$.
2. Slope in context: for each additional day of age, a calf's predicted weight increases by about 1.75 pounds. Here the intercept makes sense, because the youngest calves in the data are only 2 days old, so x = 0 is just outside the data: a newborn calf is predicted to weigh about 84 pounds.
3. (a) $\hat{y} = 84.0 + 1.75(60) = 84 + 105 = 189$ pounds.
4. (b) Residual $= 200 - 189 = 11$ pounds: this calf is 11 pounds heavier than predicted for its age.
5. (c) $r^2 = 0.810$, so $|r| = \sqrt{0.81} = 0.9$. The slope is positive, so r = 0.9.
6. (d) S = 12.5: predicted weights are typically off by about 12.5 pounds. R-Sq = 81.0%: about 81% of the variation in the calves' weights is accounted for by the linear relationship with age; the other 19% comes from other things such as breed, feed and health.

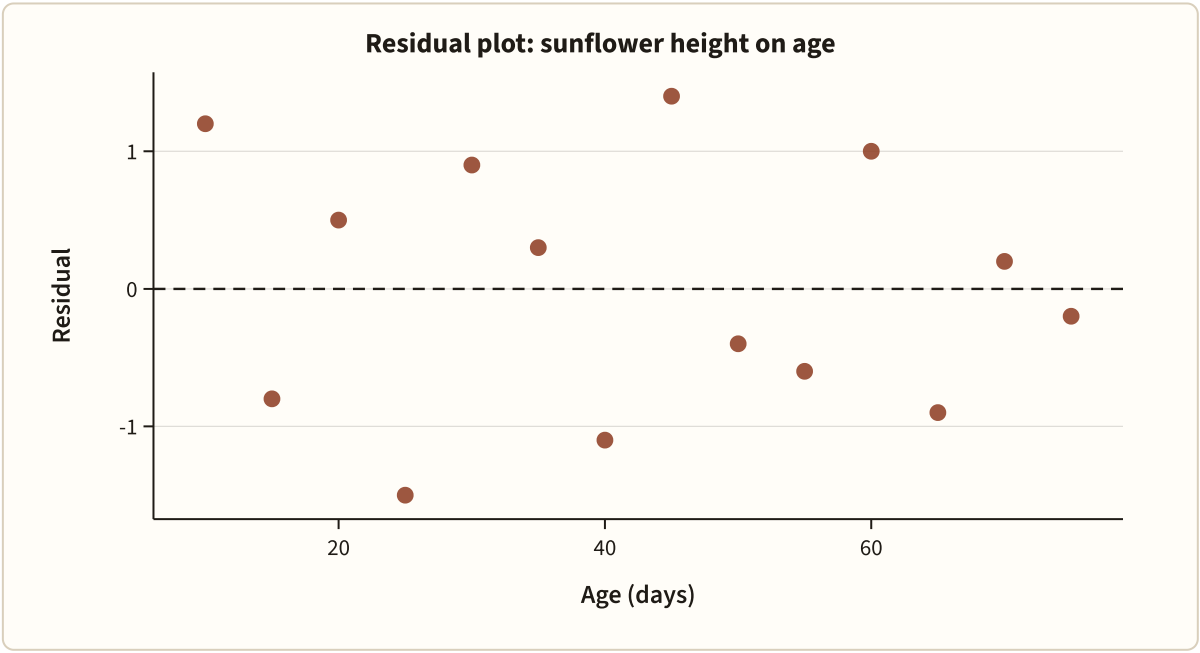
Answer: (a) $\hat{y} = 84.0 + 1.75(60) = 189$ pounds. (b) $200 - 189 = 11$ pounds. (c) $r = +\sqrt{0.81} = 0.9$. (d) Predictions are typically off by about 12.5 pounds; 81% of the variation in weight is accounted for by the linear relationship with age.

PRACTICE

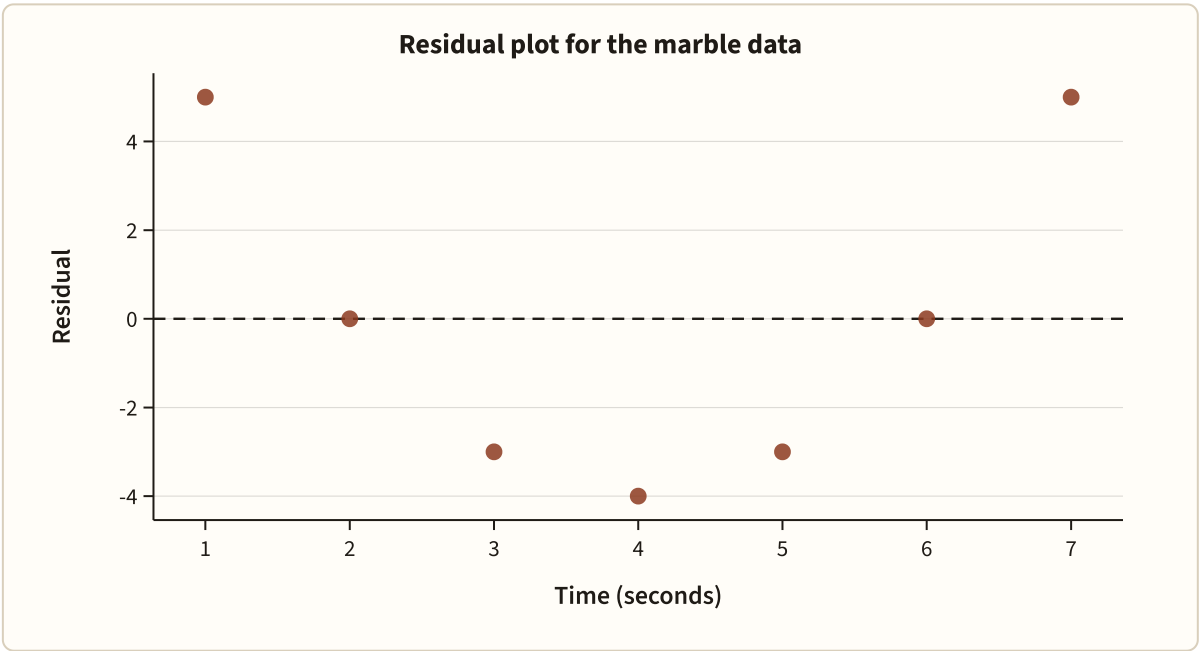
1. EASY A regression line predicts that an apple tree will yield 48.5 bushels this year. It actually yields 52 bushels. Find the residual.
2. EASY The line $\hat{y} = 10 + 2.5x$ predicts pounds of tomatoes from hours of sunshine. In a week with 4 hours of sunshine at a neighbor's garden, 18 pounds were picked. Find the residual.
3. EASY The correlation between a lamb's age and its weight is r = 0.7. What percent of the variation in weight is accounted for by the linear relationship with age?

4. **EASY** A line predicts a high school runner's 1-mile time from the runner's 400-meter time. Ruth's residual is −6 seconds. What does that mean in context?

5. **MEDIUM** The residual plot below comes from a regression of sunflower height (inches) on age (days). Is a straight line an appropriate model? Answer yes or no, and explain.



6. **MEDIUM** Ruth's marble data give r = 0.98. Henry says the high correlation proves a straight line is the right model. Using the residual plot, is a straight line appropriate? Answer yes or no, and explain.



7. MEDIUM A poultry farmer ran a regression of eggs per week on hen age (years). Use the output to write the regression equation and predict the eggs per week for a 4-year-old hen.

Regression output: eggs per week on hen age (years)

Predictor	Coef
Constant	6.8
Age	-0.55
S	0.9
R-Sq	64.0%

8. MEDIUM Use the egg-production output to find the correlation r between hen age and eggs per week.

Regression output: eggs per week on hen age (years)

Predictor	Coef
Constant	6.8
Age	-0.55
S	0.9
R-Sq	64.0%

9. MEDIUM Interpret the values S = 0.9 and R-Sq = 64.0% in the egg-production output, in context.

Regression output: eggs per week on hen age (years)

Predictor	Coef
Constant	6.8
Age	-0.55
S	0.9
R-Sq	64.0%

10. HARD For five used tractors with ages 3, 5, 6, 7, 9 years and values 24, 22, 20, 19, 15 (thousands of dollars), the least-squares line is $\hat{y} = 29 - 1.5x$. Find the five residuals, in order.

11. HARD For the tractor data in Problem 10, $\sum(y - \bar{y})^2 = 46$. (a) Find the standard deviation of the residuals s. Round to three decimal places. (b) Find r^2. Round to three decimal places.

12. **HARD** A beekeeper predicts each hive's honey harvest (pounds) from the number of frames of bees in spring for 12 hives. The total squared deviation from the mean is $\sum(y-\bar{y})^2 = 200$, and the sum of squared residuals from the least-squares line is 50. The slope is positive. (a) Find r^2. (b) Find r. Round to three decimal places.

13. **CHALLENGE** Square garden plots have sides of 1, 2, 3, 4 and 5 yards and areas of 1, 4, 9, 16 and 25 square yards. The least-squares line for predicting area from side length is $\hat{y} = -7 + 6x$, and r = 0.98. (a) Find the five residuals, in order. (b) Is a straight line an appropriate model? Answer yes or no, and explain using the residuals.

COMMON MISTAKES

- Computing predicted minus actual. A residual is actual minus predicted, $y - \hat{y}$; a point above the line has a positive residual.
- Deciding a line fits because r is close to 1. Always check the residual plot: a curve in the residuals means the wrong model, whatever r is.
- Saying "$r^2 = 0.81$ means 81% of the data are on the line" or "x causes 81% of y." Say "81% of the variation in y is accounted for by the linear relationship with x."
- Forgetting the sign when finding r from r^2. Take the square root, then give r the same sign as the slope.

ANSWERS TO PRACTICE

1. Residual $= 52 - 48.5 = 3.5$ bushels; the tree yielded 3.5 bushels more than predicted.
2. $\hat{y} = 10 + 2.5(4) = 20$; residual $= 18 - 20 = -2$ pounds (2 pounds less than predicted).
3. $r^2 = 0.7^2 = 0.49$, so 49%.
4. Ruth's actual mile time was 6 seconds less (faster) than the line predicted from her 400-meter time; her point lies below the line.
5. Yes. The residuals scatter randomly above and below 0 with no curve and about the same spread across all ages, so the line fits the pattern well.
6. No. The residuals form a clear U shape (positive at both ends, negative in the middle), so the relationship is curved: the marble speeds up. A large r shows only that a line comes close, not that a line is the right form.
7. $\hat{y} = 6.8 - 0.55x$; $\hat{y} = 6.8 - 0.55(4) = 6.8 - 2.2 = 4.6$ eggs per week.
8. $\sqrt{0.64} = 0.8$, and the slope is negative, so r = −0.8.
9. S = 0.9: predictions of a hen's weekly egg count from her age are typically off by about 0.9 egg. R-Sq = 64%: about 64% of the variation in weekly egg production is accounted for by the linear relationship with age; the remaining 36% is due to other things such as breed, feed and season.
10. Predictions: 24.5, 21.5, 20, 18.5, 15.5. Residuals: $24 - 24.5 = -0.5$, $22 - 21.5 = 0.5$, $20 - 20 = 0$, $19 - 18.5 = 0.5$, $15 - 15.5 = -0.5$ (they add to 0).

11. (a) $\sum (y-\hat{y})^2 = 0.25 + 0.25 + 0 + 0.25 + 0.25 = 1$; $s = \sqrt{\dfrac{1}{5-2}} = 0.577$ thousand dollars (about \$577). (b) $r^2 = \dfrac{46-1}{46} = 0.978$: about 97.8% of the variation in value is accounted for by the linear relationship with age.
12. (a) $r^2 = \dfrac{200-50}{200} = 0.75$. (b) $r = +\sqrt{0.75} = 0.866$ (positive because the slope is positive).
13. (a) Predictions −1, 5, 11, 17, 23; residuals $1-(-1)=2$, $4-5=-1$, $9-11=-2$, $16-17=-1$, $25-23=2$. (b) No. The residuals form a U (positive, negative, negative, negative, positive), showing a curved relationship; indeed area = side^2, a curve, not a line, despite r = 0.98.

LESSON 4.5 · 2 SESSIONS OF 45 MINUTES

Cautions about Regression and Causation

OBJECTIVES

- Distinguish an outlier from an influential point, and check a point's influence by fitting the line with and without it.
- Explain why correlation and regression describe only linear relationships and why correlations based on averages are stronger than those for individuals.
- Identify lurking variables and confounding in observational data, in context.
- Explain why correlation does not imply causation and how a randomized experiment can show cause and effect.

VIDEO

WATCH

Statistics, Storks, and Babies - Numberphile

Numberphile · 9:20

When to watch: Before Example 2: correlation is not causation (storks and babies, and the history of smoking studies).

https://www.youtube.com/watch?v=-n-d1ApBTFw

BACKUP

Leverage and Influential Points in Simple Linear Regression

jbstatistics · 7:14

When to watch: Before Example 1: outliers, leverage and influential points.

https://www.youtube.com/watch?v=xc_X9GFVuVU

KEY TERMS

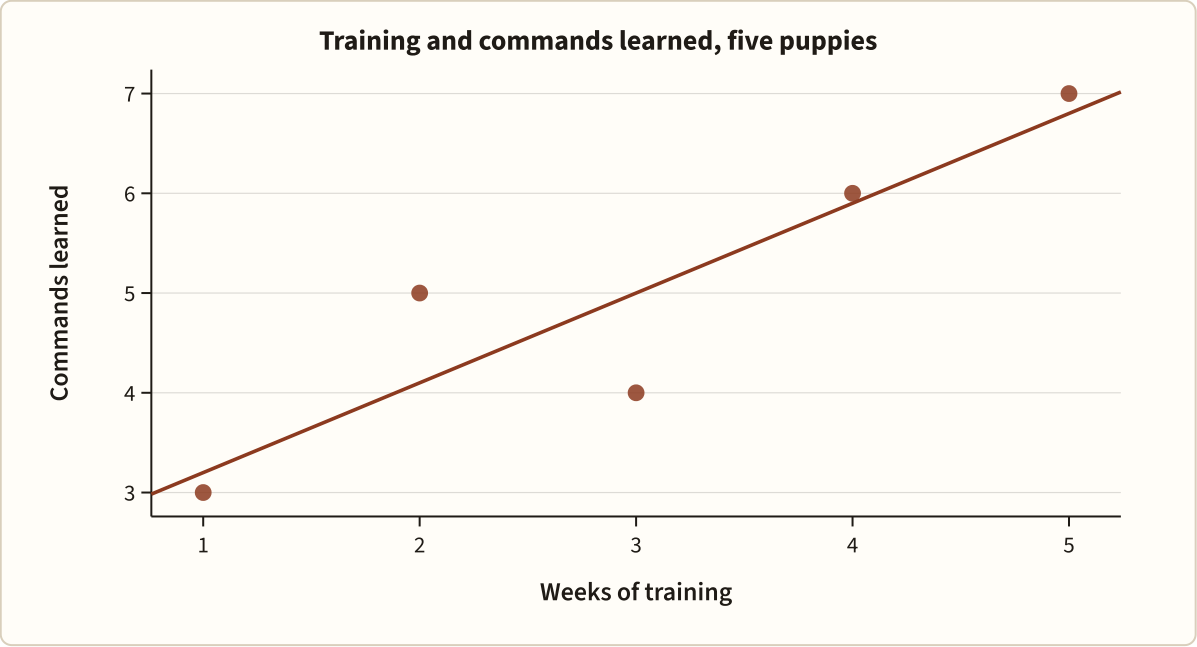
Outlier (in regression)	A point that lies far from the overall pattern of the scatterplot; a point with a large residual is an outlier in the y direction.
Influential point	A point that, if removed, would change the regression line (or r) a great deal. Points far out in the x direction are often influential.
Lurking variable	A variable not among those studied that affects the variables being studied and can create or hide an association between them.
Confounding	Two variables are confounded when their effects on a response cannot be told apart, so we cannot say which one caused a change.

Causation	Changing the explanatory variable actually produces a change in the response variable.

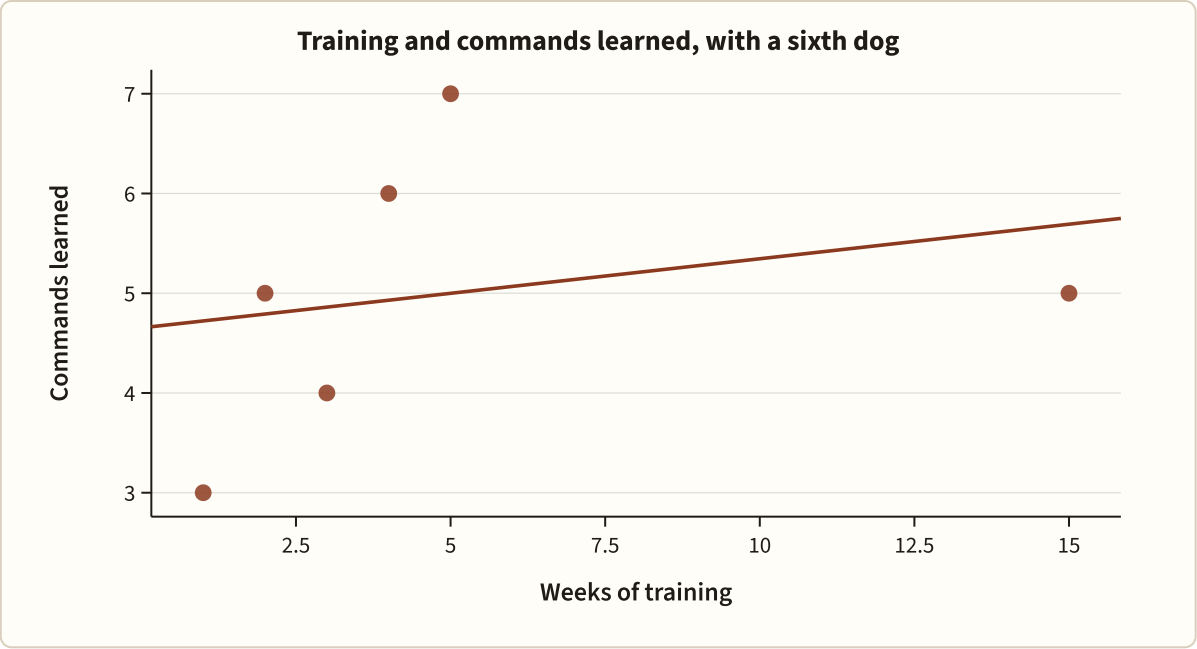
CORE IDEAS

Correlation and regression are powerful tools, but they are easy to misuse. This lesson collects the main cautions: watch for unusual points, remember that the methods describe only straight-line patterns, be careful with averaged data, and never confuse association with cause.

Outliers and influential points. A 4-H club trained five puppies for 1 to 5 weeks; they learned 3, 5, 4, 6 and 7 commands. The pattern is linear and strong: $\hat{y} = 2.3 + 0.9x$ with r = 0.9.



Now add a sixth dog, trained for 15 weeks, that knows only 5 commands. Refitting by calculator gives $\hat{y} = 4.65 + 0.069x$ and r = 0.25. One point flattened the slope from 0.9 to about 0.07 and cut r from 0.9 to 0.25. This dog is an **influential point**.



Why does one point have so much pull? It lies far out in the x direction, like a child sitting at the very end of a seesaw. The least-squares line must pass through $(\bar{x}, \bar{y})$, and a point far from $\bar{x}$ would have an enormous squared residual unless the line pivots toward it, so the line pivots. Notice that the influential point now has a fairly small residual, so you cannot always find influential points by looking at residuals. The sure check is to **fit the line with and without the point** and compare. By contrast, a point near $\bar{x}$ that sits far above or below the pattern is an outlier in y: it lowers r and shifts the line up or down, but it barely changes the slope. Never simply delete an unusual point; find out why it is unusual, and report what you did.

Only linear relationships. r and the regression line describe straight-line patterns only. A strong curved relationship can have r near 0, and a curved one can have r near 1 while a line still fits badly. Always look at the scatterplot and the residual plot. **Averaged data.** A correlation computed from averages (the average yield of each county, the average temperature of each month) is usually much stronger than the correlation for individuals, because averaging smooths away the individual-to-individual variation. Such a correlation says nothing reliable about how closely individuals follow the pattern.

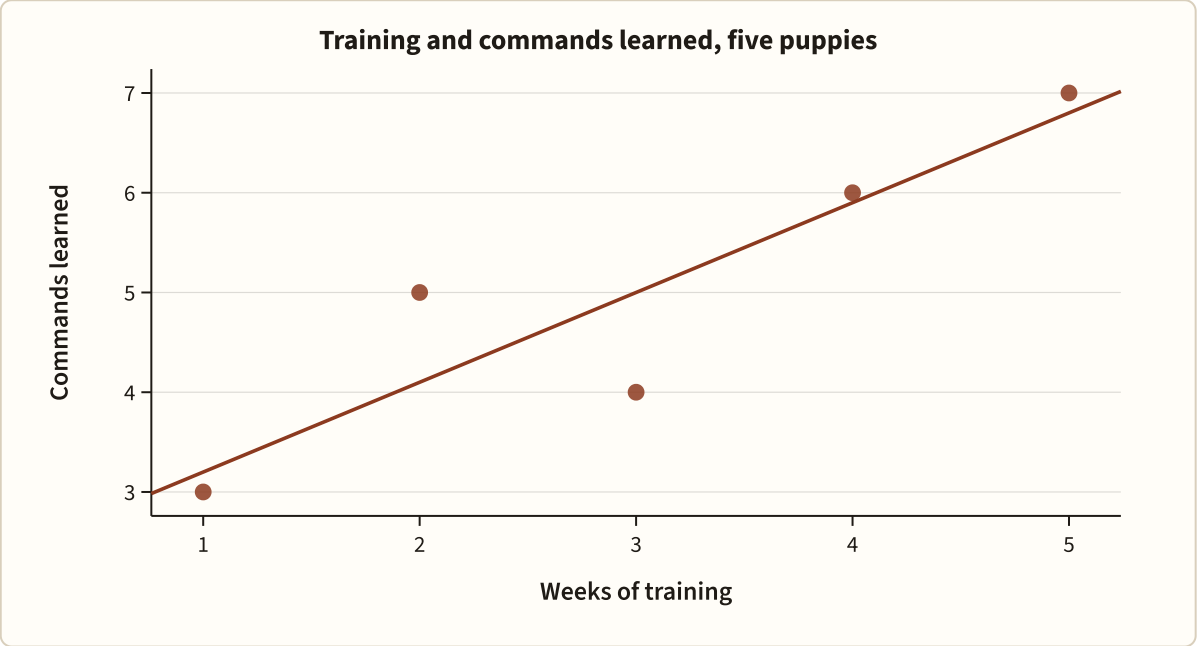
Lurking variables and confounding. In observational data, variables we did not measure may drive both x and y, or be mixed up with x so that their effects cannot be separated (confounding). Families who eat supper together more often may also differ in many other ways that affect their children's schoolwork. A strong association, even r = 0.95, cannot by itself tell us whether x causes y, y causes x, or a lurking variable causes both.

How can we show cause? By an experiment: we choose and impose the values of x on the individuals, assigning them **at random**, so that lurking variables tend to balance out between the groups. If the groups then differ in the response by more than chance would explain, the treatment is the cause. You will learn to design such experiments in Unit 5. When an experiment is impossible (we cannot assign

the weather), evidence for cause must be built carefully from many consistent studies and a believable explanation of how x could affect y.

> **Practice habit.** Ask "what else could explain this?" before believing that x causes y, and fit the line with and without any unusual point to see how much it matters.

Example 1. The five puppies in the 4-H club had (weeks of training, commands learned) = (1, 3), (2, 5), (3, 4), (4, 6), (5, 7). (a) Find the slope of the least-squares line. (b) Find its intercept. (c) When a sixth dog (15 weeks, 5 commands) is added, a calculator gives $\hat{y} = 4.65 + 0.069x$ and r = 0.25. Is the sixth dog influential? Explain.



1. Means: $\bar{x} = \dfrac{15}{5} = 3$ weeks; $\bar{y} = \dfrac{3+5+4+6+7}{5} = \dfrac{25}{5} = 5$ commands.
2. Deviations of x: −2, −1, 0, 1, 2; squares add to 10; $s_x = \sqrt{\dfrac{10}{4}} = 1.5811$.
3. Deviations of y: −2, 0, −1, 1, 2; squares add to 10; $s_y = \sqrt{\dfrac{10}{4}} = 1.5811$.
4. Products of z-scores: each is (x-deviation)(y-deviation)/2.5: 1.6, 0, 0, 0.4, 1.6; sum = 3.6. $r = \dfrac{3.6}{4} = 0.9$.
5. (a) $b = r\dfrac{s_y}{s_x} = 0.9 \times \dfrac{1.5811}{1.5811} = 0.9$ command per week.
6. (b) $a = \bar{y} - b\bar{x} = 5 - 0.9(3) = 2.3$, so $\hat{y} = 2.3 + 0.9x$.
7. (c) Adding the sixth dog changes the slope from 0.9 to 0.069 and r from 0.9 to 0.25, so the dog is very influential. It lies far to the right of the other dogs (15 weeks versus 1 to 5), so it pulls the line toward itself like a weight at the end of a lever.
8. Conclusion in context: for puppies trained 1 to 5 weeks, each extra week goes with about 0.9 more command learned. The one older dog should be reported separately (it may have had a different trainer, or progress may level off with longer training); it should not be allowed to hide the clear pattern among the five puppies, nor be silently deleted.

Answer: (a) b = 0.9. (b) a = 2.3, so $\hat{y} = 2.3 + 0.9x$. (c) Yes: the 15-week dog lies far out in x and changes the slope from 0.9 to about 0.07 and r from 0.9 to 0.25.

Example 2. A survey of families in a town finds that children from homes with more books score higher on reading tests. (a) Does this show that adding books to a home will raise a child's reading score? Answer yes or no. (b) Name a possible lurking variable. (c) Describe an experiment that could test whether giving books to families causes higher reading scores.

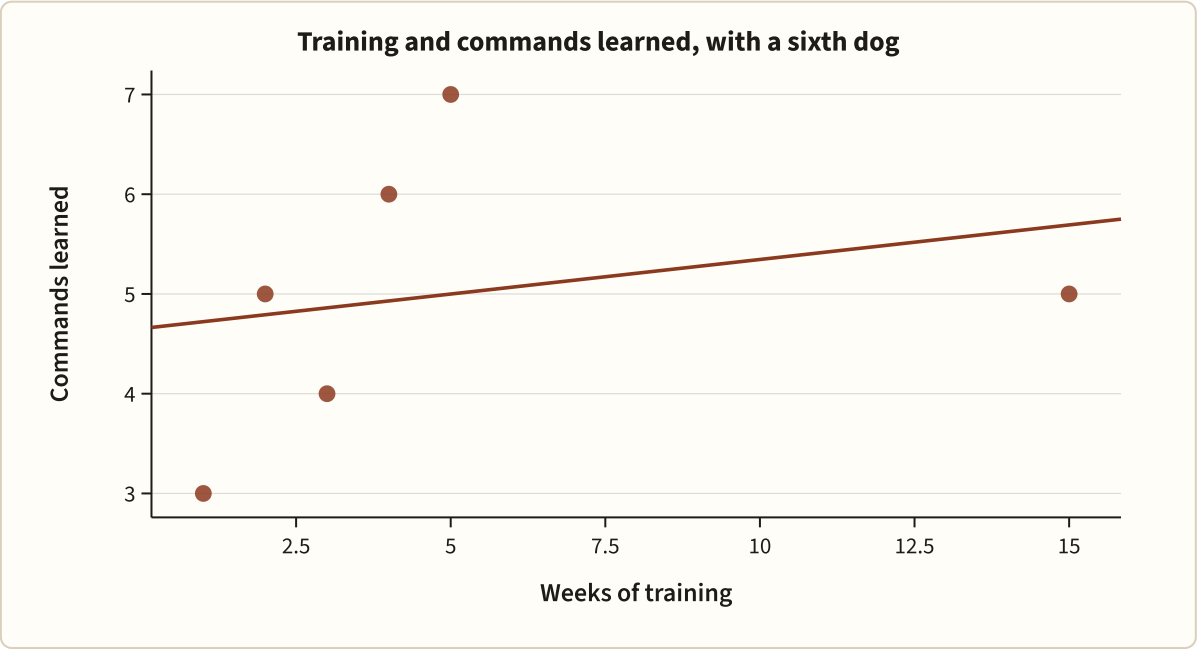
1. (a) These are observational data: nobody assigned the number of books. An association, however strong, does not show cause. Answer: no.
2. (b) A lurking variable: how much time parents spend reading with their children. Parents who read aloud often tend to own more books and also help their children read better, so this variable could create the association by itself.
3. (c) Experiment: take 100 families with young children who own few books. Randomly assign 50 to receive a set of 30 good children's books and 50 to receive none (or the books after the study). After one year, compare the children's reading scores.
4. Why it works: random assignment makes the two groups alike, on average, in parents' habits and everything else, so a clear difference in scores can be credited to the books.

Answer: (a) No. (b) For example, the time parents spend reading with their children, which may increase both the number of books and the reading scores. (c) Randomly assign families to receive books or not, then compare reading scores after a year.

PRACTICE

1. EASY In a small town, on days when the hardware store sells more electric fans, the market also sells more lemonade. Does buying fans make people buy lemonade? Name the lurking variable.
2. EASY For 365 days, the correlation between the daily high temperature and the number of glasses of lemonade sold at a stand is r = 0.6. If you instead correlate the 12 monthly averages of the two variables, would you expect r to be higher or lower than 0.6? Explain.
3. EASY Henry finds r = 0 between two variables and says, "There is no relationship between them." Is Henry necessarily right? Answer yes or no.

4. EASY In the scatterplot of six dogs, which point is the influential point? Give its coordinates (weeks, commands).



5. MEDIUM Use the five-puppy line $\hat{y} = 2.3 + 0.9x$ and the six-dog line $\hat{y} = 4.654 + 0.0692x$ to predict the number of commands learned by a puppy trained for 4 weeks. Give both predictions, rounding to two decimal places.

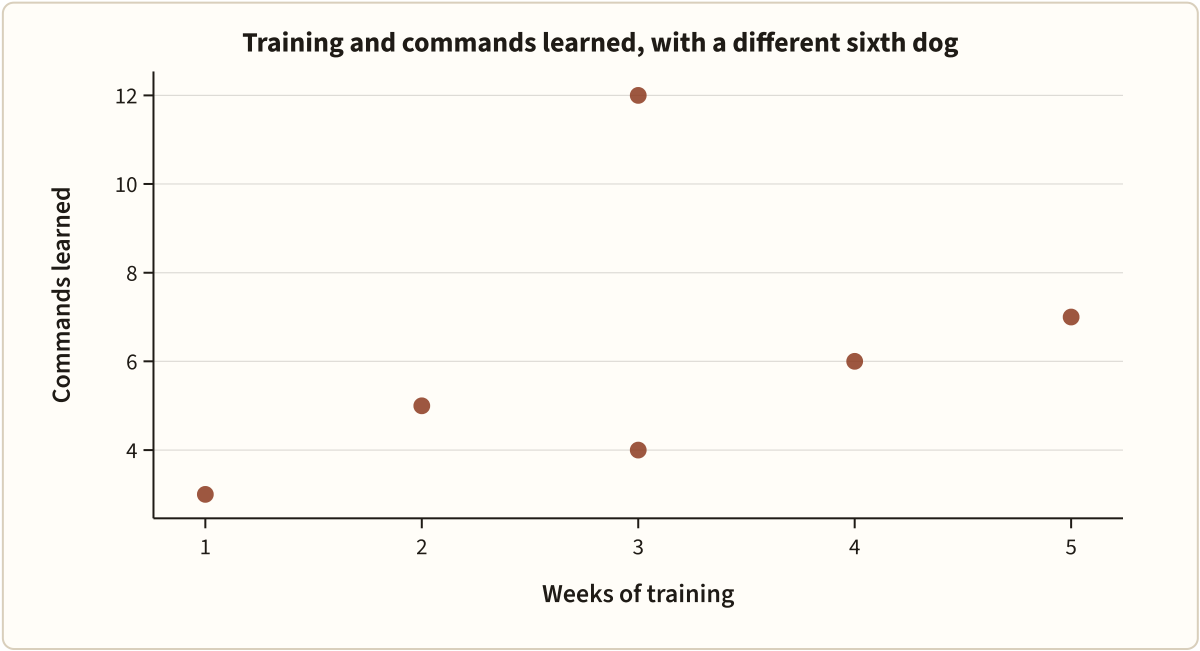
6. MEDIUM Find the residual of the 15-week dog (5 commands) with respect to the five-puppy line $\hat{y} = 2.3 + 0.9x$.

7. MEDIUM Without the 15-week dog, r = 0.9; with it, r = 0.25. Explain what this tells you about r, and what you should report.

8. MEDIUM A survey of church families finds that children who eat supper with their families more nights per week tend to earn higher grades. (a) Can we conclude that family suppers cause higher grades? Answer yes or no. (b) Name a possible lurking variable.

9. MEDIUM For lambs from birth to 4 months old, the line $\hat{y} = 10 + 12x$ predicts weight y (pounds) from age x (months). (a) What weight does it predict for a 5-year-old sheep (60 months)? (b) Why is this prediction not believable?

10. **HARD** Instead of the 15-week dog, suppose the sixth dog was trained for 3 weeks and learned 12 commands, giving the data (1, 3), (2, 5), (3, 4), (3, 12), (4, 6), (5, 7). Use a calculator to find the least-squares line. Round a and b to two decimal places.



11. **HARD** Compare the line in Problem 10 with the five-puppy line $\hat{y} = 2.3 + 0.9x$. Which changed, the slope or the intercept? Explain why this outlier, unlike the 15-week dog, is not very influential on the slope.

12. **HARD** In a county survey, orchards that sprayed their apple trees more times during the summer lost a larger share of their crop to insects. Ruth concludes that spraying causes crop loss. Give a better explanation, and say what kind of study could settle the question.

13. **CHALLENGE** For all six dogs including the 15-week dog, a calculator gives $\bar{x} = 5$, $s_x = 5.0990$, $\bar{y} = 5$, $s_y = 1.4142$ and r = 0.2496. Use the formulas $b = r\dfrac{s_y}{s_x}$ and $a = \bar{y} - b\bar{x}$ to find the line. Round b to three decimal places and a to two decimal places (use the unrounded b when you compute a), and explain in terms of s_x why the slope became so small.

COMMON MISTAKES

- Deleting an unusual point because it spoils the pattern. Investigate it, fit the line with and without it, and report both.
- Looking only at residuals to find influential points. An influential point often has a small residual because it pulls the line toward itself; check by refitting without it.
- Applying a correlation from averaged data (county or monthly averages) to individuals. Individuals vary much more than averages do.
- Concluding cause from a strong correlation in observational data. Name possible lurking variables, and remember that only a randomized experiment shows cause directly.

ANSWERS TO PRACTICE

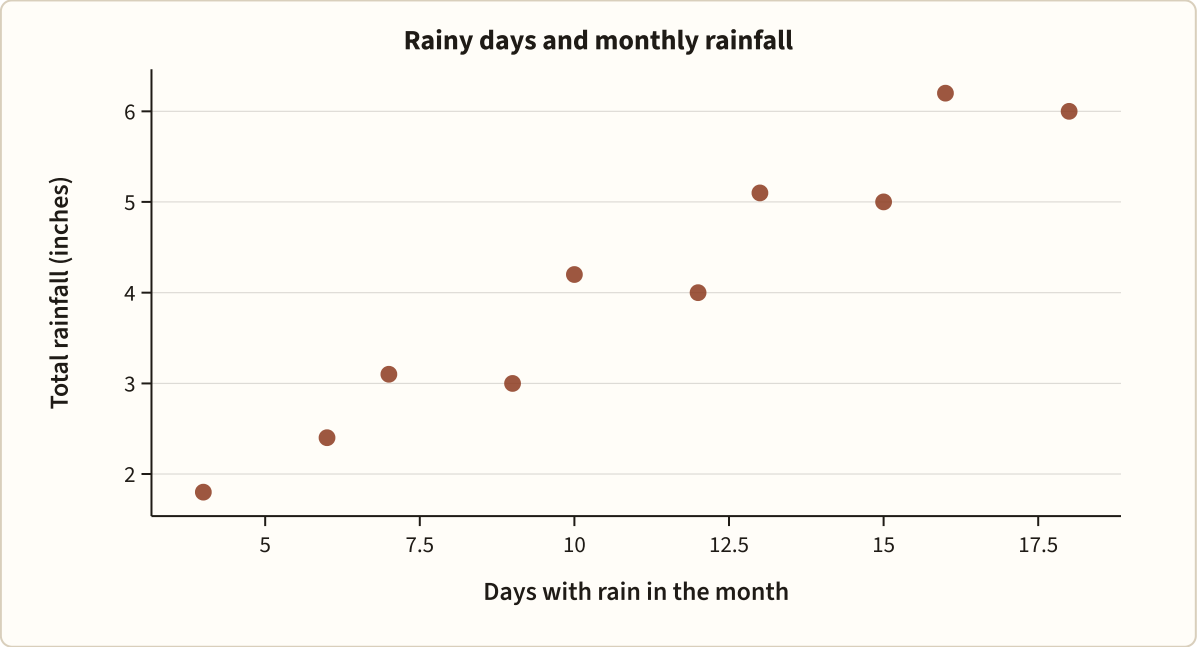
1. No. The lurking variable is the temperature: hot days lead people to buy both fans and lemonade.
2. Higher. Averaging over each month smooths out the day-to-day ups and downs (a cloudy Saturday, a busy holiday), so the 12 monthly points lie much closer to a line than the 365 daily points.
3. No. r = 0 means there is no linear relationship; there may still be a strong curved relationship (like temperature over a day).
4. (15, 5): it lies far to the right of the other points and pulls the line nearly flat.
5. Five puppies: $2.3 + 0.9(4) = 5.90$ commands. Six dogs: $4.654 + 0.0692(4) = 4.93$ commands. The one influential dog changes the prediction by about one command.
6. $\hat{y} = 2.3 + 0.9(15) = 15.8$; residual $= 5 - 15.8 = -10.8$ commands. Judged by the puppies' pattern, the older dog is far off, which is why it drags the line so much when it is included.
7. r is not resistant: a single point far out in x changed it from strong to weak. Report the analysis with and without the unusual dog, explain why it differs (more weeks than any other dog, perhaps a different trainer), and do not let one dog decide the conclusion about all puppies.
8. (a) No; these are observational data. (b) For example, the parents' work schedules or how involved the parents are in schoolwork: families with more time together may also spend more time helping with homework.
9. (a) $\hat{y} = 10 + 12(60) = 730$ pounds. (b) This is extrapolation far beyond the data (0 to 4 months). Growth slows as sheep mature; a grown sheep weighs far less than 730 pounds.
10. $\bar{x} = 3$, $\bar{y} = \dfrac{37}{6} = 6.1667$; the calculator gives b = 0.90 and a = 6.1667 − 0.9(3) = 3.47, so $\hat{y} = 3.47 + 0.90x$ (and r drops to 0.40).
11. Only the intercept changed (2.3 to 3.47); the slope stayed 0.9. The new dog sits exactly at $\bar{x} = 3$, so it has no leverage to tilt the line; it only raises $\bar{y}$ and shifts the whole line up by about 1.17. It is an outlier in y (large residual) that lowers r, but it does not pull the slope. A point far out in x, like the 15-week dog, acts on a long lever and tilts the line.
12. The amount of insect pressure is a lurking variable: orchards with heavy insect problems spray more often and also lose more fruit, so the association runs from the insects to both variables. An experiment could settle it: in orchards with similar insect pressure, randomly assign blocks of trees to different spraying schedules and compare the crop loss.
13. $b = 0.2496 \times \dfrac{1.4142}{5.0990} = 0.069$; $a = 5 - 0.069225(5) = 4.65$. $\hat{y} = 4.65 + 0.069x$. The far-out dog both lowered r and greatly increased s_x (from 1.58 to 5.10) while s_y changed little (1.58 to 1.41), and both effects shrink $b = r\dfrac{s_y}{s_x}$.

Unit 4 test

You have 55 minutes. You may use a scientific calculator with one- and two-variable statistics; no statistical tables are needed. Show your work: write each formula, the numbers you substitute and the result, and answer every interpretation question in a complete sentence in context. Keep full precision until the last step and round as each problem states; when no rounding is stated, give exact values.

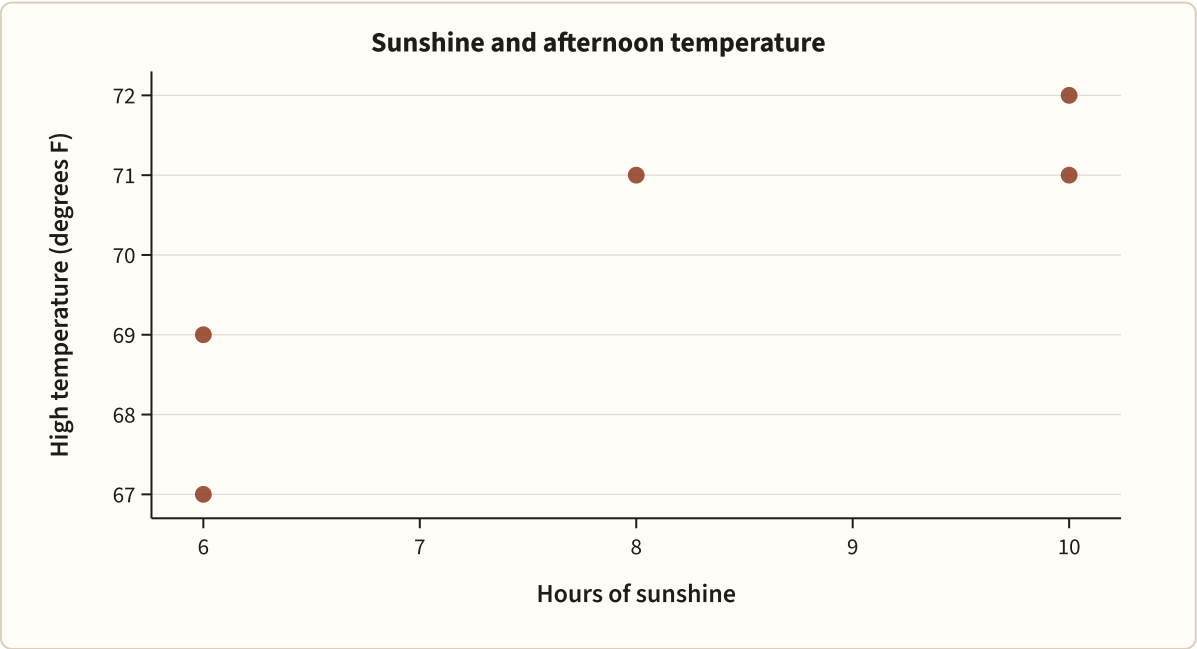
1. **5 PTS** On 12 days in March, a maple farmer recorded the overnight low temperature and the gallons of sap collected the next day. Which variable is explanatory and which is the response? Explain briefly.

2. **6 PTS** A weather station recorded, for ten months, the number of days with rain and the total rainfall in inches. Describe the relationship shown in the scatterplot, in context.



3. **5 PTS** For 25 hiking trails in a state park, the correlation between trail length in miles and the average hiking time in hours is $r = 0.88$. What is the correlation if length is measured in kilometers and time in minutes?

4. **8 PTS** On five spring days a weather watcher recorded the hours of sunshine x and the afternoon high temperature y (°F): (6, 67), (6, 69), (8, 71), (10, 71), (10, 72). Compute r by hand using z-scores. Show $\bar{x}$, $\bar{y}$, s_x and s_y.



5. **8 PTS** For lambs on a sheep farm, birth weight x has mean 9 pounds and standard deviation 1.5 pounds, weaning weight y has mean 75 pounds and standard deviation 8 pounds, and r = 0.6. Find the least-squares regression line for predicting weaning weight from birth weight.

6. **6 PTS** Using the lamb line $\hat{y} = 46.2 + 3.2x$: (a) predict the weaning weight of a lamb that weighed 10 pounds at birth; (b) interpret the slope in context.

7. **5 PTS** Should the intercept 46.2 of the lamb line be interpreted as the weaning weight of a lamb that weighed 0 pounds at birth? Explain.

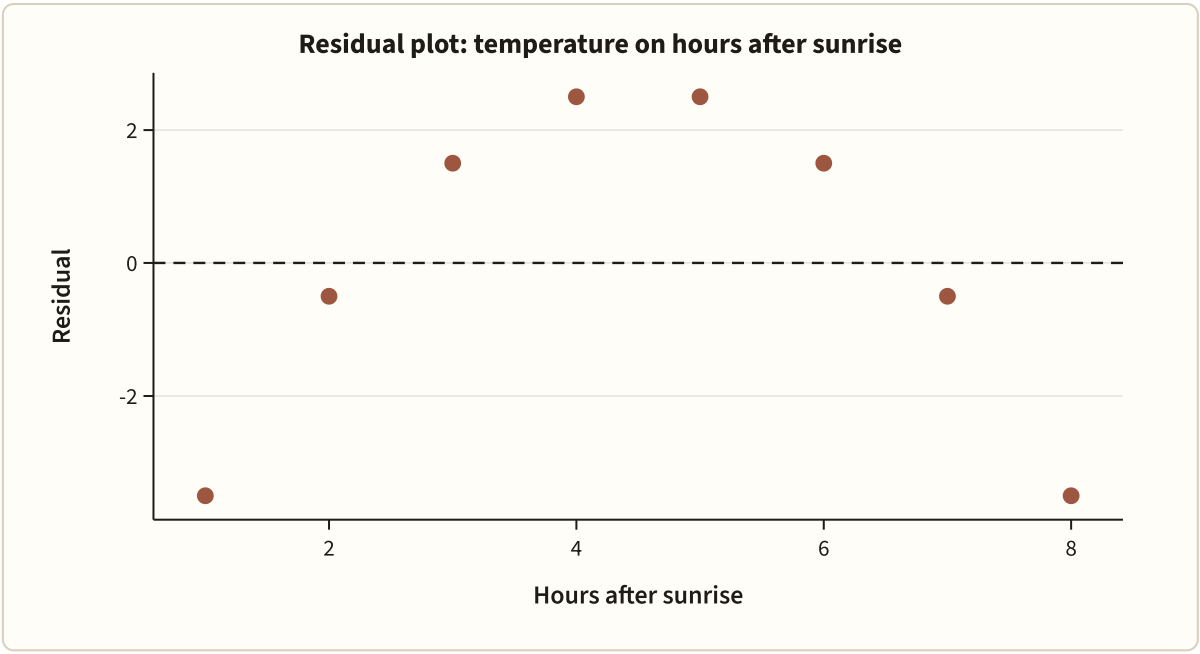
8. **6 PTS** A lamb weighed 8 pounds at birth and 70 pounds at weaning. Find its residual from the line $\hat{y} = 46.2 + 3.2x$ and say what it means.

9. **8 PTS** A family that heats with wood regressed the cords of firewood burned each month on the month's average outdoor temperature (°F). Use the output. (a) Predict the firewood used in a month averaging 30°F. (b) Find the correlation r, rounded to three decimal places. (c) Interpret R-Sq in context.

Regression output: firewood used (cords per month) on average temperature (degrees F)

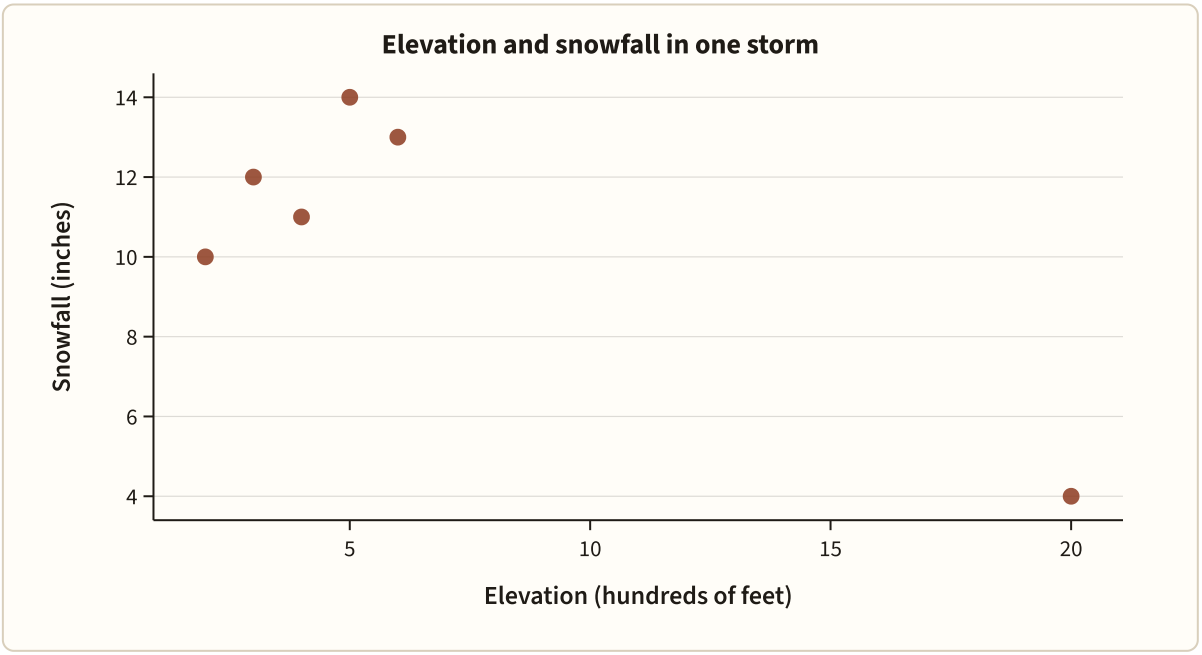
Predictor	Coef
Constant	1.6
Temperature	-0.025
S	0.12
R-Sq	90.3%

10. **6 PTS** Grace fit a line to predict the temperature from the number of hours after sunrise on a summer day. The residual plot is shown. Is a straight line an appropriate model? Answer yes or no, and explain.



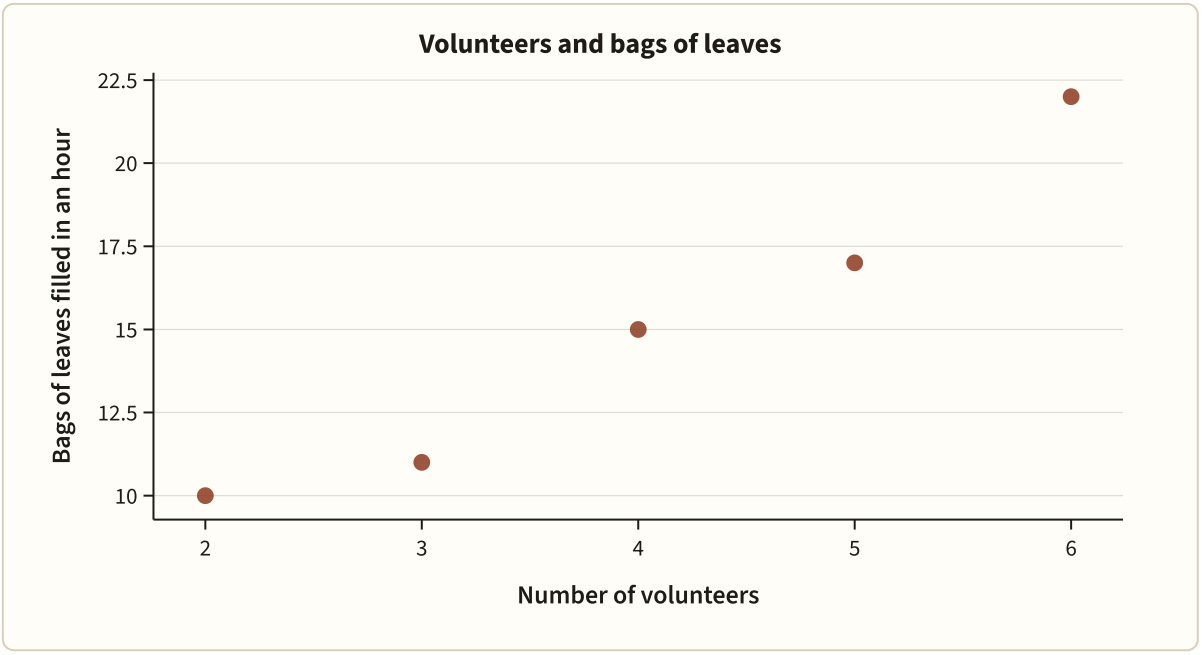
11. **6 PTS** A regression to predict the weight of a hay bale (pounds) from its length (inches) has $r^2 = 0.72$ and s = 3.5 pounds. Interpret both values in context.

12. **8 PTS** After one snowstorm, snowfall was measured at six places at different elevations. For the five lower places, the least-squares slope is 0.8 inch per hundred feet; with all six places it is about −0.44. (a) Which point is influential? Give its coordinates (elevation, snowfall). (b) Explain why that point has so much effect on the line.



13. **8 PTS** Over 30 years in a farming county, the number of new tractors sold each year and the average corn yield per acre both rose steadily, and r = 0.9. (a) Can we conclude that buying tractors caused the higher yields? Answer yes or no. (b) Name a lurking variable and explain how it could produce this association.

14. **8 PTS** A church youth group raked leaves for neighbors. On five Saturdays they recorded the number of volunteers x and the number of bags of leaves filled in one hour y: (2, 10), (3, 11), (4, 15), (5, 17), (6, 22). Find the least-squares regression line and interpret its slope.



15. 7 PTS In an archery contest, scores in the first and second rounds both have mean 70 and standard deviation 10, and the correlation between a contestant's two scores is r = 0.6. Samuel scored 85 in the first round. (a) Use the regression line to predict Samuel's second-round score. (b) Explain why the prediction is closer to 70 than 85 is.

UNIT 4 TEST: ANSWER KEY

1. Explanatory: overnight low temperature. Response: gallons of sap. The night's temperature helps determine how much sap flows the next day, not the other way around.
2. There is a strong, positive, linear association between the number of rainy days and total rainfall: months with more rainy days tend to have more rain, about 0.3 inch more for each additional rainy day, with no outliers.
3. r = 0.88. Changing units does not change the z-scores, so r is unchanged.
4. $\bar{x} = 8$, $\bar{y} = 70$. Deviations of x: −2, −2, 0, 2, 2 (squares sum to 16, $s_x = \sqrt{16/4} = 2$); deviations of y: −3, −1, 1, 1, 2 (squares sum to 16, $s_y = 2$). z_x: −1, −1, 0, 1, 1; z_y: −1.5, −0.5, 0.5, 0.5, 1. Products: 1.5, 0.5, 0, 0.5, 1; sum = 3.5. $r = \dfrac{3.5}{4} = 0.875$: a strong positive linear relationship.
5. $b = 0.6 \times \dfrac{8}{1.5} = 3.2$; $a = 75 - 3.2(9) = 75 - 28.8 = 46.2$. $\hat{y} = 46.2 + 3.2x$.
6. (a) $\hat{y} = 46.2 + 3.2(10) = 78.2$ pounds. (b) For each additional pound of birth weight, the predicted weaning weight increases by about 3.2 pounds.
7. No. A birth weight of 0 pounds is impossible and far outside the data (birth weights near 9 pounds). The intercept only sets the height of the line; it has no meaning here.
8. $\hat{y} = 46.2 + 3.2(8) = 71.8$ pounds; residual $= 70 - 71.8 = -1.8$ pounds. The lamb weighed 1.8 pounds less at weaning than predicted from its birth weight.
9. (a) $\hat{y} = 1.60 - 0.025(30) = 1.60 - 0.75 = 0.85$ cord. (b) $\sqrt{0.903} = 0.950$, and the slope is negative, so r = −0.950. (c) About 90.3% of the month-to-month variation in firewood burned is accounted for by the linear relationship with average temperature.
10. No. The residuals form an upside-down U (negative at both ends, positive in the middle), so the relationship is curved: the temperature rises quickly in the morning and levels off in the afternoon. A curve would fit better than a line.
11. $r^2 = 0.72$: about 72% of the variation in bale weight is accounted for by the linear relationship with bale length. s = 3.5: predictions of a bale's weight from its length are typically off by about 3.5 pounds.
12. (a) (20, 4): 2,000 feet of elevation and 4 inches of snow. (b) It lies far out in the x direction (20 hundred feet, versus 2 to 6 for the others), so like a weight at the end of a lever it pulls the line toward itself, turning a positive slope into a negative one. The line should be reported with and without it, and the reason for the low snowfall (perhaps the storm's wind pattern) should be investigated.
13. (a) No. (b) Time and the general improvement of farming: over 30 years, better seed varieties, fertilizers and methods raised yields, while growing farms and prosperity led to more tractor purchases. Both variables rose with the years, so they are correlated even if tractors had little direct effect on yield.

14. $\bar{x} = 4$, $\bar{y} = 15$, $s_x = 1.5811$, $s_y = 4.8477$, r = 0.97849. $b = 0.97849 \times \dfrac{4.8477}{1.5811} = 3$; $a = 15 - 3(4) = 3$. $\hat{y} = 3 + 3x$: each additional volunteer adds about 3 bags to the predicted number filled in an hour.
15. (a) $z_x = \dfrac{85 - 70}{10} = 1.5$; predicted $z_y = 0.6 \times 1.5 = 0.9$; $\hat{y} = 70 + 0.9(10) = 79$. (b) Because r is less than 1, the prediction is only 0.9 standard deviation above the mean instead of 1.5: regression toward the mean. A high first-round score reflects skill plus some good fortune, and the good fortune is not expected to repeat.

UNIT 5

Collecting Data: Samples and Experiments

In Units 1 to 4 you learned to describe data with graphs and numbers. This unit steps back and asks a question that comes first: where did the data come from? A conclusion can be no better than the data behind it, and the way data are collected decides what they can honestly tell us. You will learn how to choose samples that represent a population, how surveys go wrong, and how to design experiments that can show cause and effect.

The central idea is that chance, used on purpose, is our friend. Letting chance choose a sample protects against bias, so we can generalize from the sample to the population; letting chance assign treatments makes the groups in an experiment alike, so we can conclude that a treatment caused a difference. You will meet the four principles of experimental design (comparison, random assignment, replication and control) along with blocking and matched pairs, ideas Ronald Fisher worked out in agricultural experiments in the 1920s.

These ideas connect backward to Unit 4, where you learned that association is not causation, and forward to the rest of the course. Probability (Unit 6) describes the behavior of chance, sampling distributions (Unit 9) describe how random samples vary, and every confidence interval and significance test in Units 10 to 13 depends on data collected by random sampling or random assignment. After this unit you will also plan and begin collecting the data for your statistics project.

PREREQUISITES

- Compute fractions, decimals and percents of a whole (Algebra I).
- Compute the mean of a small data set and the range (Lessons 2.1 and 2.2).
- Read dotplots, boxplots, bar graphs and tables (Units 1 and 2).
- Explain why association does not imply causation and what a lurking variable is (Lessons 4.1 and 4.5).

MASTERY CHECKLIST

☐ I can identify the population and the sample in a study and explain why we usually sample instead of taking a census.
☐ I can explain what bias is, recognize convenience and voluntary response samples, and say which way they are likely to be biased.
☐ I can choose a simple random sample with a table of random digits or a random number generator (label, randomize, select).
☐ I can describe stratified, cluster, systematic and multistage samples and explain why stratifying gives more precise results.

☐ I can tell sampling error from nonsampling errors (undercoverage, nonresponse, response bias, wording) and compute a response rate.
☐ I can write a fair, clear survey question and explain why a larger sample does not fix bias.
☐ I can tell an observational study from an experiment and identify factors, treatments, experimental units and the response variable.
☐ I can explain confounding and name a plausible confounding variable in context.
☐ I can state and explain the four principles of experimental design and carry out random assignment for a completely randomized design.
☐ I can explain placebos, the placebo effect, and single- and double-blind experiments, and interpret 'statistically significant'.
☐ I can describe randomized block and matched pairs designs and analyze matched pairs results with the differences.
☐ I can state the scope of a study's conclusions: random sampling allows generalizing, random assignment allows concluding cause.

LESSON 5.1 · 2 SESSIONS OF 45 MINUTES

Populations, Samples and Bias

OBJECTIVES

- Identify the population and the sample in a statistical study, and explain why we usually sample instead of taking a census.
- Explain what bias is and why taking a larger sample by the same method does not remove it.
- Recognize convenience samples and voluntary response samples, and explain in context which way each is likely to be biased.
- Explain why letting chance choose the sample protects against bias.

VIDEO

WATCH

Census vs. Sample (1 of 2: Comparison)

Eddie Woo · 10:38

When to watch: Before Example 1: census versus sample.

https://www.youtube.com/watch?v=4oNyyfBsWZw

BACKUP

Census vs. Sample (2 of 2: A Story Illustrating Bias)

Eddie Woo · 3:08

When to watch: After Example 1: a short story of a biased sample (survivorship bias in wartime).

https://www.youtube.com/watch?v=zuKQ7JVPQhc

KEY TERMS

population	The entire group of individuals we want information about, such as all the apples in an orchard or all the families in a parish.
sample	The part of the population that we actually examine in order to learn about the whole population.
bias	A systematic error in a method of collecting data: the method tends to miss the truth in the same direction again and again, not just by bad luck once.
convenience sample	A sample made of the individuals that are easiest to reach. It usually leaves out whole kinds of individuals, so it is usually biased.

voluntary response sample	A sample made of people who choose themselves by answering a general invitation, such as a call-in radio question, a comment box or a website poll. People with strong opinions are overrepresented.
random sample	A sample chosen by a chance process (drawing names from a hat, random digits, a random number generator), so that no one's judgment or convenience decides who is included.

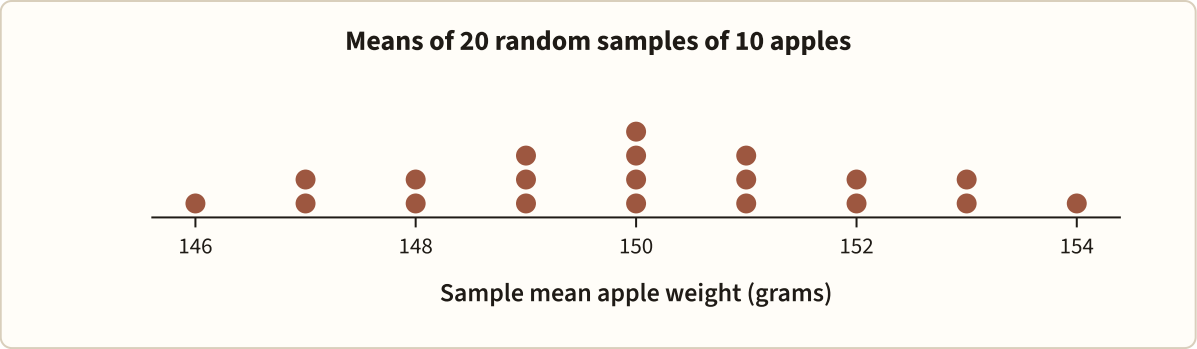
CORE IDEAS

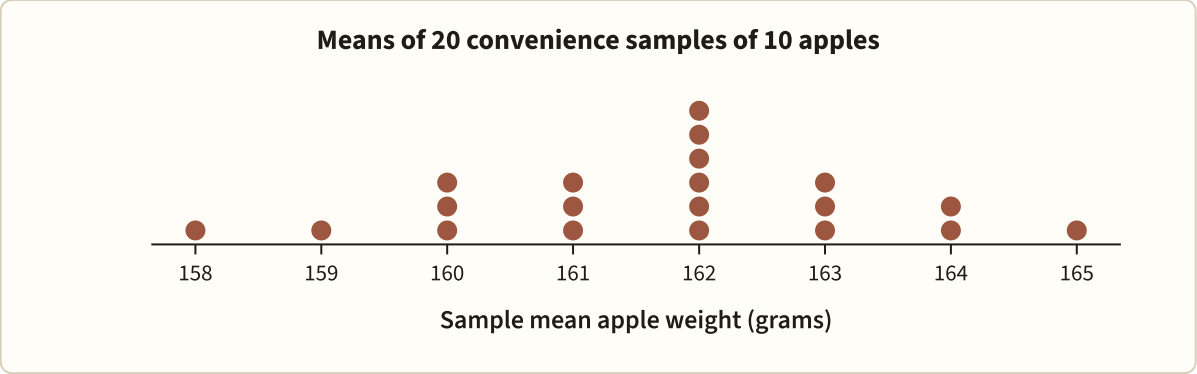
Every statistical question is about some group. The **population** is the whole group we want to learn about: all the apples in an orchard, all 300 families in a parish directory, all the boards a sawmill cuts this week. Examining every individual in the population is a **census**. A census is usually too slow, too costly or simply impossible, so we examine a **sample**, a part of the population, and use what we find to draw conclusions about the whole. Reasoning from the part to the whole is the heart of statistics.

Why not just measure everything? Sometimes measuring destroys what is measured: to learn how much weight a rope can hold you must break it, and to check the sweetness of apples you must cut them. Sometimes the population is huge or still growing (every egg a farm will lay this year). And a well-chosen sample of a few hundred can give results nearly as good as a census for a tiny part of the cost. The catch is that a sample is useful only if it *represents* the population, and that depends entirely on how the sample was chosen.

A method of sampling is **biased** if it systematically favors certain outcomes, so that its results tend to miss the truth in the same direction every time it is used. Think of a bathroom scale that reads 3 pounds heavy. Weighing yourself ten times and averaging does not help, because every reading is off in the same direction. Bias is a flaw in the *method*, not bad luck with one sample, and that is why collecting more data by the same method cannot fix it.

The easiest sample to get is a **convenience sample**: the individuals that are easiest to reach. Suppose a farmer estimates the mean weight of the apples in her orchard by picking 10 apples from the low branches of the trees along the lane. Those trees get extra sun and water, and the apples within easy reach tend to be the big ones that catch her eye. The two dotplots show the sample mean from 20 different random samples of 10 apples and from 20 different convenience samples of 10 apples, taken from an orchard whose apples really average 150 grams.





The random samples scatter on both sides of 150 grams, some a little high and some a little low, with no favorites. The convenience samples cluster around 162 grams, about 12 grams too high, again and again. That is what bias looks like: not more scatter, but scatter centered in the wrong place.

A **voluntary response sample** is made of people who choose themselves by answering a general invitation: a call-in question on the radio, a comment box, a poll on a website. People who care strongly, and especially people who are unhappy, are much more likely to respond, so strong opinions are overrepresented. If a library asks visitors to 'leave a note if you would like longer hours', the notes come mostly from people who want longer hours; the many patrons who are content say nothing.

The cure for bias in choosing a sample is to let chance choose. In a **random sample**, a chance process such as drawing names from a hat, a table of random digits or a random number generator picks the individuals; neither the researcher's judgment nor the individuals' own wishes decide who is in. Chance has no favorites: it does not prefer big apples, nearby trees or people with strong opinions. Random samples still differ from one another, but their errors go both ways, and, as you will see in later units, the size of those errors can be predicted. Lesson 5.2 shows how to choose random samples in practice.

> **Practice habit.** Before you trust any number that comes from a sample, ask two questions: Who was in the sample? How were they chosen? If people chose themselves or were chosen for convenience, expect bias and ask which direction it runs.

WORKED EXAMPLES

Example 1. The Hope Valley Library wants to know whether its patrons would like the library to open on Saturday afternoons. The librarian puts a card box on the front desk with a sign: 'Would you like Saturday afternoon hours? Tell us!' In two weeks, 60 cards are turned in, and 48 of them say yes. (a) Describe the population and the sample. (b) What percent of the cards said yes? (c) Is this percent likely to be too high or too low as an estimate for all patrons? Explain.

1. (a) The population is the whole group the library wants to learn about: all of the library's patrons.
2. The sample is the group actually examined: the 60 patrons who turned in cards.
3. (b) Percent yes $= \dfrac{48}{60} = 0.80$, which is 80%.
4. (c) The patrons chose themselves by deciding to fill out a card, so this is a voluntary response sample.
5. Patrons who want Saturday hours have a reason to write a card; patrons who are content with the present hours mostly do not bother.
6. So people who want the change are overrepresented, and 80% is probably too high as an estimate of the percent of all patrons who want Saturday hours.

Answer: (a) Population: all patrons of the library; sample: the 60 patrons who turned in cards. (b) 80%. (c) Probably too high: this is a voluntary response sample, and patrons who want Saturday hours are much more likely to respond than patrons who are satisfied.

Example 2. The Miller family has 8 pear trees. The table shows each tree's harvest last fall. Trees A, B and C stand next to the house, where rain from the roof gutters waters them. (a) Find the mean harvest of all 8 trees (the population mean). (b) Mr. Miller estimates the mean by checking only trees A, B and C, because they are closest. Find the mean of this convenience sample. (c) By how many bushels does the convenience sample miss the population mean, and why would choosing 3 trees by drawing letters from a hat be better?

Miller family pear harvest

Tree	Location	Harvest (bushels)
A	Next to house	14
B	Next to house	15
C	Next to house	13
D	Orchard	9
E	Orchard	10
F	Orchard	8
G	Orchard	11
H	Orchard	8

1. (a) Add all 8 harvests: $14 + 15 + 13 + 9 + 10 + 8 + 11 + 8 = 88$ bushels.
2. Population mean $= \dfrac{88}{8} = 11$ bushels per tree.
3. (b) Sample mean $= \dfrac{14 + 15 + 13}{3} = \dfrac{42}{3} = 14$ bushels per tree.
4. (c) The convenience sample misses by $14 - 11 = 3$ bushels, too high.
5. The trees near the house get extra water, so they yield more than the other trees. Choosing the closest trees always picks exactly those trees, so this method is off in the same direction every time: it is biased.
6. Drawing letters from a hat gives every tree the same chance. Some random samples would come out high and some low, but the method does not favor the watered trees, so it is not biased.

Answer: (a) 11 bushels per tree. (b) 14 bushels per tree. (c) It is 3 bushels too high. The method always picks the well-watered trees, so it is biased upward; drawing from a hat lets chance choose, giving every tree the same chance, so the errors go both ways instead of always up.

PRACTICE

1. EASY A beekeeper has 80 hives. She draws 12 hive numbers from a hat and checks those 12 hives for mites. (a) How many individuals are in the population? (b) How many are in the sample?
2. EASY Is each of these a census? Answer yes or no. (a) An usher counts every person at the Sunday service to record attendance. (b) A lumberyard checks 5 boards from each truckload for warping.
3. EASY A rope company wants to know how much weight its climbing ropes can hold before they break. Explain why the company must use a sample rather than a census.
4. EASY The county fair's website posted the question 'Should the fair add a third day?' Of the 150 visitors who answered, 132 said yes. (a) What percent of those who answered said yes? (b) What kind of sample is this, and why should the fair board be cautious about the result?
5. MEDIUM To estimate the average depth of snow on his 200-acre farm after a storm, Henry measures the depth at 10 spots along the north wall of his barn, where the wind piles up drifts. (a) What type of sample is this? (b) Will his estimate probably be too high or too low? Explain.
6. MEDIUM Grace wants to know what percent of the strawberries in her family's large patch have been pecked by birds. She checks the 60 berries closest to the woods, where the birds perch, and finds 18 pecked. (a) What percent of her sample was pecked? (b) Is this likely to overestimate or underestimate the percent for the whole patch? Explain.
7. MEDIUM A farmer's six wheat fields had the yields shown in the table. Fields P and Q are beside the road, and they get the most attention. (a) Find the mean yield of all six fields. (b) Find the mean yield of the convenience sample made of fields P and Q. (c) By how many bushels per acre does the convenience sample overestimate the mean of all six fields?

Wheat yields of six fields

Field	Beside the road?	Yield (bushels per acre)
P	Yes	52
Q	Yes	50
R	No	40
S	No	38
T	No	42
U	No	42

8. MEDIUM The youth group leader needs 6 of the 30 members to answer some questions about next summer's service trip. He plans to choose 'a good mix' of members himself. Explain why writing all 30 names on equal slips, mixing them in a hat and drawing 6 is a better method.

9. MEDIUM A radio station asked listeners to call in and say whether its Saturday morning bluegrass program should be replaced by a gardening show. 200 people called, and 70% said no. The station manager says, 'Next time we will get 2,000 calls, so the result will be accurate.' Will a larger voluntary response sample remove the bias? Answer yes or no, and explain.

10. HARD The county newspaper, which has 1,200 subscribers, printed a coupon asking readers whether it should add a weekly recipe page. 90 readers mailed the coupon back, and 72 of them said yes. (a) What percent of those who responded said yes? (b) What percent of the subscribers responded? (c) Explain why the newspaper should not conclude that about 80% of its subscribers want a recipe page.

11. HARD Ruth wants to estimate the mean number of hours per week that the 480 students at her school spend reading for pleasure. She surveys the 25 students who are in the school library during lunch on Monday. Identify the population and the sample, name the type of sample, and explain which way the estimate is likely to be biased.

12. HARD A feed store's scale reads 4 pounds too heavy for every sack. A farmer weighs 10 sacks of oats on it and gets a mean of 54 pounds. (a) What is the best estimate of the true mean weight of these 10 sacks? (b) If he weighed 100 sacks on the same scale, would the bias go away? Answer yes or no.

13. CHALLENGE The parish council wants to know how many of its 300 families would come to a Saturday workday to repair the church hall. One member suggests asking the families who stay for coffee after the Sunday service. Explain why this sample is likely to be biased and in which direction, and describe a better way to choose 30 families.

COMMON MISTAKES

- Thinking a bigger sample fixes bias. A larger sample gathered by the same biased method is just as biased; fix the method first, then worry about the size.
- Mixing up the sample and the population. The population is everyone you want to learn about; the sample is only the individuals you actually examine.
- Calling a haphazard choice 'random'. Picking whatever comes to hand, or 'a good mix', is not random; random means a chance process (a hat, random digits, a random number generator) does the choosing.
- Saying a voluntary response sample is fair because 'anyone could respond'. The respondents chose themselves, so people with strong opinions are overrepresented.

ANSWERS TO PRACTICE

1. (a) 80 hives (all of her hives). (b) 12 hives.
2. (a) Yes: every individual in the population (everyone at the service) is counted. (b) No: only some of the boards are checked, so this is a sample.

3. Finding the breaking strength of a rope means pulling it until it breaks, which destroys it. A census would destroy every rope the company makes, leaving nothing to sell, so the company tests a sample of ropes and uses the results to describe all of them.
4. (a) $\frac{132}{150} = 0.88$, so 88%. (b) A voluntary response sample: visitors chose whether to answer, and people who are enthusiastic about the fair (and want more of it) are the most likely to bother, so 88% probably overstates support among all fairgoers.
5. (a) A convenience sample: he measured where it was easy, near the barn. (b) Too high. Wind piles snow into drifts against the barn wall, so those spots are deeper than the open fields, and every spot he chose is a drift spot.
6. (a) $\frac{18}{60} = 0.30$, so 30%. (b) Overestimate: berries next to the woods are the easiest for birds to reach, so they are pecked more often than berries farther from the trees.
7. (a) $\frac{52 + 50 + 40 + 38 + 42 + 42}{6} = \frac{264}{6} = 44$ bushels per acre. (b) $\frac{52 + 50}{2} = 51$ bushels per acre. (c) $51 - 44 = 7$ bushels per acre.
8. When the leader chooses, his own judgment decides who is in, and without meaning to he may pick members he knows well, who talk the most or who are most involved, so their views would be overrepresented. Drawing from a hat lets chance choose: every member has the same chance, nobody is favored, and any error in the result is due to chance rather than to the leader's preferences.
9. No. A larger voluntary response sample is still made of people who chose to call, mostly those with strong feelings (here, devoted bluegrass fans). More calls make the result steadier, but it stays off in the same direction; bias comes from the method, not from the sample size.
10. (a) $\frac{72}{90} = 0.80$, so 80%. (b) $\frac{90}{1200} = 0.075$, so 7.5%. (c) Only 7.5% of subscribers chose to respond, and readers who love recipes had the most reason to clip and mail the coupon. This voluntary response sample probably overstates interest, and it tells us almost nothing about the 92.5% who did not answer.
11. Population: all 480 students at the school. Sample: the 25 students in the library at lunch on Monday. This is a convenience sample. Students who spend their lunch in the library are probably more eager readers than students in general, so her estimate of the mean reading time is likely too high.
12. (a) Each reading is 4 pounds too high, so the true mean is about $54 - 4 = 50$ pounds. (b) No. Every reading would still be 4 pounds too heavy. Weighing more sacks reduces chance variation, but the 4-pound bias stays; only fixing the scale (the method) removes it.

13. Families who stay for coffee hour are usually the most involved in parish life, and they are more likely than other families (for example, those with small children or long drives home) to volunteer for a workday. This convenience sample would probably overestimate how many families will come. Better: number the families in the directory 001 to 300 and let chance choose 30 of them, by drawing numbered slips from a hat or by using random digits, skipping repeats; then contact all 30 chosen families. Every family has the same chance, so the council's choices do not favor the most active families.

LESSON 5.2 · 2 SESSIONS OF 45 MINUTES

Random Sampling Methods

OBJECTIVES

- Select a simple random sample with a table of random digits or a random number generator, using the steps label, randomize, select.
- Explain how a stratified random sample is chosen and why strata of similar individuals give more precise results.
- Describe cluster, systematic and multistage samples and tell them apart from stratified samples.
- Choose a sensible sampling method for a real setting such as a farm, a school or a church directory.

VIDEO

WATCH

Sampling (3 of 5: Random Sampling with Calculators and Grid of Numbers)

Eddie Woo · 7:48

When to watch: Before Example 1: choosing an SRS with random digits.

https://www.youtube.com/watch?v=F6rnH1fztDo

BACKUP

Sampling (4 of 5: Introductory Examples of Stratified Random Sampling)

Eddie Woo · 9:00

When to watch: Before Example 2: stratified random sampling.

https://www.youtube.com/watch?v=PYLfQr9_hBk

KEY TERMS

simple random sample (SRS)	A sample of n individuals chosen so that every possible group of n individuals in the population has the same chance to be the sample.
table of random digits	A long list of the digits 0 to 9 produced by a chance process: every position is equally likely to hold any digit, and knowing some digits tells you nothing about the others.
stratified random sample	A sample chosen by first dividing the population into groups of similar individuals, called strata, and then choosing a separate SRS from each stratum.

cluster sample	A sample chosen by dividing the population into groups (clusters) that are each a mix of individuals, choosing an SRS of clusters, and including every individual in the chosen clusters.
systematic random sample	A sample chosen from a list by picking a random starting point among the first k individuals and then taking every kth individual after it.
multistage sample	A sample chosen in stages, for example an SRS of counties, then an SRS of farms in each chosen county, then an SRS of animals on each chosen farm.

CORE IDEAS

The basic random sample is the **simple random sample (SRS)**, the careful version of drawing names from a hat. Every individual has the same chance to be chosen, and, more than that, every possible *group* of n individuals has the same chance to be the sample. Because no one's judgment enters, an SRS cannot favor any kind of individual, and that is exactly why it avoids the bias of convenience and voluntary response samples.

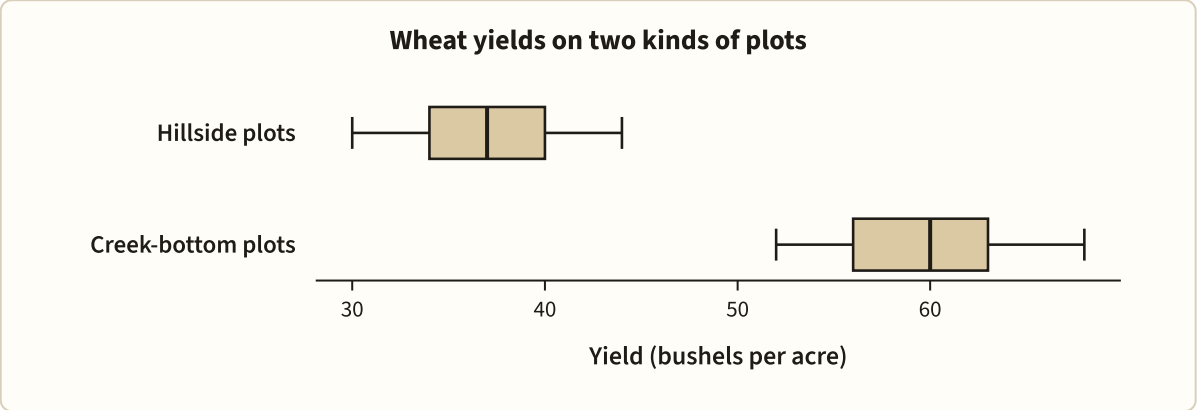
To choose an SRS with a table of random digits, follow three steps. **Label:** give each individual a numerical label with the same number of digits (01 to 30 for 30 families, 001 to 250 for 250 boards). **Randomize:** read the table in groups of that many digits, moving along the line. **Select:** keep each label that belongs to an individual, skip groups that are not labels (such as 47 when the labels stop at 30), skip repeats, and stop when you have n individuals. Equal-length labels matter: every two-digit group from 00 to 99 is equally likely to appear, so every label gets the same chance.

Table of random digits

Line	Digits 1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40
101	48271	90365	12804	77539	26018	54392	81746	03957
102	63518	20947	85132	46079	31685	97240	15863	72094
103	07624	58391	42716	09853	64170	28935	50481	36627
104	91045	36782	70259	13846	82593	41607	69318	25740
105	25930	71468	03587	92614	47082	16359	88720	54193
106	84713	05296	61840	37925	19564	72038	40675	98312

A random number generator does the same job faster: on a calculator or a website, ask for random whole numbers from 1 to 30, ignore repeats, and continue until you have n different numbers. Either way chance does the choosing, and you should record exactly what you did so someone else could check your sample.

Sometimes we know ahead of time that the population is made of different kinds of individuals: a farm has stony hillside plots and rich creek-bottom plots, a school has four grades, a parish has children, youth and adults. A **stratified random sample** divides the population into these groups, called **strata** (one is a *stratum*), and takes a separate SRS from each. Usually each stratum's share of the sample matches its share of the population. The boxplots show wheat yields on the two kinds of plots on one farm.



Yields *within* each stratum are close together, but the two strata are far apart. Suppose half of the farm's plots are on the hillside. An SRS of 10 plots from the whole farm might by bad luck contain 7 hillside plots, and its estimate of the farm's mean yield would come out much too low. Sampling each stratum separately guarantees each kind of plot its proper share, so the luck of the draw matters much less. That is why stratifying by groups of *similar* individuals gives more precise estimates (estimates that vary less from one sample to the next) than an SRS of the same size.

Three other methods save time and travel. A **cluster sample** chooses whole groups at random (a few egg cartons from a shipment, a few rows of an orchard) and measures every individual in them; it works well when each cluster is a mix that looks like the whole population. Notice the contrast: strata are alike *inside* and different from each other, while clusters are mixed inside and alike from one to the next. A **systematic random sample** picks a random start among the first k names on a list and then takes every kth name; it is quick, but beware a list with a repeating pattern. A **multistage sample** uses chance at several stages, as large farm surveys do: counties, then farms, then animals.

Random sampling methods

Method	How chance is used	Works best when
Simple random sample	Choose n individuals at random from the whole list	A full list exists and individuals are easy to reach
Stratified	Choose a separate SRS from each stratum	The population has groups that differ from one another
Cluster	Choose whole groups at random; measure everyone in them	Individuals come in natural groups that are each a mix
Systematic	Random start among the first k, then every kth	A list exists with no repeating pattern
Multistage	Random choices at each of several stages	The population is large and spread out

Choosing a method is a practical judgment. Use an SRS when you have a complete list and the individuals are easy to reach; stratify when you know of groups that differ in the thing you are measuring; use clusters when individuals come in natural groups and reaching them one by one is costly. Whatever method you choose, chance must make the final choice of individuals.

KEY FORMULAS

$$\text{stratum's share of the sample} = n \times \frac{\text{stratum size}}{\text{population size}}$$

$$\text{systematic sampling: } k = \frac{\text{population size}}{\text{sample size}},\ \text{random start from 1 to } k,\ \text{then every } k\text{th individual}$$

From history. Before computers, statisticians needed printed tables of random digits. L. H. C. Tippett published one of the first, with 41,600 digits, in 1927. In 1955 the RAND Corporation published an entire book, A Million Random Digits with 100,000 Normal Deviates, produced with an electronic machine; for years it was a standard tool for choosing samples and assigning treatments.

Practice habit. When you choose a sample with random digits, write out the three steps (Label, Randomize, Select) and say exactly which line of the table you started on, so anyone can check your sample.

WORKED EXAMPLES

Example 1. Pastor Miller wants to ask an SRS of 4 of the 30 families in the church directory whether they would prefer a 5:30 or a 6:30 start for the Wednesday supper. The families are listed alphabetically. Use line 101 of the table of random digits to choose the sample.

Table of random digits

Line	Digits 1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40
101	48271	90365	12804	77539	26018	54392	81746	03957
102	63518	20947	85132	46079	31685	97240	15863	72094
103	07624	58391	42716	09853	64170	28935	50481	36627
104	91045	36782	70259	13846	82593	41607	69318	25740
105	25930	71468	03587	92614	47082	16359	88720	54193
106	84713	05296	61840	37925	19564	72038	40675	98312

1. Label: number the families 01 to 30 in alphabetical order, so every label has two digits.
2. Randomize: read line 101 in pairs: 48, 27, 19, 03, 65, 12, ...
3. Select: 48 is not a label, so skip it. Keep 27. Keep 19. Keep 03. Skip 65. Keep 12.
4. We now have 4 different labels, so we stop: families 27, 19, 03 and 12.
5. Conclusion: the pastor asks families 27, 19, 03 and 12. Because random digits did the choosing, every group of 4 families had the same chance to be the sample, and nobody's preferences decided who was asked.

Answer: Families 27, 19, 03 and 12.

Example 2. An orchard has 300 apple trees: 180 Honeycrisp and 120 Gala. The two varieties produce very different amounts of fruit. The owner wants a stratified random sample of 20 trees, with each variety's share of the sample equal to its share of the orchard. (a) How many trees of each variety should be sampled? (b) The Gala trees are labeled 001 to 120. Use line 103 of the table of random digits to choose the first 3 Gala trees. (c) Why is stratifying by variety a good idea here?

Table of random digits

Line	Digits 1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40
101	48271	90365	12804	77539	26018	54392	81746	03957
102	63518	20947	85132	46079	31685	97240	15863	72094
103	07624	58391	42716	09853	64170	28935	50481	36627
104	91045	36782	70259	13846	82593	41607	69318	25740
105	25930	71468	03587	92614	47082	16359	88720	54193
106	84713	05296	61840	37925	19564	72038	40675	98312

1. (a) Honeycrisp: $20 \times \dfrac{180}{300} = 20 \times 0.6 = 12$ trees.
2. Gala: $20 \times \dfrac{120}{300} = 20 \times 0.4 = 8$ trees. Check: $12 + 8 = 20$.
3. (b) Label: the Gala trees already have three-digit labels 001 to 120.
4. Randomize: read line 103 in groups of three: 076, 245, 839, 142, 716, 098, 536, 417, 028, ...
5. Select: keep 076; skip 245, 839, 142 and 716 (all greater than 120); keep 098; skip 536 and 417; keep 028.
6. The first three Gala trees are 076, 098 and 028. (Continue the same way until 8 Gala trees are chosen, and choose the 12 Honeycrisp trees with a separate SRS.)
7. (c) Trees of the same variety yield about the same, but the two varieties differ a lot. Sampling each variety separately guarantees the right mix of 12 and 8, so a sample cannot by chance be mostly one variety, and the estimate of the orchard's mean yield varies less from sample to sample.

Answer: (a) 12 Honeycrisp and 8 Gala. (b) Gala trees 076, 098 and 028. (c) Trees within a variety are similar while the varieties differ, so stratifying guarantees each variety its proper share and gives a more precise estimate than an SRS of 20.

PRACTICE

1. EASY Name the sampling method in each case. (a) A beekeeper numbers her 60 hives, rolls a die to pick a starting hive from 1 to 6, and then checks every 6th hive. (b) A building inspector chooses 3 of 40 pallets of bricks at random and checks every brick on those pallets. (c) A school chooses 10 students at random from each of its four grades. (d) A choir director uses a random number generator to choose 15 of the 120 choir members.
2. EASY A sawmill wants to label its 250 boards so it can choose an SRS with a table of random digits. How many digits should each label have?
3. EASY Ruth has 40 hens, labeled 01 to 40. Use line 102 of the table of random digits to choose an SRS of 3 hens to weigh.

Table of random digits

Line	Digits 1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40
101	48271	90365	12804	77539	26018	54392	81746	03957
102	63518	20947	85132	46079	31685	97240	15863	72094
103	07624	58391	42716	09853	64170	28935	50481	36627
104	91045	36782	70259	13846	82593	41607	69318	25740
105	25930	71468	03587	92614	47082	16359	88720	54193
106	84713	05296	61840	37925	19564	72038	40675	98312

4. EASY A 4-H club has a list of 200 members and wants a systematic random sample of 20. (a) What is k, the spacing between chosen members? (b) If the random start is member 7, what is the number of the last member chosen?

5. MEDIUM A tree farm labels its 95 Christmas trees 01 to 95. Use line 104 of the table of random digits to choose an SRS of 8 trees to measure.

Table of random digits

Line	Digits 1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40
101	48271	90365	12804	77539	26018	54392	81746	03957
102	63518	20947	85132	46079	31685	97240	15863	72094
103	07624	58391	42716	09853	64170	28935	50481	36627
104	91045	36782	70259	13846	82593	41607	69318	25740
105	25930	71468	03587	92614	47082	16359	88720	54193
106	84713	05296	61840	37925	19564	72038	40675	98312

6. MEDIUM A church has 360 members: 90 children, 72 youth and 198 adults. The potluck committee wants a stratified random sample of 40 members to ask about the menu, with each group's share of the sample equal to its share of the membership. How many children, youth and adults should be sampled?

7. MEDIUM A dairy farmer has 50 Holstein cows and 30 Jersey cows. Holsteins give much more milk per day than Jerseys. Explain why a sample stratified by breed would give a more precise estimate of the herd's mean daily milk yield than an SRS of the same size.

8. MEDIUM An egg farm ships 500 cartons of 12 eggs. An inspector chooses 5 cartons at random and checks every egg in them for cracks. (a) How many eggs are in the sample? (b) Why might this cluster sample be easier than an SRS of the same number of eggs?

9. MEDIUM A teacher's classroom has 6 rows of 5 desks. She rolls a die to choose one row and asks all 5 students in that row about a field trip. Is this a simple random sample of 5 students? Answer yes or no, and explain.

10. HARD A sawmill has 300 boards from one day's cutting, labeled 001 to 300. Use line 105 of the table of random digits to choose an SRS of 3 boards to test for moisture.

Table of random digits

Line	Digits 1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40
101	48271	90365	12804	77539	26018	54392	81746	03957
102	63518	20947	85132	46079	31685	97240	15863	72094
103	07624	58391	42716	09853	64170	28935	50481	36627
104	91045	36782	70259	13846	82593	41607	69318	25740
105	25930	71468	03587	92614	47082	16359	88720	54193
106	84713	05296	61840	37925	19564	72038	40675	98312

11. HARD An orchard has 40 rows of 10 trees. The trees are listed row by row, so trees 10, 20, 30, ... are all at the east end of a row, next to the windbreak. Explain what could go wrong with a systematic sample that takes every 10th tree, and suggest a better method.

12. HARD A state farm survey randomly selects 5 of its 40 counties, then randomly selects 4 dairy farms in each chosen county, then randomly selects 10 cows on each chosen farm. (a) How many cows are in the sample? (b) What kind of sample is this, and why would a surveyor use it instead of an SRS of all the cows in the state?

13. CHALLENGE A flock has 50 Leghorn hens that each lay about 6 eggs a week and 50 Bantam hens that each lay about 3 eggs a week. (a) What is the mean number of eggs per hen per week for the whole flock? (b) An SRS of 10 hens happens to contain 8 Leghorns and 2 Bantams. If each hen lays exactly her breed's typical number, what is the sample mean? (c) Explain how a stratified sample would prevent this kind of error.

COMMON MISTAKES

- Using labels of different lengths (1, 2, ..., 30 instead of 01, 02, ..., 30). With unequal labels you cannot read the table in equal groups, and some labels get more chances than others. Give every label the same number of digits.
- Forgetting to skip repeats and groups that are not labels (such as 00 or 31 when the labels are 01 to 30). Cross off each label as it is chosen and keep reading until you have n different individuals.
- Mixing up strata and clusters. Strata are groups of similar individuals, and you sample some from every stratum; clusters are mixed groups, and you take all of a few clusters.
- Using a systematic sample on a list with a repeating pattern, which can make every chosen individual the same kind. Check the list for patterns first.

ANSWERS TO PRACTICE

1. (a) Systematic random sample. (b) Cluster sample (the pallets are the clusters). (c) Stratified random sample (the grades are the strata). (d) Simple random sample.
2. 3 digits: the labels run from 001 to 250.
3. Pairs: 63 (skip), 51 (skip), 82 (skip), 09 (keep), 47 (skip), 85 (skip), 13 (keep), 24 (keep). The sample is hens 09, 13 and 24.
4. (a) $k = \dfrac{200}{20} = 10$. (b) The sample is 7, 17, 27, ..., so the 20th member chosen is $7 + 19 \times 10 = 197$.
5. Pairs: 91, 04, 53, 67, 82, 70, 25 (all keep), 91 (repeat, skip), 38 (keep). The sample is trees 91, 04, 53, 67, 82, 70, 25 and 38.
6. Children: $40 \times \dfrac{90}{360} = 10$. Youth: $40 \times \dfrac{72}{360} = 8$. Adults: $40 \times \dfrac{198}{360} = 22$. Check: $10 + 8 + 22 = 40$.
7. Cows of the same breed give similar amounts of milk, but the breeds differ a lot. An SRS might by chance include too many Holsteins (estimate too high) or too many Jerseys (estimate too low). Stratifying guarantees each breed its proper share, so that source of chance error is removed and estimates vary less from sample to sample.
8. (a) $5 \times 12 = 60$ eggs. (b) For an SRS of 60 eggs the inspector would have to number all 6,000 eggs and open up to 60 different cartons; with clusters she opens only 5 cartons and checks every egg in them.
9. No. In an SRS every group of 5 students must have the same chance to be the sample, but here only 6 groups (the six rows) are possible; for example, two students in different rows can never both be chosen. This is a cluster sample.
10. Groups of three: 259 (keep), 307 (skip), 146 (keep), 803 (skip), 587 (skip), 926 (skip), 144 (keep). The sample is boards 259, 146 and 144.
11. Because the list repeats every 10 trees, a systematic sample with k = 10 takes trees in the same position in every row. If the random start is 10, every chosen tree stands by the windbreak, where conditions differ from the rest of the orchard, and the sample would be biased. A better method is an SRS of all 400 trees, or a stratified sample that chooses trees at random from each row.
12. (a) $5 \times 4 \times 10 = 200$ cows. (b) A multistage sample. There is no single list of every cow in the state, and visiting cows scattered over 40 counties would be very costly; choosing counties and then farms keeps the work to 20 farm visits while chance still makes every choice.

13. (a) $\dfrac{50 \times 6 + 50 \times 3}{100} = \dfrac{450}{100} = 4.5$ eggs. (b) $\dfrac{8 \times 6 + 2 \times 3}{10} = \dfrac{54}{10} = 5.4$ eggs, far too high. (c) A stratified sample would take an SRS of 5 Leghorns and an SRS of 5 Bantams, matching each breed's half of the flock, so the sample can never be lopsided by breed; the only chance error left is the small variation among hens of the same breed.

LESSON 5.3 · 2 SESSIONS OF 45 MINUTES

Problems with Surveys

OBJECTIVES

- Distinguish sampling error (chance variation from sample to sample) from nonsampling errors such as undercoverage, nonresponse, response bias and poor wording.
- Identify the likely source and direction of error in a survey described in context, and compute a response rate.
- Write survey questions that are clear, neutral and ask about only one thing.
- Explain why a larger random sample reduces chance variation but does not remove bias.

VIDEO

WATCH

Cautions about Sample Surveys, Causes of Bias, and Inference defined

ProfRobBob · 14:57

When to watch: Before Example 1: undercoverage, nonresponse, voluntary response and response bias.

https://www.youtube.com/watch?v=xM1dzLI0uKM

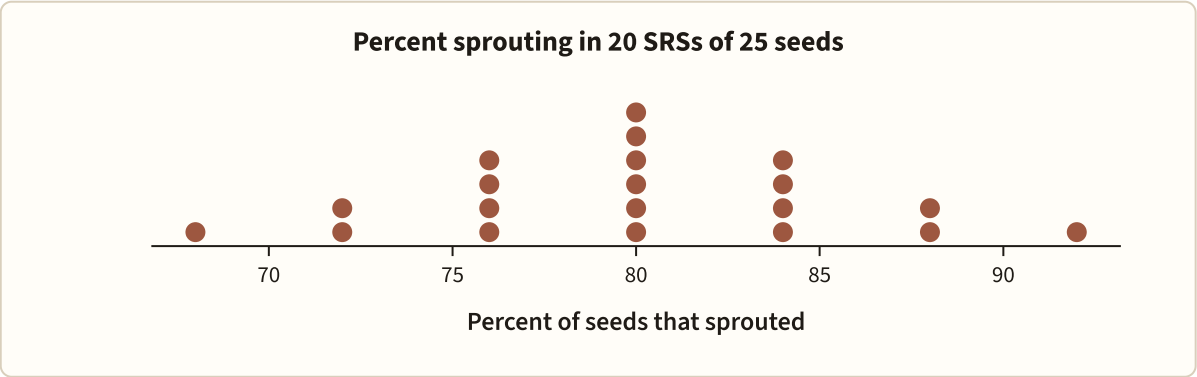
KEY TERMS

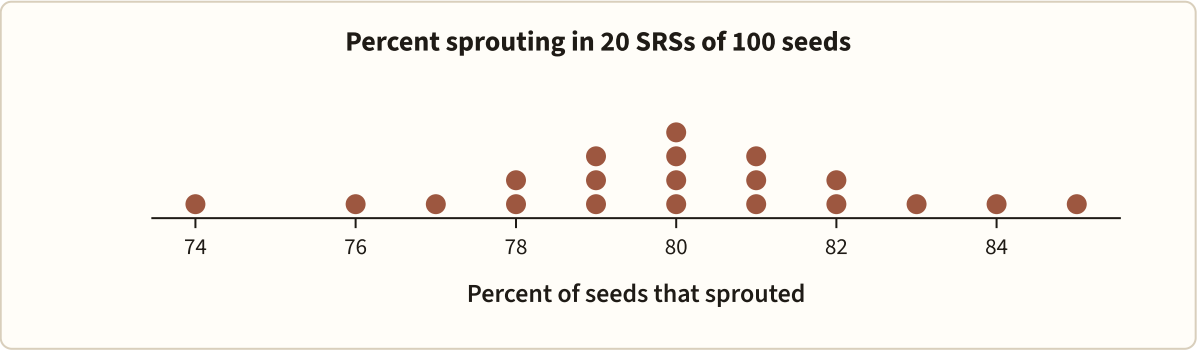
sampling error	The chance difference between a sample result and the true population value that comes from examining only part of the population. It occurs even in a perfectly chosen random sample, and it shrinks as the sample gets larger.
nonsampling error	An error that comes from how a survey is carried out (who is left out, who answers, how questions are asked) rather than from chance. A larger sample does not reduce it.
undercoverage	When some groups in the population have no chance to be in the sample, often because the list used to choose the sample leaves them out.
nonresponse	When individuals chosen for the sample cannot be contacted or refuse to answer. The response rate is the percent of the chosen sample that actually responds.
response bias	A systematic pattern of inaccurate answers, caused for example by people wanting to look good, misremembering, or leaning toward the answer they think the interviewer wants.

wording bias Bias caused by the way a question is asked: leading or loaded words, confusing phrasing, or asking about two things at once.

CORE IDEAS

Suppose a seed company takes a perfectly chosen SRS of 25 seeds from a large seed lot and plants them. Even if exactly 80% of all the seeds in the lot will sprout, this sample might give 76%, or 84%, or even 72%, and a second SRS would give a different answer. This chance variation from sample to sample is called **sampling error**. It is not a mistake anyone made; it is the price of looking at only part of the population. The dotplots show the sprouting percents in 20 different SRSs of 25 seeds and in 20 different SRSs of 100 seeds from the same lot.



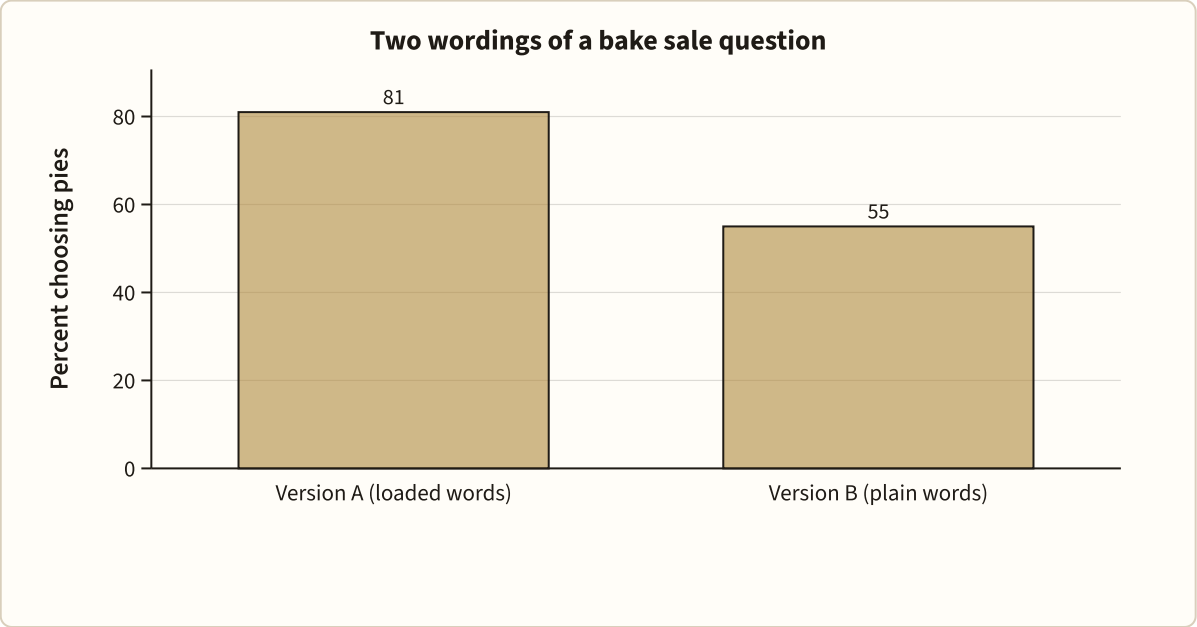


Both sets of results center near the true 80%: random sampling has no favorites. But the samples of 100 vary much less (from 74% to 85%) than the samples of 25 (from 68% to 92%). In a larger sample, the unusually good and unusually poor seeds have more chances to balance each other out, so larger random samples give results that are more consistent and usually closer to the truth.

Other errors have nothing to do with chance; they come from how the survey is carried out. These **nonsampling errors** do not shrink as the sample grows. **Undercoverage** happens when part of the population has no chance to be chosen: a survey about library hours mailed only to card holders misses town residents who have stopped using the library, perhaps because the hours don't suit them. **Nonresponse** happens when chosen people cannot be reached or decline to answer; if they differ from those who do answer (busy farm families during harvest are harder to reach), the results are biased. The **response rate** tells what percent of the chosen sample actually answered.

Response bias comes from the answers themselves. People tend to give the answer that sounds good ('Yes, I read to my children every night'), they forget ('How many days did it rain last month?'), and they lean toward the answer they think the interviewer wants. Asking in private, asking about specific recent events, and using a neutral interviewer all help.

Wording may be the most powerful influence of all. A question that leads the respondent ('Wouldn't you agree ...?'), uses loaded words ('fresh, homemade' against 'store-bought'), is confusing, or asks two things at once ('Do you like the choir's new music and the new practice time?') produces answers that measure the wording as much as the opinion. The bar graph shows the percent choosing pies when two randomly chosen halves of a congregation were asked two versions of one question. Version A: 'Should the bake sale offer store-bought cookies or fresh, homemade pies?' Version B: 'Should the bake sale offer cookies or pies?'



The two halves were alike, and the choice was the same, yet the results differ by $81 - 55 = 26$ percentage points, caused only by the words. A fair question is short and clear, uses neutral words, asks about one thing, and describes every choice in the same plain way. Before using a survey, try it on a few people and ask them what they thought each question meant.

Here is the key lesson of this unit so far: **a larger sample reduces sampling error, but it does nothing about bias.** A survey of 5,000 people who answered a website poll, or who were asked a leading question, simply gives a very steady wrong answer. Fix the method first; then choose the size.

KEY FORMULAS

$$\text{response rate} = \frac{\text{number who respond}}{\text{number chosen for the sample}} \times 100\%$$

Practice habit. Read every survey question out loud and ask: would a person on either side of this question think it is fair? If the wording leans, rewrite it before you collect a single answer.

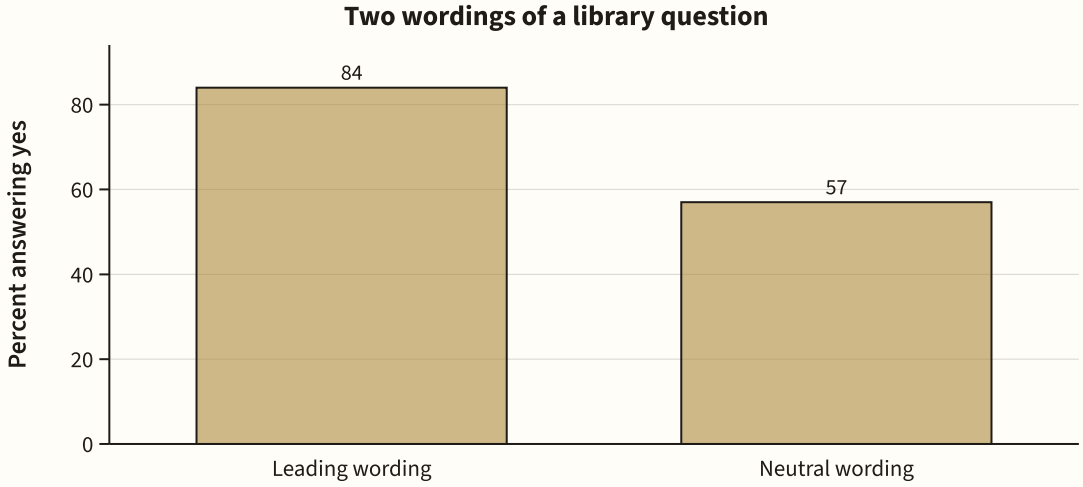
WORKED EXAMPLES

Example 1. The county fair board mailed a survey to an SRS of 400 households in the county asking whether the fair should move from July to August. Only 140 households returned it, and 98 of those favored moving the fair. (a) What was the response rate? (b) What percent of those who responded favored moving the fair? (c) Explain how nonresponse could bias this result.

1. (a) Response rate $= \dfrac{140}{400} \times 100\% = 35\%$.
2. (b) Percent in favor among respondents $= \dfrac{98}{140} = 0.70$, which is 70%.
3. (c) 260 of the 400 chosen households, almost two-thirds, did not answer. The 70% describes only the households that chose to respond.
4. Households that want the change have a reason to send the survey back; households that are happy with July, or that rarely attend the fair, may not bother.
5. If the nonresponders differ from the responders in this way, 70% overstates support for moving the fair among all households. The random sample was a good start, but nonresponse can bring bias back in.

Answer: (a) 35%. (b) 70%. (c) Nearly two-thirds of the chosen households did not answer, and those who care about a change are more likely to respond, so the 70% may overstate support among all households in the county.

Example 2. The parish library committee surveyed two random halves of the congregation. Half were asked, 'Don't you agree that the library should stay open on Tuesday evenings so that busy families have time to borrow books?' The other half were asked, 'Should the library be open on Tuesday evenings?' The bar graph shows the results. (a) What kind of problem does the first question have? (b) By how many percentage points did the wording change the result? (c) Write a fair version of the question.



1. (a) 'Don't you agree ...' invites a yes, and the reason 'so that busy families have time to borrow books' argues for one side. This is a leading question, a source of wording bias.
2. (b) The leading version got 84% yes and the neutral version 57% yes.
3. Difference $= 84 - 57 = 27$ percentage points.
4. Since the two halves were chosen at random, the groups were alike, so the 27-point gap is due to the wording, not to the people.
5. (c) A fair question names the choices plainly and gives no reasons for either side: 'Should the library add hours on Tuesday evenings from 6 to 8 p.m., or keep its current hours?'
6. Conclusion: the committee should report the neutral result (57%), not the inflated 84%.

Answer: (a) It is a leading question (wording bias): it asks for agreement and gives a reason for one side. (b) $84 - 57 = 27$ percentage points. (c) For example: 'Should the library add hours on Tuesday evenings from 6 to 8 p.m., or keep its current hours?'

PRACTICE

1. EASY A town library chooses people to call about new hours from an old printed phone book of landline numbers. Many younger families have only cell phones. Name the type of error and explain how it might affect the results.

2. **EASY** The county extension office mailed a survey about pasture management to 600 randomly chosen farm families, and 186 were returned. What was the response rate, as a percent?

3. **EASY** A basketball coach asks each player, face to face, 'How many hours did you practice free throws on your own this week?' What kind of error is likely, and in which direction?

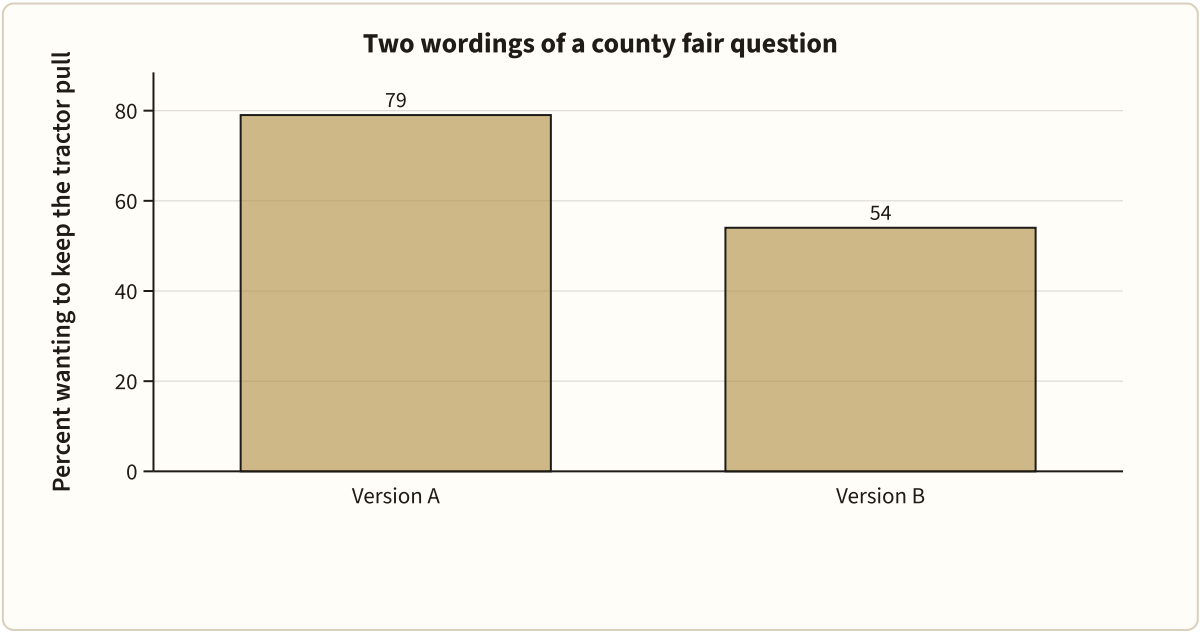
4. **EASY** Two different SRSs of 50 apples from the same orchard find 42% and 48% of the apples blemished. Is the difference between the two results a sign that one of the samples is biased? Answer yes or no, and explain.

5. **MEDIUM** A survey question reads: 'Wouldn't a new swing set be a wonderful gift for the children of our parish?' Explain what is wrong with the question and rewrite it fairly.

6. **MEDIUM** The choir director asks, 'Do you like the choir's new music and the new Thursday practice time?' Explain the problem with this question and fix it.

7. **MEDIUM** A weather service chose 250 farms at random to ask about frost damage. 40 farms could not be reached and 60 refused to answer. (a) How many farms responded? (b) What was the response rate, as a percent?

8. **MEDIUM** The county fair board asked two random halves of fairgoers about the tractor pull. Version A: 'Should the fair keep its popular, long-standing tractor pull?' Version B: 'Should the fair keep the tractor pull or replace it with a horse show?' The bar graph gives the percent who wanted to keep the tractor pull. (a) By how many percentage points did the wording change the result? (b) Which version is fairer, and why?



9. **MEDIUM** The fair board's website poll about parking got 300 responses. A board member wants to keep the poll open until it has 3,000 responses 'so the results will be reliable.' Will the larger number of responses remove the bias? Answer yes or no, and explain.

10. HARD To learn how many hours a week teenagers in the county spend volunteering, a club leader hands a questionnaire to the teens at a Saturday service project and waits beside each one while they fill it in. Name two different sources of bias and the direction of each.

11. HARD An orchard growers' association mailed the question 'Would you attend a pruning workshop in March?' to an SRS of 400 member families. Only 240 families answered, and 144 of them said yes. (a) What percent of those who answered said yes? (b) If all 160 families who did not answer would have said no, what percent of the full sample of 400 would have said yes? (c) If all 160 would have said yes, what percent would that be?

12. HARD A farmer wants the mean weight of the potatoes in a large bin. When a bin is jostled, large potatoes work their way to the top. Method A takes an SRS of 20 potatoes from the whole bin; Method B scoops 200 potatoes from the top. Each method was tried five times, with the results in the table. The true mean weight of the potatoes in the bin is 170 grams. (a) Find the mean of Method A's five results. (b) Find the mean of Method B's five results. (c) Which method is biased, which varies less, and which should the farmer use?

Mean potato weight from two methods, five tries each

Try	Method A: SRS of 20 (grams)	Method B: 200 from the top (grams)
1	164	192
2	176	194
3	168	191
4	173	193
5	169	190

13. CHALLENGE The church library committee wants to learn which evening of the week would be best for extra library hours. Describe how to choose the sample, write one fair question, and describe two things the committee could do to reduce nonresponse.

COMMON MISTAKES

- Calling sampling error a 'mistake'. Chance variation between random samples is expected and can be measured; bias is the real danger. Ask whether the method favors certain answers.
- Believing a large sample guarantees accuracy. A large sample reduces chance variation only; undercoverage, nonresponse, response bias and poor wording remain no matter how many people answer.
- Ignoring nonresponse: reporting the percent of respondents as if it described everyone chosen. Always report the response rate and ask how the nonresponders might differ.
- Writing questions that lean or that ask two things at once. Use neutral words, one idea per question, and list every choice in the same plain way.

ANSWERS TO PRACTICE

1. Undercoverage: families with only cell phones have no chance to be chosen. Younger families may well want different hours (for example, evenings or Saturdays), so their views would be underrepresented.
2. $\frac{186}{600} \times 100\% = 31\%$.
3. Response bias: players want to look dedicated in front of their coach, so the reported hours are likely too high.
4. No. Different random samples give different results just by chance; this is sampling error, not bias. A difference of a few percentage points between two samples of 50 is expected.
5. It is a leading question: 'Wouldn't ...' invites agreement, and 'wonderful gift for the children' makes it hard to say no. A fair version: 'Should the parish spend money from this year's budget on a new swing set? (Yes / No / No opinion)'.
6. It asks about two things at once. A member who likes the music but not the time cannot answer honestly with one yes or no. Ask two separate questions: 'Do you like the choir's new music?' and 'Does the Thursday practice time work for you?'
7. (a) $250 - 40 - 60 = 150$ farms. (b) $\frac{150}{250} \times 100\% = 60\%$.
8. (a) $79 - 54 = 25$ percentage points. (b) Version B: it names both choices plainly. Version A uses the loaded words 'popular, long-standing' and does not mention any alternative.
9. No. It is still a voluntary response sample: people who are upset about parking are the most likely to answer. More responses of the same kind make the result steadier, but they do not make the respondents represent all fairgoers.
10. First, a convenience sample: teens who are at a service project are more likely than other teens to volunteer, so the estimate will be too high. Second, response bias: with the leader standing beside them, teens may report more volunteer hours than they really give, again making the estimate too high.
11. (a) $\frac{144}{240} = 0.60$, so 60%. (b) $\frac{144}{400} = 0.36$, so 36%. (c) $\frac{144 + 160}{400} = \frac{304}{400} = 0.76$, so 76%. Because of nonresponse, the true sample percent could be anywhere from 36% to 76%, which shows how much nonresponse can matter.
12. (a) $\frac{164 + 176 + 168 + 173 + 169}{5} = \frac{850}{5} = 170$ grams. (b) $\frac{192 + 194 + 191 + 193 + 190}{5} = \frac{960}{5} = 192$ grams. (c) Method B is biased: its results cluster near 192 grams, far above the truth, because it takes only the big potatoes on top. Method B varies less (its results are within 4 grams of each other) because its samples are large, but that only makes it consistently wrong. Method A is centered on the truth, so the farmer should use Method A, ideally with a larger SRS to reduce its variation.

13. Sample: number every household in the church directory and choose an SRS (for example 40 households) with random digits or a random number generator, so no one is chosen for convenience and no one volunteers. Question: 'If the library added one evening of hours from 6 to 8 p.m., which evening would you be most likely to use: Monday, Tuesday, Wednesday, Thursday, Friday, or none?' (neutral, one topic, every choice listed plainly). To reduce nonresponse: keep the survey short, and follow up with a phone call or a second note to every chosen household that has not answered, rather than replacing them with easier-to-reach households.

LESSON 5.4 · 2 SESSIONS OF 45 MINUTES

Observational Studies and Experiments

OBJECTIVES

- Distinguish an observational study from an experiment by asking who decided which individuals received which treatment.
- Identify the explanatory and response variables, factors and levels, treatments, and experimental units or subjects in an experiment.
- Explain confounding and name a plausible confounding variable in context.
- Explain why a well-designed experiment can show cause and effect while an observational study usually cannot.

VIDEO

WATCH

Causation Defined & 5 Key Checks for Signs of Causattion

ProfRobBob · 10:16

When to watch: Before Example 2: lurking variables, confounding, and why only a well-designed experiment shows cause and effect.

https://www.youtube.com/watch?v=b8TR3FXWN8A

KEY TERMS

observational study	A study that observes individuals and measures variables without trying to influence the responses. The researchers do not decide who gets which treatment.
experiment	A study that deliberately imposes treatments on individuals in order to observe their responses.
factor	An explanatory variable in an experiment. The particular values of a factor that are tried are called its levels (for example, 0, 40 and 80 pounds of fertilizer per acre).
treatment	A specific condition applied to the individuals in an experiment. With two or more factors, a treatment is a combination of one level of each factor.
experimental unit	The object to which a treatment is applied, such as a plot, a plant, a calf or a loaf of bread. When the units are people, they are called subjects.

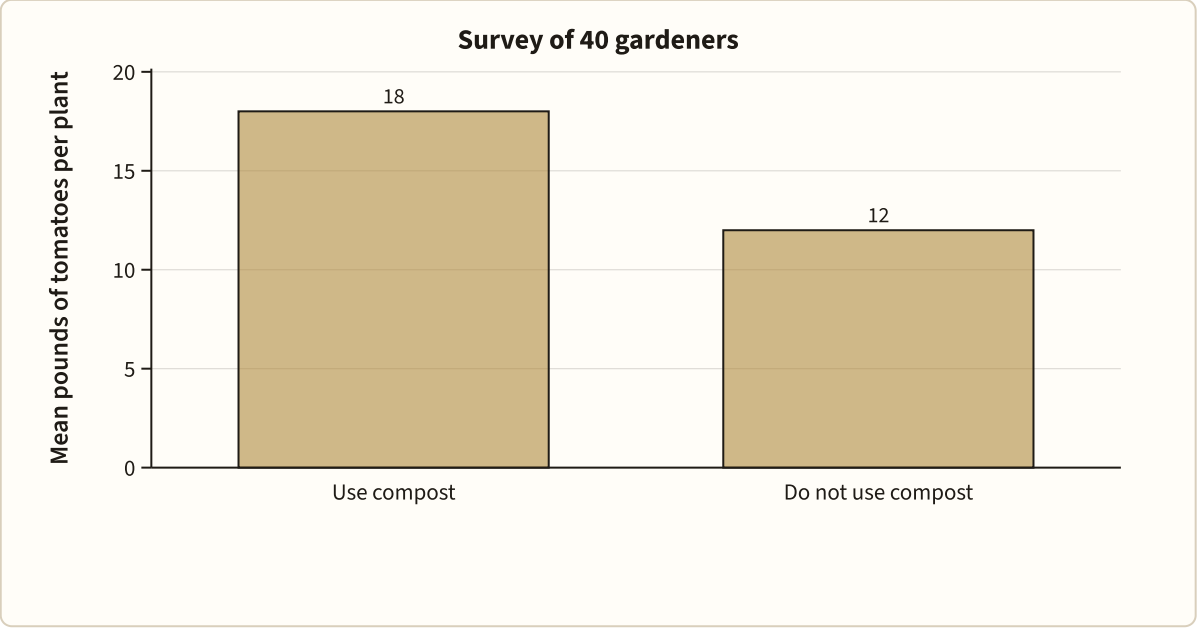
confounding Two variables are confounded when their effects on the response cannot be told apart, so we cannot say which one caused a difference.

CORE IDEAS

In Unit 4 you learned that association is not causation. This lesson introduces the tool that *can* show causation. In an **observational study** we observe individuals and record variables without trying to change anything: we find out which gardeners use compost and how many tomatoes they pick. In an **experiment** we deliberately do something to the individuals and watch the response: *we* decide which garden plots get compost.

Experiments have their own vocabulary. The response variable is what we measure (pounds of tomatoes per plant). An explanatory variable that we control is called a **factor**, and the values we try are its **levels**. A **treatment** is what a unit actually receives: with one factor, the treatments are its levels; with two factors, each treatment is a combination. Testing 3 tomato varieties with and without straw mulch gives $3 \times 2 = 6$ treatments (Roma with mulch, Roma without, and so on). The things that receive the treatments are the **experimental units**; when they are people, we call them **subjects**.

Why not simply observe? Suppose a survey of 40 gardeners finds that those who use compost harvest more tomatoes per plant.



Does the compost cause the bigger harvest? Not necessarily. Gardeners who take the trouble to make compost may also be more experienced, water more faithfully and plant in sunnier spots. Experience, watering and sunlight are **confounded** with compost use: their effects on the harvest are mixed together with the effect of compost, and nothing in the data can pull them apart. A confounding variable is a lurking variable (Lesson 4.5) that travels together with the explanatory variable. An observational study can measure the association, but it cannot say how much of the 6-pound difference, if any, is due to compost.

An experiment cuts this knot. If *we* decide which plots get compost, and we decide by chance (Lesson 5.5), then well-tended and poorly tended plots, sunny and shady spots, and rich and poor soil are spread across both groups about evenly. The groups start out alike in every way except the treatment. So if the compost plots then produce clearly more, the compost is the only good explanation left, apart from chance, and the size of chance differences can be measured. That is why a well-designed experiment is the best way to show cause and effect.

Observational studies are still valuable. Some questions cannot be studied by experiment: we cannot assign rainfall to farms or a birth month to a calf, and some treatments would be wrong to impose on people. Observational studies describe groups, uncover associations worth testing, and answer questions where experiments are impossible. But their conclusions about cause must be cautious: the more confounding variables a study can rule out, the stronger its case, yet it can never rule out every one.

To tell the two kinds of study apart, ask one question: **Did the researchers decide which individuals received which treatment?** If yes, it is an experiment. If they only recorded what individuals already did or already were, it is an observational study, no matter how large the sample or how carefully the data were measured.

KEY FORMULAS

$$\text{number of treatments} = (\text{levels of factor 1}) \times (\text{levels of factor 2}) \times \cdots$$

Practice habit. For every study, ask first: did the researchers assign the treatments? If not, it is observational, and name at least one possible confounding variable before you say anything about cause.

WORKED EXAMPLES

Example 1. An agriculture student wants to know whether the amount of a fertilizer (0, 40 or 80 pounds per acre) and the planting date (early May or late May) affect corn yield. She has 24 plots and will use each treatment on the same number of plots, measuring the yield of each plot in bushels per acre. (a) Name the factors and their levels. (b) How many treatments are there? (c) How many plots get each treatment? (d) Name the experimental units and the response variable.

The six treatments

Fertilizer	Planted early May	Planted late May
0 lb per acre	Treatment 1	Treatment 2
40 lb per acre	Treatment 3	Treatment 4
80 lb per acre	Treatment 5	Treatment 6

1. (a) Factor 1: amount of fertilizer, with 3 levels (0, 40, 80 pounds per acre). Factor 2: planting date, with 2 levels (early May, late May).
2. (b) Each treatment combines one level of each factor: number of treatments $= 3 \times 2 = 6$, as the table shows.
3. (c) Plots per treatment $= \dfrac{24}{6} = 4$.
4. (d) The experimental units are the 24 plots; the response variable is corn yield in bushels per acre.
5. Conclusion: the student will compare the mean yields of 6 groups of 4 plots, which lets her study both factors, and how they work together, in one experiment.

Answer: (a) Fertilizer amount (0, 40, 80 lb per acre) and planting date (early May, late May). (b) 6 treatments. (c) 4 plots each. (d) Units: the 24 plots; response: corn yield (bushels per acre).

Example 2. A county survey of 40 beef farms recorded whether each farm feeds its calves a new mineral supplement and the calves' mean weight at weaning. The results are in the table. (a) How much heavier, on average, were the calves on supplement farms? (b) Does this study show that the supplement causes heavier calves? Answer yes or no. (c) Explain, using a confounding variable from the table.

County survey of beef farms

Group	Number of farms	Mean weaning weight (lb)	Percent with irrigated pasture
Use supplement	18	560	72%
No supplement	22	520	27%

1. (a) Difference $= 560 - 520 = 40$ pounds.
2. (b) The farmers chose for themselves whether to use the supplement; the county only recorded what they did. This is an observational study, so it cannot by itself show cause.
3. (c) 72% of supplement farms have irrigated pasture, compared with only 27% of the other farms.
4. Irrigated pasture grows more grass, and better grass alone could make calves heavier.
5. Pasture quality is confounded with supplement use: we cannot tell how much of the 40-pound difference comes from the supplement and how much from the pasture. (Farmers who buy a supplement may also manage their herds more carefully in other ways.)
6. Conclusion: the supplement is associated with heavier calves, but to learn whether it causes the gain we need an experiment that assigns the supplement by chance.

Answer: (a) 40 pounds. (b) No. (c) This is an observational study. Supplement farms are much more likely to have irrigated pasture (72% versus 27%), and better pasture could explain the heavier calves; pasture is confounded with the supplement.

PRACTICE

1. EASY Is each study an observational study or an experiment? (a) A weather club records each day's high temperature and the number of people at the town swimming pool. (b) A baker bakes 10 loaves sweetened with honey and 10 sweetened with sugar and compares how high they rise.
2. EASY A student gives bean seedlings 1, 2 or 3 cups of water a day and measures their height after three weeks. Name the explanatory variable and the response variable.

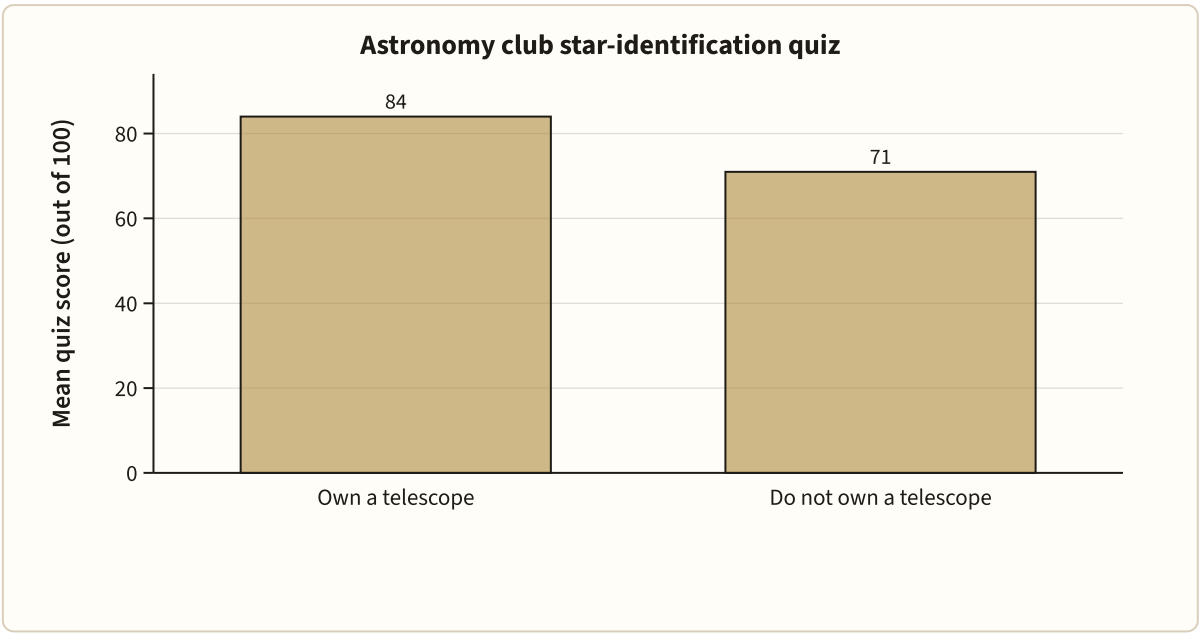
3. EASY An experiment tests 3 amounts of fertilizer (0, 50 and 100 pounds per acre) on corn, and each amount is applied to 5 plots. (a) How many treatments are there? (b) How many plots are needed?

4. EASY A coach randomly assigns 24 basketball players to practice free throws with one of two routines for a month, then records the percent of 50 free throws each player makes. Identify (a) the experimental units, (b) the treatments, (c) the response variable.

5. MEDIUM A bakery tests 4 rising times and 2 oven temperatures for a new bread, baking 3 loaves for each treatment. (a) How many treatments are there? (b) How many loaves are baked?

6. MEDIUM A study finds that students who take music lessons have higher average math grades than students who do not. Name a variable that might be confounded with taking music lessons, and explain how it could produce this result even if music lessons had no effect on math.

7. MEDIUM An astronomy club compared the scores on a star-identification quiz of members who own a telescope and members who do not. (a) How many points higher was the mean score of the telescope owners? (b) Can we conclude that owning a telescope causes higher scores? Answer yes or no. (c) Name a possible confounding variable.



8. MEDIUM The county extension office wants to know whether a new oat variety yields more than the old one. Explain why a randomized experiment on test plots could show that the new variety causes higher yields, but a survey comparing farms that already grow each variety could not.

9. MEDIUM A dairy researcher compares the milk yields of cows on farms that already milk three times a day with cows on farms that milk twice a day. (a) Is this an experiment? Answer yes or no. (b) Name a variable that might be confounded with the milking schedule.

10. HARD A researcher records the summer rainfall and the corn yield on 50 farms. (a) Why must this be an observational study? (b) Describe how an experiment could study the effect of water on corn yield.

11. HARD A farmer gives a new feed to the 15 calves in his north pasture and his usual feed to the 15 calves in his south pasture, then compares weight gains. Explain the flaw in this experiment.
12. HARD A sheep farmer wants to test whether a new hoof-care spray reduces lameness. She has 40 ewes. Describe an experiment, naming the treatments, the experimental units and the response variable.
13. CHALLENGE An experiment on wheat uses 2 varieties, 3 fertilizer levels and 2 watering schedules, with 4 plots for each treatment. (a) How many treatments are there? (b) How many plots are needed?
14. CHALLENGE An observational study found that farms using drip irrigation had higher vegetable yields than farms using sprinklers. Give two different confounding variables, and explain how each could create the difference even if drip irrigation itself had no effect.

COMMON MISTAKES

- Calling a study an experiment because it measured carefully or used a large sample. It is an experiment only if the researchers assigned the treatments.
- Confusing the experimental units with the response. The units are what the treatments are applied to (the plots, the calves); the response is what is measured on them (yield, weight gain).
- Adding levels instead of multiplying to count treatments. With 3 levels of one factor and 2 of another there are $3 \times 2 = 6$ treatments, not 5.
- Concluding cause from an observational study. Say 'is associated with', and name a possible confounding variable.

ANSWERS TO PRACTICE

1. (a) Observational study: the club only records what happens. (b) Experiment: the baker decides which loaves get honey and which get sugar.
2. Explanatory variable (the factor): amount of water per day, with levels 1, 2 and 3 cups. Response variable: seedling height after three weeks.
3. (a) 3 treatments (one factor with 3 levels). (b) $3 \times 5 = 15$ plots.
4. (a) The 24 players (subjects). (b) The two practice routines. (c) The percent of 50 free throws made.
5. (a) $4 \times 2 = 8$ treatments. (b) $8 \times 3 = 24$ loaves.
6. One possibility is how much time and help a student has at home. Families who arrange and pay for music lessons may also help more with homework and expect steady study. Students with that support could earn higher math grades whether or not they took music, so the effect of lessons is mixed up with the effect of home support.

7. (a) $84 - 71 = 13$ points. (b) No: members chose whether to buy telescopes, so this is an observational study. (c) Interest or experience in astronomy: members who are more interested are more likely to buy a telescope and also to have studied the stars, so interest is confounded with telescope ownership.
8. In the survey, the farmers chose their varieties, and farmers who try new varieties may also have better land or equipment, so those differences are confounded with variety. In the experiment, chance decides which plots get each variety, so soil and other conditions are spread evenly; the groups differ only in variety, so a clear difference in yield can be credited to the variety.
9. (a) No: the researcher did not decide which farms milk three times a day; the farms had already chosen. (b) For example, farm size or herd management: larger, more intensive farms are more likely to milk three times a day and may also feed richer rations, which raises milk yield by itself.
10. (a) No one can decide how much rain falls on a farm, so rainfall cannot be imposed as a treatment. (b) Use irrigation instead: on a set of test plots, assign different amounts of irrigation water (for example 0, 1 or 2 inches a week) to the plots by chance, keep everything else the same, and compare the yields.
11. Pasture is confounded with feed. The two pastures may differ in grass, water, shade or shelter, so if the north-pasture calves gain more, we cannot tell whether the feed or the pasture made the difference. The feeds should be assigned to calves by chance, with calves on each feed in the same pasture (or in both pastures equally).
12. Treatments: the new spray and the usual hoof care (a comparison group). Experimental units: the 40 ewes. Assign 20 ewes to each treatment by chance (for example, draw 20 ear-tag numbers from a hat). Keep pasture, feed and handling the same for all. Response variable: whether each ewe becomes lame (or the number of lame days) over the next three months. Compare the two groups.
13. (a) $2 \times 3 \times 2 = 12$ treatments. (b) $12 \times 4 = 48$ plots.
14. First, farm resources: drip systems cost more to install, so farms that use them may be larger or better funded and may also use better seed, fertilizer and equipment, raising yields on their own. Second, soil or location: drip irrigation is common in certain regions or on certain soils; if those places also have richer soil or a better growing climate, yields would be higher regardless of the watering method. In both cases the variable travels with drip irrigation, so its effect cannot be separated from the irrigation's.

LESSON 5.5 · 2 SESSIONS OF 45 MINUTES

Designing Experiments

OBJECTIVES

- State the four principles of experimental design (comparison, random assignment, replication, control) and explain why each matters.
- Carry out the random assignment for a completely randomized design with a table of random digits or by drawing names.
- Explain the placebo effect and the purpose of single-blind and double-blind experiments.
- Interpret 'statistically significant' as a difference too large to be explained by chance alone.

VIDEO

WATCH

Experimental Design Part 1

ProfRobBob · 14:01

When to watch: Before Example 1: control, randomization and replication; treatments, factors and placebos.

https://www.youtube.com/watch?v=v-xnPVCi9wM

KEY TERMS

control group	A group in an experiment that receives no treatment, a placebo, or the usual treatment, so the new treatment can be compared with it.
random assignment	Using a chance process (random digits, names drawn from a hat, a random number generator) to decide which experimental units receive which treatment.
completely randomized design	An experimental design in which all the experimental units are assigned to the treatments completely at random.
placebo	A dummy treatment that looks like the real one but has no active ingredient. The improvement people often feel from any treatment at all is called the placebo effect.
double-blind	An experiment in which neither the subjects nor the people who work with them and measure the response know which treatment each subject received. In a single-blind experiment, only one side is kept from knowing.

statistically significant	Describes an observed effect that is so large it would rarely occur by chance alone, so chance is not a believable explanation.

CORE IDEAS

A good experiment follows four principles. **Comparison:** compare two or more treatments, often including a **control group**, so the effect of the new treatment can be separated from everything else going on (if a farmer puts a new fertilizer on all his corn and harvests a record crop, was it the fertilizer or the warm, rainy summer? Only a control group grown the same summer can tell). **Random assignment:** let chance decide which units get which treatment. **Replication:** use enough units in each group that chance differences between units average out. **Control:** keep other variables the same for every group (same planting date, same watering, same way of measuring).

Why randomize? If the farmer chooses which plots get the new fertilizer, he may, without meaning to, give it to his best plots; then soil quality is confounded with the treatment. When chance makes the assignment, good and poor plots, and every other difference we know about *or don't know about*, tend to be split evenly between the groups. The groups start out alike, so a large difference at the end points to the treatment. This is the reason a randomized experiment can show cause and effect.

In a **completely randomized design**, the units are assigned to the treatments completely at random. Suppose 24 plots are to be split between a new fertilizer and the usual one. Label the plots 01 to 24, read pairs of random digits, and give the new fertilizer to the first 12 different labels from 01 to 24 that appear; the other 12 plots get the usual fertilizer. Or write the 24 labels on equal slips, mix them well in a hat, and draw 12. The outline below is the plan in three steps.

A completely randomized design

Step	What happens
1. Random assignment	The 24 plots are split by chance into two groups of 12
2. Treatments	Group 1 gets the new fertilizer; Group 2 gets the usual fertilizer
3. Compare	Measure the corn yield of every plot and compare the two groups' mean yields

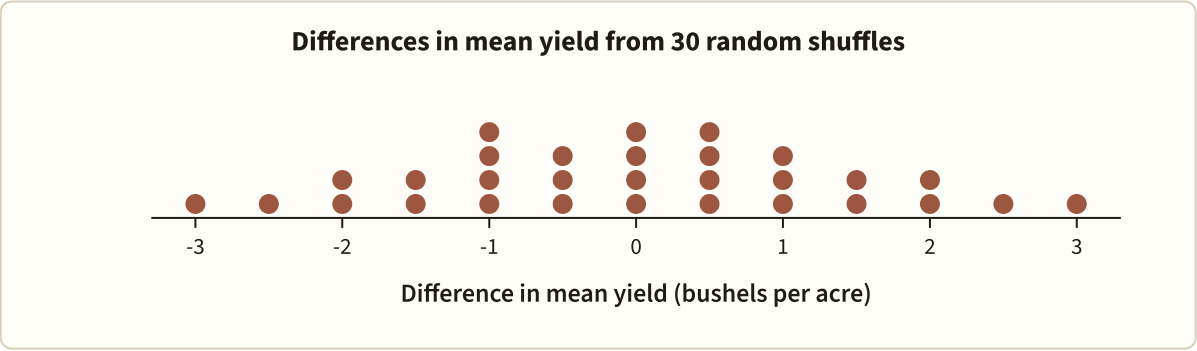
Table of random digits

Line	Digits 1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40
101	48271	90365	12804	77539	26018	54392	81746	03957
102	63518	20947	85132	46079	31685	97240	15863	72094
103	07624	58391	42716	09853	64170	28935	50481	36627
104	91045	36782	70259	13846	82593	41607	69318	25740
105	25930	71468	03587	92614	47082	16359	88720	54193
106	84713	05296	61840	37925	19564	72038	40675	98312

Why replicate? With one plot per treatment, one plot might simply have better soil, and we could not tell luck from a real effect. With 12 plots per treatment, the good and poor plots average out, and chance differences between the groups become small. The more units, the smaller the effect we can detect.

People respond to the mere fact of being treated. A **placebo** is a dummy treatment, such as a cream with no medicine in it, and the improvement people often feel from any treatment at all is the **placebo effect**. To compare fairly, the control group gets the placebo, so both groups have the same expectations. In a **single-blind** experiment the subjects do not know which treatment they have; in a **double-blind** experiment neither the subjects nor the people who measure the response know. Blinding keeps hopes and expectations, on either side, from becoming a confounding variable.

Even when a treatment does nothing, two randomly formed groups differ a little by chance. So when an experiment shows a difference, we ask: is it bigger than chance alone would usually produce? Suppose the fertilizer plots in our experiment averaged 5 bushels per acre more than the usual plots. To see what chance alone can do, the 24 yields were shuffled into two random groups of 12 again and again, 30 times, as if the fertilizer made no difference. Each dot is the difference in mean yield from one shuffle.



Chance alone produced differences between −3 and 3 bushels per acre, never close to 5. A difference of 5 is very unlikely to be explained by chance alone, so we call it **statistically significant**, and because

the treatments were assigned at random, we may conclude that the fertilizer caused the increase. Unit 11 will make this reasoning precise.

From history. Ronald A. Fisher developed the randomized design of experiments in the 1920s while studying crop experiments at the Rothamsted agricultural research station in England. He insisted that treatments be assigned to plots by chance, and he set out the principles of comparison, randomization and replication in his book The Design of Experiments (1935).

Practice habit. Know the four principles of design by heart: Compare, Randomize, Replicate, Control. Check every experiment you read about, and every one you design, against all four.

Example 1. Henry has 40 garden plots and wants to compare a new pole-bean variety with his usual variety, 20 plots each. He labels the plots 01 to 40. Use line 106 of the table of random digits to find the first five plots that will receive the new variety, and explain how the assignment is finished.

Table of random digits

Line	Digits 1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40
101	48271	90365	12804	77539	26018	54392	81746	03957
102	63518	20947	85132	46079	31685	97240	15863	72094
103	07624	58391	42716	09853	64170	28935	50481	36627
104	91045	36782	70259	13846	82593	41607	69318	25740
105	25930	71468	03587	92614	47082	16359	88720	54193
106	84713	05296	61840	37925	19564	72038	40675	98312

1. Label: the plots are already labeled 01 to 40 (two digits each).
2. Randomize: read line 106 in pairs: 84, 71, 30, 52, 96, 61, 84, 03, 79, 25, 19, 56, 47, 20, ...
3. Select: skip 84 and 71 (greater than 40); keep 30; skip 52, 96, 61, 84; keep 03; skip 79; keep 25; keep 19; skip 56 and 47; keep 20.
4. The first five plots for the new variety are 30, 03, 25, 19 and 20.
5. Keep reading, skipping repeats and numbers above 40, until 20 different plots are chosen; the other 20 plots get the usual variety.
6. Conclusion: because chance chose the plots, sunny and shady spots and good and poor soil should be spread about evenly between the two varieties, so a clear difference in harvest can be credited to the variety.

Answer: Plots 30, 03, 25, 19 and 20 are the first five for the new variety. Continue until 20 different plots are chosen; the remaining 20 get the usual variety.

Example 2. A cross-country coach wants to know whether a new sore-muscle cream reduces soreness the day after a hard workout. Thirty runners volunteer. (a) Describe a completely randomized, double-blind design. (b) The results are in the table. Find the difference in mean soreness ratings (placebo minus cream). (c) The coach shuffled the 30 ratings into two random groups of 15 one hundred times, and a difference as large as the one observed appeared only once. Is the result statistically significant? Answer yes or no, and explain.

Sore-muscle cream experiment

Group	Number of runners	Mean soreness rating (0 to 10)
New cream	15	3.1
Placebo cream	15	4.6

1. (a) Label the runners 01 to 30 and use random digits (or names drawn from a hat) to choose 15 for the new cream; the other 15 get a placebo cream that looks and smells the same.
2. The jars are marked only with codes, so neither the runners nor the coach who records the soreness ratings knows who has which cream: the experiment is double-blind.
3. After the same hard workout, every runner rates soreness the next morning on a 0 to 10 scale, and the two groups' mean ratings are compared.
4. (b) Difference $= 4.6 - 3.1 = 1.5$ points.
5. (c) When the ratings were shuffled as if the cream did nothing, chance produced a difference this large only 1 time in 100.
6. So chance alone is not a believable explanation: the result is statistically significant.
7. Conclusion: because the creams were assigned at random and the experiment was double-blind, there is good evidence that the new cream reduces next-day soreness for runners like these.

Answer: (a) Randomly assign 15 runners to the new cream and 15 to an identical-looking placebo; code the jars so neither the runners nor the coach recording soreness knows who got which; compare mean next-morning soreness ratings. (b) 1.5 points. (c) Yes: a difference this large happened by chance only once in 100 shuffles, so it is too large to be explained by chance alone.

PRACTICE

1. EASY Name the four principles of experimental design and give a one-sentence reason for each.
2. EASY A farmer used a new fertilizer on all of his wheat this year and harvested 15% more than last year. Explain why this does not show that the fertilizer works, and how a control group would help.

3. EASY A 4-H member has 20 hens, labeled 01 to 20, and will give a new feed to 10 of them. Use line 104 of the table of random digits to find the first 3 hens that get the new feed.

Table of random digits

Line	Digits 1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40
101	48271	90365	12804	77539	26018	54392	81746	03957
102	63518	20947	85132	46079	31685	97240	15863	72094
103	07624	58391	42716	09853	64170	28935	50481	36627
104	91045	36782	70259	13846	82593	41607	69318	25740
105	25930	71468	03587	92614	47082	16359	88720	54193
106	84713	05296	61840	37925	19564	72038	40675	98312

4. EASY A completely randomized experiment compares 3 tomato fertilizers on 36 plants, with the same number of plants for each fertilizer. How many plants get each fertilizer?

5. MEDIUM In an experiment comparing two throat lozenges for choir members with scratchy throats, the singers do not know which lozenge they are given, but the choir director who asks how their throats feel does know. Is the experiment single-blind or double-blind? Why might this matter?

6. MEDIUM A gardener has 30 rose cuttings, labeled 01 to 30, and will try 3 rooting powders, 10 cuttings each. Use line 103 of the table of random digits to find the first 4 cuttings assigned to powder 1.

Table of random digits

Line	Digits 1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40
101	48271	90365	12804	77539	26018	54392	81746	03957
102	63518	20947	85132	46079	31685	97240	15863	72094
103	07624	58391	42716	09853	64170	28935	50481	36627
104	91045	36782	70259	13846	82593	41607	69318	25740
105	25930	71468	03587	92614	47082	16359	88720	54193
106	84713	05296	61840	37925	19564	72038	40675	98312

7. MEDIUM A farmer testing a new seed variety says, 'I'll plant the new seed in the plots I know are best, so it gets a fair chance.' Explain why he should assign the plots at random instead.

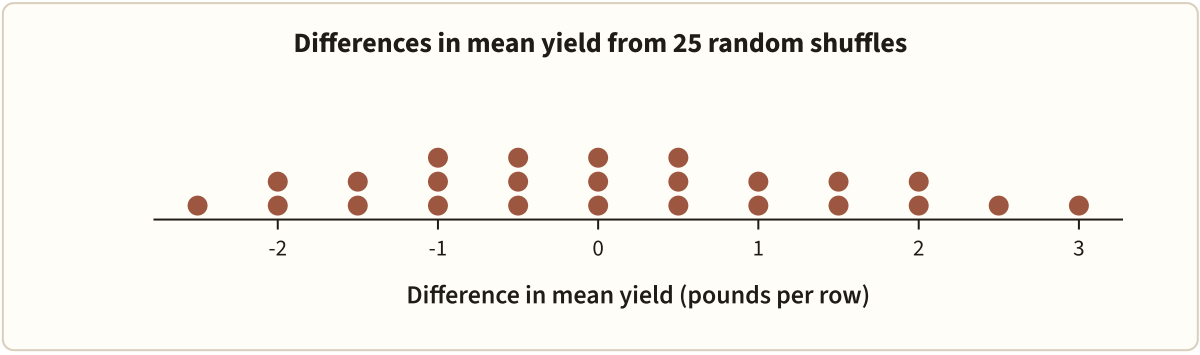
8. MEDIUM Ruth wants to compare two potting soils by growing one pepper plant in each. Explain why one plant per soil is not enough.

9. MEDIUM A coach randomly assigned players to two free-throw routines and reports that the difference in their mean free-throw percentages was statistically significant. Explain what that means in this context.

10. HARD An agricultural student has 24 plots and wants to compare 3 oat varieties. Describe a completely randomized design, including how to do the random assignment.

11. HARD A riding stable wants to test whether a new hoof ointment reduces cracked hooves in its 24 horses. Explain how to make this experiment as blind as possible, and why it matters even though horses do not have expectations.

12. HARD In an experiment on two pole-bean varieties, the mean yield of variety A was 1.5 pounds per row more than variety B. To see what chance could do, the yields were shuffled into two random groups 25 times; the dotplot shows the 25 differences. (a) How many of the shuffled differences were 1.5 or more? (b) Is the observed difference of 1.5 statistically significant? Answer yes or no.



13. CHALLENGE Grace wants to know whether a new bread recipe rises higher than her usual one. She bakes 6 loaves of the new recipe on a dry Monday and 6 loaves of the usual recipe on a humid Tuesday, and the new loaves rise higher. Identify the flaw, explain why it matters, and describe a better design.

COMMON MISTAKES

- Assigning treatments by judgment ('the healthiest calves get the new feed') or haphazardly. Only a chance process makes the groups alike, so always randomize.
- Confusing random assignment with random sampling. Random sampling decides who is in the study; random assignment decides which treatment each one gets. Experiments need random assignment.
- Leaving out the control group or the placebo. Without a comparison, the weather, the season or the placebo effect could explain any change.
- Reading 'statistically significant' as 'large' or 'important'. It means only that the difference is too big to be explained by chance alone.

ANSWERS TO PRACTICE

1. Comparison: compare treatments (often with a control group) so the treatment's effect can be separated from other influences. Random assignment: let chance assign treatments so the groups start out alike. Replication: use enough units that chance differences between units average out. Control: keep other variables the same for all groups so they cannot cause a difference.
2. This year's weather, pests or seed could have caused the increase, and with no comparison group there is no way to separate those effects from the fertilizer's. A control group of plots grown the same year with the usual fertilizer would face the same weather, so a difference between the groups could be credited to the fertilizer.
3. Pairs: 91, 04, 53, 67, 82, 70, 25, 91, 38, 46, 82, 59, 34, 16, 07. Keep only labels 01 to 20: hens 04, 16 and 07.
4. $\frac{36}{3} = 12$ plants.
5. Single-blind: only the singers are kept from knowing. The director's expectations could affect how she asks the question or records the answers, so knowing the treatment could bias the results. Keeping her unaware too (double-blind) would remove that risk.
6. Pairs: 07, 62, 45, 83, 91, 42, 71, 60, 98, 53, 64, 17, 02, 89, 35, 50, 48, 13. Keep only labels 01 to 30: cuttings 07, 17, 02 and 13.
7. If the new seed goes on the best plots, soil quality is confounded with the variety: a higher yield could be due to the better soil, not the seed. Random assignment spreads good and poor plots evenly between the varieties, so the groups start out alike and any clear difference can be credited to the seed.
8. Plants differ naturally: one seedling might simply be stronger than the other. With one plant per soil, a difference could be due to that luck rather than the soil. Replication, using many plants per soil, lets these chance differences average out so a real difference between the soils can show.
9. It means the difference between the two groups' mean free-throw percentages was too large to be easily explained by the chance of which players happened to land in which group. Since chance alone would rarely produce a difference that big, there is good evidence that the routines really differ in effect.
10. Label the plots 01 to 24. Using a table of random digits (or slips drawn from a hat), choose 8 different labels for variety A, then 8 more of the remaining labels for variety B; the last 8 plots get variety C. Plant, care for and harvest all plots in the same way. Response: oat yield of each plot. Compare the mean yields of the three varieties.

11. Randomly assign 12 horses to the new ointment and 12 to a plain ointment that looks the same, with the jars marked only by codes. The person who examines the hooves and judges the cracking should not know which horse got which ointment (and neither should the grooms who apply it, if possible). The horses cannot be fooled or helped by expectations, but the people judging the hooves can be: someone who knows which horses got the new ointment might, without meaning to, rate their hooves as better.
12. (a) The values 1.5, 1.5, 2, 2, 2.5 and 3: 6 of the 25 shuffles. (b) No: chance alone produced a difference at least this large in about 1 out of every 4 shuffles, so the observed difference could easily be explained by chance. This does not prove the varieties are the same; the experiment just did not give convincing evidence of a difference.
13. The day (and its humidity) is confounded with the recipe: every new-recipe loaf was baked on Monday, so a difference in rise could be due to the weather rather than the recipe. Better: bake both recipes on the same days and randomize, for example by labeling 12 baking slots across the two days 01 to 12 and using random digits to choose which 6 slots get the new recipe; keep the oven, pans and rising time the same, and compare the mean rise.

LESSON 5.6 · 2 SESSIONS OF 45 MINUTES

Blocking, Matched Pairs and the Scope of Conclusions

OBJECTIVES

- Describe a randomized block design and explain why blocking makes the comparison of treatments more precise.
- Describe a matched pairs design and summarize its results using the differences within pairs.
- Determine the scope of a study's conclusions: when we may generalize to a population and when we may conclude cause and effect.
- Explain why honest reporting of how data were collected is necessary for others to judge a conclusion.

VIDEO

WATCH

Experimental Design Part 2

ProfRobBob · 13:55

When to watch: Before Example 1: block designs, matched pairs and double-blind experiments.

https://www.youtube.com/watch?v=cKj4TULve0s

KEY TERMS

block	A group of experimental units that are similar in some way, known before the experiment, that is expected to affect the response (for example, plots in the same field, or lambs of the same breed).
randomized block design	An experimental design in which the units are first divided into blocks, and then the treatments are randomly assigned separately within each block.
matched pairs design	A randomized block design with blocks of two: either each subject receives both treatments in a random order, or two very similar units are paired and the treatments are randomly assigned within each pair.
scope of inference	How far the conclusions of a study can reach: random sampling allows generalizing to the population sampled, and random assignment allows concluding cause and effect.

CORE IDEAS

Sometimes we know before an experiment begins that the units differ in a way that will affect the response. Wheat on rich creek-bottom land yields far more than wheat on a stony hillside, whatever the variety. A **block** is a group of units that are alike in such a way. In a **randomized block design**, we form the blocks first and then randomly assign the treatments separately within each block. Blocking in an experiment is like stratifying in a sample: in both, we group similar individuals on purpose and let chance work inside each group.

Why block? The table shows the yields when two wheat varieties were each planted, by random choice, in one half of each of three very different fields.

Wheat yields in three fields (bushels per acre)

Field (block)	Variety A	Variety B	Difference (B - A)
Stony hillside	38	42	4
Creek bottom	60	64	4
Upland	48	51	3

The fields differ by more than 20 bushels per acre, but within every field variety B beats variety A by 3 or 4 bushels. If the varieties had been assigned completely at random across all the plots, the large field-to-field differences would be mixed into the comparison, and a 3- or 4-bushel difference could be hidden among them. Comparing the varieties *within* each field removes the field differences from the comparison, the same way you would compare two runners fairly by having them race on the same track on the same day. Blocking does not make the units alike; it makes each comparison between like units.

A **matched pairs design** is blocking with blocks of two. Either each subject receives both treatments in a random order (each runner tries both warm-up routines on different days, with a coin flip deciding which comes first), or we pair up two very similar units (twin lambs, two halves of the same field) and randomly assign one treatment to each member of the pair. Then we analyze the *differences* within the pairs. Each subject is compared with itself, so the differences between subjects (some runners are simply faster) drop out. The random order matters: if every runner tried routine B second, practice or fresher weather on the second day would be confounded with routine B.

Chance plays two different roles in collecting data, and each gives a different power. **Random sampling**, choosing the individuals from a population at random, lets us generalize from the sample to that population. **Random assignment**, giving out the treatments at random, lets us conclude that the treatments caused the differences we see. A study can have one, both, or neither, as the table shows.

The scope of inference

	Treatments assigned at random	Treatments not assigned at random
Individuals chosen at random	Cause and effect; generalize to the population	Generalize to the population; no conclusion about cause
Individuals not chosen at random	Cause and effect for individuals like these only	Describes only the individuals studied; no cause

Most experiments use volunteers or the units at hand (the plots on one farm, the runners on one team) rather than a random sample, so their cause-and-effect conclusions apply to units like those in the study. Most surveys use random samples but assign nothing, so they describe the population but cannot show cause. When neither is present, the data describe only the individuals studied.

Because the reach of a conclusion depends on how the data were collected, honest reporting must say how: how the individuals were chosen, how the treatments were assigned, how many units dropped out, and every comparison that was made, not only the ones that came out well. Running many comparisons and reporting only the one that looked impressive misleads readers, since by chance alone some comparisons will look impressive. Truthfulness with data is part of doing statistics well.

KEY FORMULAS

$$d = (\text{treatment 1 result}) - (\text{treatment 2 result}) \text{ for each pair}, \qquad \bar{d} = \frac{\sum d}{n}$$

Practice habit. Ask two separate questions about every study: Were the individuals chosen at random? (Then we may generalize to the population.) Were the treatments assigned at random? (Then we may conclude cause and effect.)

WORKED EXAMPLES

Example 1. A sheep farmer wants to compare two feeds on 30 lambs: 12 Suffolk lambs and 18 Dorset lambs. Suffolk lambs naturally gain weight faster than Dorsets. (a) Describe a randomized block design that gives each feed to half of each breed. How many Suffolk lambs and how many Dorset lambs get feed A? (b) Explain why blocking by breed is better than a completely randomized design here.

1. (a) The blocks are the breeds, because breed is known in advance to affect weight gain.
2. Suffolk block: label the 12 Suffolk lambs 01 to 12 and use random digits to choose $\frac{12}{2} = 6$ for feed A; the other 6 get feed B.
3. Dorset block: label the 18 Dorset lambs 01 to 18 and choose $\frac{18}{2} = 9$ for feed A; the other 9 get feed B.
4. After the feeding period, compare the weight gains on the two feeds within each breed.
5. (b) In a completely randomized design, chance might put, say, 8 of the 12 fast-gaining Suffolks on feed A. Feed A would then look better partly because of breed.
6. Blocking guarantees each feed gets exactly half of each breed, and comparing within breeds removes the breed differences from the comparison.
7. Conclusion: the block design gives a fairer and more precise comparison of the two feeds.

Answer: (a) Within each breed, randomly assign half the lambs to each feed: 6 Suffolk and 9 Dorset lambs get feed A (and 6 and 9 get feed B). (b) Breed affects weight gain, so a completely random assignment could, by chance, put more Suffolks on one feed. Blocking balances the breeds and compares the feeds within each breed, removing breed differences from the comparison.

Example 2. Six runners on a cross-country team each ran 400 meters twice, a week apart: once after warm-up routine A and once after routine B, with a coin flip deciding which routine each runner used first. The table gives their times. (a) Find each runner's difference, A − B. (b) Find the mean difference. (c) Explain why this matched pairs design is better than assigning three runners to each routine.

400-meter times after two warm-up routines

Runner	Routine A (seconds)	Routine B (seconds)
Samuel	62	60
Ruth	70	69
Henry	58	57
Grace	66	63
Thomas	74	72
Anna	60	57

1. (a) Differences A − B: Samuel $62 - 60 = 2$, Ruth $70 - 69 = 1$, Henry $58 - 57 = 1$, Grace $66 - 63 = 3$, Thomas $74 - 72 = 2$, Anna $60 - 57 = 3$.
2. (b) Sum of differences $= 2 + 1 + 1 + 3 + 2 + 3 = 12$.
3. Mean difference $\bar{d} = \dfrac{12}{6} = 2.0$ seconds.
4. (c) The runners themselves differ a lot: their times range from 57 to 74 seconds, a spread of 17 seconds.
5. With three runners per routine, chance could put the faster runners in one group, and that difference could easily swamp a 2-second effect.
6. In the matched pairs design each runner is compared with himself or herself, so differences in speed between runners cancel out; the differences all fall between 1 and 3 seconds.
7. Conclusion: on average, these runners were 2.0 seconds faster after routine B, and all six were faster with B.

Answer: (a) 2, 1, 1, 3, 2, 3 seconds. (b) 2.0 seconds (routine B faster). (c) Each runner serves as his or her own comparison, so the large differences in speed between runners (57 to 74 seconds) drop out, and the small, consistent effect of the routine can be seen.

PRACTICE

1. EASY Name the design. (a) 20 tomato plants are split at random into two groups of 10, one for each fertilizer. (b) Ten pairs of twin lambs: a coin flip decides which twin in each pair gets the new feed. (c) An orchard's trees are grouped by variety, and within each variety half the trees are chosen at random for a new spray.
2. EASY An orchard has 40 trees for an experiment: 24 Gala and 16 Fuji. In a randomized block design by variety, half of each variety gets a new spray. How many Gala trees and how many Fuji trees get the new spray?
3. EASY A survey of a random sample of 100 of a county's 2,000 farms found that 38% grow hay. Can this result be generalized to all 2,000 farms in the county? Answer yes or no.
4. EASY What does random assignment of treatments allow us to conclude that random sampling alone does not?
5. MEDIUM Five dairy cows were each milked for one week on schedule 1 and one week on schedule 2, in a random order for each cow. The table gives each cow's average daily milk. Find the mean difference (schedule 2 − schedule 1), to one decimal place.

Average daily milk (pounds)

Cow	Schedule 1	Schedule 2
Daisy	56	60
Bess	62	64
Clover	50	56
Rosie	68	70
Molly	60	64

6. MEDIUM Forty volunteers from a town's walking club were randomly assigned to wear either a new or an old style of walking shoe for a month, and the numbers of blisters were compared. (a) Can the club conclude that the shoe style caused a clear difference in blisters? (b) Can the results be generalized to all adults in the town? Answer yes or no to each.
7. MEDIUM An agriculture student is comparing two oat varieties on a farm with four fields of very different soil. Explain why she should block by field rather than assign varieties to plots completely at random across the farm.
8. MEDIUM A company wants to compare two hand creams for dry, cracked hands among 20 farm workers. Describe a matched pairs design and explain the role of the coin flip.

9. MEDIUM A magazine invited readers to report how many hours a week they garden and how rested they feel. Readers who gardened more reported feeling more rested. (a) Can we conclude that gardening causes people to feel more rested? (b) Can we generalize the result to all readers of the magazine? Answer yes or no to each.

10. HARD Two oat varieties were planted in each of four fields, with a coin flip deciding which half of each field got variety A. The table gives the yields. (a) Find the mean of the differences (B − A). (b) Find the range of variety A's yields across the four fields. (c) Use your answers to explain why blocking by field was a good idea.

Oat yields in four fields (bushels per acre)

Field (block)	Variety A	Variety B
Field 1	45	49
Field 2	62	65
Field 3	38	43
Field 4	55	59

11. HARD A feed company tested its new calf feed in five separate experiments. Only one showed a statistically significant gain, and the company's advertisement reports only that one. Explain why this is misleading.

12. HARD A poultry farm with 600 hens chose an SRS of 60 hens, then randomly assigned 30 of them to a new feed and 30 to the usual feed. The new-feed hens laid significantly more eggs. (a) Can the farm conclude that the new feed caused the increase? (b) Can the farm generalize this to all 600 hens? Answer yes or no to each.

13. CHALLENGE An orchard owner wants to compare two pruning methods on 20 apple trees that range in age from 5 to 30 years. Older trees produce much more fruit. Describe a matched pairs design, explain how to randomize, and explain why it is better than a completely randomized design.

COMMON MISTAKES

- Thinking of blocks as treatments. Blocks are groups you form before the experiment from differences that already exist (field, breed, age); treatments are what you assign within the blocks.
- Forgetting to randomize within each block, or to randomize the order of treatments in a matched pairs design. Without the coin flip, order or position can be confounded with the treatment.
- Mixing up the two powers of chance. Random sampling lets you generalize to a population; random assignment lets you conclude cause. Check for each one separately.
- Analyzing matched pairs data as if the two treatments were given to separate groups. Work with the differences within the pairs; that is where the design's advantage lies.

ANSWERS TO PRACTICE

1. (a) Completely randomized design. (b) Matched pairs design. (c) Randomized block design (blocked by variety).
2. Gala: $\frac{24}{2} = 12$. Fuji: $\frac{16}{2} = 8$.
3. Yes. The farms were chosen at random from all 2,000, so the sample result can be generalized to the county's farms (allowing for chance variation).
4. Random assignment allows us to conclude cause and effect: because chance made the groups alike except for the treatment, a clear difference in the response can be credited to the treatment. Random sampling only allows us to generalize from the sample to the population.
5. Differences: $60 - 56 = 4$, $64 - 62 = 2$, $56 - 50 = 6$, $70 - 68 = 2$, $64 - 60 = 4$. Mean $= \frac{4 + 2 + 6 + 2 + 4}{5} = \frac{18}{5} = 3.6$ pounds per day more on schedule 2.
6. (a) Yes: the shoes were assigned at random, so a clear difference can be credited to the shoe style. (b) No: the subjects were volunteers from a walking club, not a random sample of the town's adults, so the results apply to walkers like these.
7. Soil differences between fields will cause large differences in yield. If varieties are assigned at random across the whole farm, one variety might by chance get more plots in the best field, and the field effect would be mixed into the comparison. Blocking by field puts both varieties in every field and compares them within each field, removing the field-to-field differences so a smaller difference between the varieties can be detected.
8. Each worker uses both creams, one on each hand, for two weeks; then the dryness of each hand is rated and the difference between the two hands is found for each worker. A coin flip decides for each worker which hand gets cream A. Without it, cream A might always go on the right hand, which most people use more for work, so hand use would be confounded with the cream. Each worker serves as his or her own comparison, removing differences between workers' skin.
9. (a) No: no treatment was assigned, so this is an observational study, and other variables (free time, health) could explain the link. (b) No: the readers chose themselves (a voluntary response sample), so they do not represent all readers.
10. (a) Differences: $49 - 45 = 4$, $65 - 62 = 3$, $43 - 38 = 5$, $59 - 55 = 4$; mean $= \frac{16}{4} = 4.0$ bushels per acre. (b) $62 - 38 = 24$ bushels per acre. (c) The fields differ by as much as 24 bushels per acre, six times the 4-bushel difference between varieties. Without blocking, those field differences could hide or fake a difference between varieties; within each field, B beat A every time, by 3 to 5 bushels.

11. Even if the feed has no effect, chance alone will sometimes produce a difference large enough to be called significant. Reporting one success out of five, and hiding the four that showed nothing, makes chance look like a real effect. Honest reporting would describe all five experiments and their results.
12. (a) Yes: the feeds were assigned at random, so the significant difference can be credited to the feed. (b) Yes: the 60 hens were a random sample of the 600, so the conclusion applies to the whole flock.
13. Rank the 20 trees by age and pair them: the two youngest, the next two, and so on, giving 10 pairs of trees of nearly the same age. In each pair, flip a coin to decide which tree gets pruning method 1; the other gets method 2. At harvest, find the difference in fruit yield within each pair and average the 10 differences. Age has a large effect on yield; in a completely randomized design, chance could give one method more old trees, and the age effect would be mixed into the comparison. Pairing compares the methods on trees of the same age, removing that variation.

Unit 5 test

You have 60 minutes. A calculator is allowed, and the table of random digits you need is printed with the problems. Show your work for every calculation, and answer explanation questions in complete sentences in the context of the problem. Give percents exactly or to one decimal place, and means to one more decimal place than the data, unless a problem says otherwise. The test is worth 100 points.

1. **5 PTS** A sawmill inspector measures the thickness of 25 of the 1,500 boards cut on Tuesday. (a) How many individuals are in the population? (b) How many are in the sample?
2. **6 PTS** A local weather station's website asked visitors, 'Was last week's storm forecast accurate?' Of the 220 visitors who answered, 176 said no. (a) What percent said no? (b) Name the type of sample and explain why the result may not describe all of the station's viewers.
3. **6 PTS** The table gives the weights of the 8 pumpkins grown in a 4-H plot. Pumpkins 1, 2 and 3 grew next to the water spigot. (a) Find the mean weight of all 8 pumpkins. (b) A member estimates the mean by weighing only the three pumpkins by the spigot, which are closest to the gate. Find the mean of this sample. (c) Explain why this sampling method is biased.

Pumpkins in a 4-H plot

Pumpkin	Location	Weight (pounds)
1	By the spigot	18
2	By the spigot	22
3	By the spigot	14
4	Far end	12
5	Far end	10
6	Far end	13
7	Far end	11
8	Far end	12

4. **7 PTS** A choir has 45 members, labeled 01 to 45. The director wants an SRS of 4 members to try out new music folders. Use line 107 of the table of random digits to choose the sample, and show which pairs you kept and skipped.

Table of random digits

Line	Digits 1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40
107	52806	13947	69025	38174	90613	27458	06392	81570
108	31792	64058	17326	95481	03769	28514	70936	45201

5. **6 PTS** At the county fair, 300 young people are showing exhibits: 150 livestock, 90 crops and 60 crafts. The fair board wants a stratified random sample of 30 exhibitors, with each group's share of the sample equal to its share of all exhibitors. How many exhibitors should be chosen from each group?

6. **5 PTS** A seed company wants a systematic random sample of 60 customers from its list of 1,200. (a) What is k, the spacing between chosen customers? (b) If the random start is customer 13, what is the number of the third customer chosen?

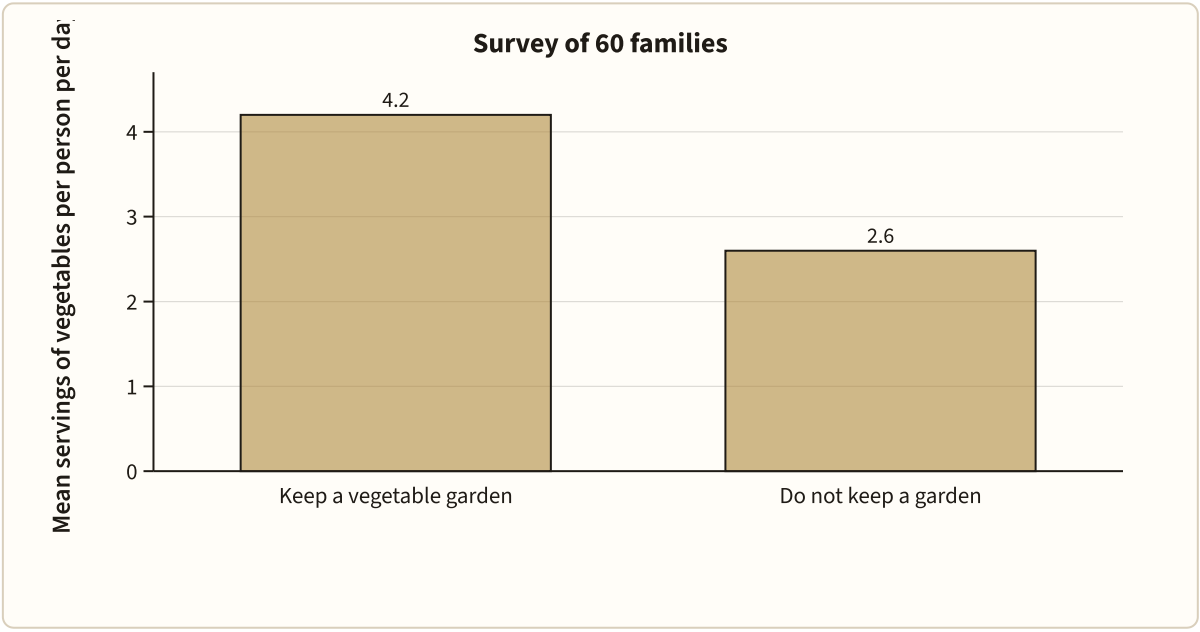
7. **6 PTS** A beekeeper has 30 hives in a clover meadow and 50 hives at the edge of the woods. Hives in the clover produce much more honey. Explain why a sample stratified by location would estimate the mean honey yield per hive more precisely than an SRS of the same size.

8. **5 PTS** The county historical society mailed a questionnaire to an SRS of 500 of its members, and 320 were returned. What was the response rate, as a percent?

9. **6 PTS** The historical society's questionnaire asked: 'Shouldn't the society spend its savings to restore the beautiful old schoolhouse before it is lost forever?' Name the problem with this question and rewrite it fairly.

10. **6 PTS** A golf course left cards in the clubhouse asking golfers whether the greens are too fast. It received 85 cards. The manager plans to keep collecting cards the same way until there are 850. Will the larger number of cards remove the bias? Answer yes or no, and explain.

11. **7 PTS** A survey of 60 families compared families that keep a vegetable garden with families that do not. The bar graph shows the mean number of servings of vegetables each person eats per day. (a) How many more servings per day did members of gardening families eat, on average? (b) Does this study show that keeping a garden causes people to eat more vegetables? Answer yes or no. (c) Name a possible confounding variable and explain how it could produce the difference.



12. **6 PTS** A greenhouse experiment tests 3 watering schedules and 2 fertilizers on pepper plants, with 5 pots for each treatment. (a) How many treatments are there? (b) How many pots are needed?

13. **7 PTS** Samuel has 24 tomato seedlings, labeled 01 to 24, and will plant 12 in a new potting soil and 12 in his usual soil. (a) Use line 108 of the table of random digits to find the first 4 seedlings assigned to the new soil. (b) Explain why he should assign the seedlings at random rather than putting the healthiest-looking seedlings in the new soil.

Table of random digits

Line	Digits 1-5	6-10	11-15	16-20	21-25	26-30	31-35	36-40
107	52806	13947	69025	38174	90613	27458	06392	81570
108	31792	64058	17326	95481	03769	28514	70936	45201

14. **7 PTS** A company wants to test whether its new cold remedy shortens colds. Fifty adults with fresh colds volunteer. Describe a completely randomized, double-blind experiment, and explain why a placebo is needed.

15. **8 PTS** Henry wanted to compare two chainsaw blades. On each of 5 logs he made one cut with blade A and one with blade B, with a coin flip deciding which blade cut first. The table gives the cutting times. (a) Find the mean of the differences (A − B), to one decimal place. (b) Explain why this matched pairs design is better than cutting 5 logs with blade A and 5 different logs with blade B.

Time to cut through each log (seconds)

Log	Blade A	Blade B
1	42	39
2	55	51
3	38	37
4	61	56
5	49	47

16. **7 PTS** A large high school chose a random sample of 80 of its 800 students. Students in the sample who play a sport reported more hours of sleep, on average, than those who do not. (a) Can the result be generalized to all 800 students? (b) Can we conclude that playing a sport causes students to sleep more? Answer yes or no to each, and explain briefly.

UNIT 5 TEST: ANSWER KEY

1. (a) 1,500 boards (all boards cut on Tuesday). (b) 25 boards.
2. (a) $\frac{176}{220} = 0.80$, so 80%. (b) A voluntary response sample: visitors chose whether to answer, and viewers who were unhappy with the forecast are the most likely to bother, so 80% probably overstates how many viewers think the forecast was wrong.
3. (a) $\frac{18+22+14+12+10+13+11+12}{8} = \frac{112}{8} = 14$ pounds. (b) $\frac{18+22+14}{3} = \frac{54}{3} = 18$ pounds. (c) It is a convenience sample that always takes the pumpkins by the spigot, which get the most water and grow largest, so it overestimates the mean every time (here by 4 pounds).
4. Pairs: 52 (skip), 80 (skip), 61 (skip), 39 (keep), 47 (skip), 69 (skip), 02 (keep), 53, 81, 74, 90, 61 (skip), 32 (keep), 74, 58 (skip), 06 (keep). The sample is members 39, 02, 32 and 06.
5. Livestock: $30 \times \frac{150}{300} = 15$. Crops: $30 \times \frac{90}{300} = 9$. Crafts: $30 \times \frac{60}{300} = 6$. Check: $15 + 9 + 6 = 30$.
6. (a) $k = \frac{1200}{60} = 20$. (b) The sample is 13, 33, 53, ..., so the third customer is 53.
7. Hives in the same location produce similar amounts of honey, but the two locations differ a lot. An SRS might by chance include too many clover hives (estimate too high) or too many woods hives (estimate too low). Stratifying guarantees each location its proper share of the sample, so that source of chance error is removed and the estimate varies less from sample to sample.
8. $\frac{320}{500} \times 100\% = 64\%$.
9. It is a leading question (wording bias): 'Shouldn't ...' invites agreement, and 'beautiful' and 'lost forever' push toward yes. A fair version: 'Should the society use its savings to restore the old schoolhouse? (Yes / No / No opinion)'.
10. No. Golfers still choose whether to fill out a card, so those with strong complaints are overrepresented. Ten times as many cards would make the result steadier but would leave it off in the same direction; only a different method, such as asking a random sample of golfers, would fix the bias.
11. (a) $4.2 - 2.6 = 1.6$ servings per day. (b) No: the families chose whether to garden, so this is an observational study. (c) For example, a family's interest in healthy eating: families who already value vegetables are more likely both to plant a garden and to eat vegetables, whether home-grown or bought. That interest is confounded with gardening.
12. (a) $3 \times 2 = 6$ treatments. (b) $6 \times 5 = 30$ pots.

13. (a) Pairs: 31, 79, 26, 40, 58 (skip), 17 (keep), 32, 69, 54, 81 (skip), 03 (keep), 76, 92, 85 (skip), 14 (keep), 70, 93, 64, 52 (skip), 01 (keep). The first four are seedlings 17, 03, 14 and 01. (b) If the healthiest seedlings go into the new soil, seedling health is confounded with soil, and better growth could be due to the seedlings rather than the soil. Random assignment spreads strong and weak seedlings evenly, so the groups start out alike and a clear difference can be credited to the soil.
14. Label the volunteers 01 to 50 and use random digits (or names drawn from a hat) to choose 25 for the remedy; the other 25 get a placebo that looks and tastes the same. Code the packages so neither the volunteers nor the people recording their symptoms know who got which (double-blind). Response: the number of days until symptoms are gone; compare the two groups' means. A placebo is needed because people often feel better simply from being treated (the placebo effect); giving the control group a placebo makes expectations the same in both groups, so any extra improvement can be credited to the remedy.
15. (a) Differences: $42 - 39 = 3$, $55 - 51 = 4$, $38 - 37 = 1$, $61 - 56 = 5$, $49 - 47 = 2$; mean $= \frac{15}{5} = 3.0$ seconds (blade B faster). (b) Logs differ a lot in thickness and hardness (times range from 37 to 61 seconds). Cutting the same log with both blades compares them on equal wood, so the log-to-log differences drop out and the small, steady difference between blades shows clearly.
16. (a) Yes: the students were a random sample of the 800, so the result can be generalized to the school (allowing for chance variation). (b) No: students chose whether to play sports; nothing was assigned at random, so other variables (such as daily schedules or health) could explain the difference.

UNIT 6

Probability

Every statistical conclusion rests on one question: could this result have happened by chance alone? To answer it, we need a careful language for chance. In this unit you will learn that probability describes the long-run pattern of random outcomes. You will estimate probabilities by simulation, and you will work with the basic rules: complements, the addition rules, conditional probability, independence and the multiplication rules, tree diagrams, and counting with permutations and combinations.

The unit builds on the two-way tables of Unit 1 and the random digits and random sampling of Unit 5. The emphasis is on reasoning, not formulas alone: why the complement makes "at least one" easy, why the overlap is subtracted in the addition rule, why P(A | B) is not P(B | A), and why a test for a rare condition can give many false alarms. Parents will find the contexts familiar: seeds and harvests, weather forecasts, free throws, choir committees and county fairs.

These ideas carry the rest of the course. Unit 7 turns probabilities into random variables with means and standard deviations, Unit 8 uses the multiplication rule and combinations to build the binomial distribution, and from Unit 9 on, every confidence interval and significance test is a statement about probability.

PREREQUISITES

- Working with fractions, decimals and percents, and converting among them
- Reading two-way tables and computing conditional distributions (Lesson 1.3)
- Choosing a simple random sample with a table of random digits (Lesson 5.2)
- The idea of random assignment and of results "too large to be explained by chance" (Lesson 5.5)
- Exponents and order of operations with a calculator

MASTERY CHECKLIST

☐ I can explain probability as a long-run relative frequency and state the law of large numbers.
☐ I can explain why the "law of averages" is a myth: a coin has no memory.
☐ I can design and carry out a simulation with random digits (state the model, assign digits, simulate, conclude) and interpret the result.
☐ I can list a sample space and find probabilities when outcomes are equally likely.
☐ I can use the complement rule and the addition rules, and explain why the overlap is subtracted.
☐ I can find a conditional probability from a two-way table or with $P(A \mid B) =$

$\frac{P(A \text{ and } B)}{P(B)}$, and explain why P(A | B) is usually not P(B | A).

☐ I can use the general multiplication rule, including for draws without replacement.

☐ I can check whether two events are independent and use the multiplication rule for independent events.

☐ I can solve "at least one" problems with $1 - P(\text{none})$, and I know that mutually exclusive events (each with positive probability) are not independent.

☐ I can use a tree diagram to find the probability of an event and a reverse conditional probability, and interpret it in context.

☐ I can count with the multiplication principle, factorials, permutations and combinations, and use counts to find probabilities.

☐ I can explain why $\binom{n}{r} = \frac{{}_nP_r}{r!}$ and read values of $\binom{n}{r}$ from Pascal's triangle.

LESSON 6.1 · 2 SESSIONS OF 45 MINUTES

Randomness and Simulation

OBJECTIVES

- Explain what makes a phenomenon random and interpret a probability as a long-run relative frequency.
- State the law of large numbers and explain why the "law of averages" is a myth.
- Carry out a simulation with random digits: state the model, assign digits, simulate, and conclude.
- Interpret simulation results in context and explain why more trials give a more trustworthy estimate.

VIDEO

WATCH

Experimental Probability

Eddie Woo · 13:39

When to watch: Before Example 1: experimental probability and long-run relative frequency.

https://www.youtube.com/watch?v=xrLrqo7VJT8

BACKUP

Experimental vs. Theoretical Probability Explained!

Mario's Math Tutoring · 4:10

When to watch: Review: experimental versus theoretical probability.

https://www.youtube.com/watch?v=J0ijhF-cdFU

KEY TERMS

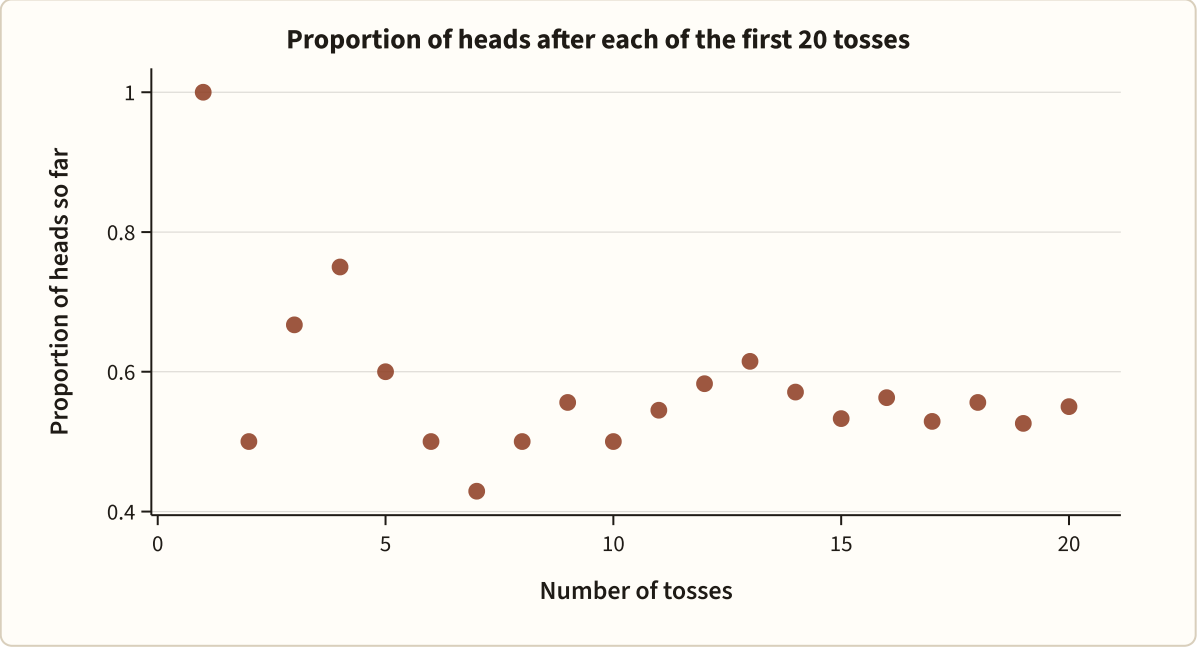
random phenomenon	A process whose single outcomes cannot be predicted, but which shows a regular pattern over many repetitions (a coin toss, whether a seed sprouts, whether a free throw goes in).
probability	The proportion of times an outcome would occur in a very long series of repetitions; a number from 0 (never) to 1 (always).
law of large numbers	As a random process is repeated more and more times, the proportion of times an outcome occurs gets closer and closer to its probability.
simulation	Imitating a chance process with random digits or a random number generator, many times, to estimate a probability.
random digits	A list of the digits 0 to 9 in which each digit is equally likely in every position and has no connection to the digits around it.

trial One complete run of a simulation, from start to the point where you record the result.

CORE IDEAS

Toss a coin once and you cannot say whether it will land heads or tails. Toss it a thousand times, though, and the proportion of heads settles down close to one half. A **random phenomenon** behaves like this: any single outcome is uncertain, but there is a regular pattern in the long run. Tomorrow's weather, whether one seed sprouts, and whether one free throw goes in all behave this way.

The **probability** of an outcome is the proportion of times it would occur in a very long series of repetitions. When a seed company says the probability that a seed sprouts is 0.85, it promises nothing about any one seed; it means that in a very large batch, about 85% of the seeds sprout. A probability is always between 0 and 1.



The graph shows the proportion of heads after each of Samuel's first 20 tosses. At first it jumps around wildly: 1 after one toss, 0.5 after two, 0.75 after four. Each new toss moves the proportion less and less, because it is one toss among more and more. The table follows the same coin much further.

Samuel's coin tosses

Number of tosses	Number of heads	Proportion of heads
10	5	0.5
20	11	0.55
100	54	0.54
500	241	0.482
1000	507	0.507
5000	2489	0.4978

This is the **law of large numbers**: as the number of repetitions grows, the proportion of times an outcome occurs gets closer and closer to its probability. Notice that the *count* of heads does not get close to half the tosses: after 5000 tosses Samuel had 2489 heads, 11 short of 2500. It is the *proportion* that settles down, because 11 out of 5000 is only 0.0022.

The law of large numbers does not work by evening things out. A coin has no memory: after five tails in a row, the probability of heads on the next toss is still one half. Why, then, does the proportion settle down? Suppose the first 10 tosses give 7 heads, 2 more than half. If the next 990 tosses come out exactly half heads (495), the total is $7 + 495 = 502$ heads in 1000 tosses, a proportion of 0.502. The extra 2 heads were never cancelled; they were swamped by many new tosses. Believing that a run of tails makes heads "due" is the **law of averages** myth.

Many probabilities are hard to work out with rules, but we can estimate them by **simulation**. Use four steps. **State the model**: describe the chance process and its probabilities. **Assign digits** so the probabilities match: for a probability of 0.7, let the seven digits 0–6 mean success and 7–9 failure; for 0.37, use two-digit pairs, 00–36 for success and 37–99 for failure. Say what one trial is and what you will record. **Simulate** many trials. **Conclude**: the proportion of trials in which the event happened is the estimate of its probability, stated in context. Why does this work? If the digits are assigned correctly, each simulated trial behaves just like the real process, and the law of large numbers promises that the proportion over many trials comes close to the true probability. Ten trials give a rough idea; a computer running thousands of trials gives an estimate that is usually very close.

KEY FORMULAS

$$\text{probability} \approx \frac{\text{number of times the outcome occurs}}{\text{number of repetitions}} \quad \text{(many repetitions)}$$

$$\text{estimated probability} = \frac{\text{number of trials in which the event happened}}{\text{total number of trials}}$$

From history. The first rules of probability were worked out in 1654 in letters between two French mathematicians, Blaise Pascal and Pierre de Fermat, who were solving a puzzle about how to divide a prize fairly when a game is stopped before it is finished. The Swiss mathematician Jacob Bernoulli proved the first law of large numbers; it appeared in his book Ars Conjectandi (The Art of Conjecturing), published in 1713, after his death.

Practice habit. Before you simulate, write the model in one sentence: what each digit stands for, what one trial is, and what you will record. A clearly described simulation can be checked and repeated by someone else.

WORKED EXAMPLES

Example 1. Henry makes 70% of his free throws, and one shot does not affect the next. In a game he will shoot 4 free throws. Estimate the probability that he makes all 4 by simulation. Let the digits 0–6 stand for a made shot and 7–9 for a miss; one trial is a group of 4 digits. Carry out 10 trials using lines 101 and 102 of the random digit table.

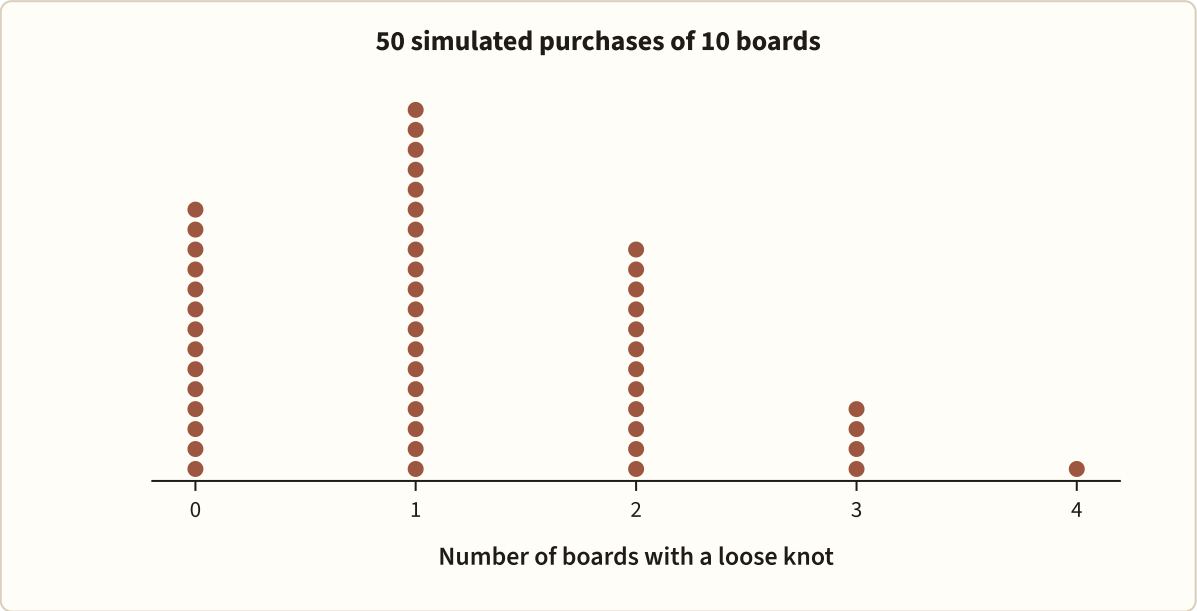
Random digits

Line	Random digits
101	31508 26404 73692 15036
102	24181 95273 04426 10583

1. State the model: each shot is made with probability 0.7, and the shots do not affect one another.
2. Assign digits: the seven digits 0, 1, 2, 3, 4, 5, 6 mean a made shot and the three digits 7, 8, 9 mean a miss, so a made shot has probability $\frac{7}{10} = 0.7$.
3. Simulate: read the digits in groups of 4: 3150, 8264, 0473, 6921, 5036, 2418, 1952, 7304, 4261, 0583.
4. A trial is a success only if it has no 7, 8 or 9. The successes are 3150, 5036 and 4261: 3 of the 10 trials.
5. Conclude: the estimated probability is $\frac{3}{10} = 0.3$.
6. Ten trials give only a rough estimate. A computer simulation of 1000 trials gave 243 trials with all 4 shots made, an estimate of $\frac{243}{1000} = 0.243$. (In Lesson 6.4 you will see that the exact probability is about 0.24.)
7. In context: even a good 70% shooter makes all 4 free throws in only about one game out of four.

Answer: From 10 trials, about $\frac{3}{10} = 0.3$ (a larger simulation gives about 0.24). Henry makes all 4 free throws in only about a quarter of such games.

Example 2. At a sawmill, 12% of pine boards have a loose knot. A carpenter buys 10 boards chosen at random. How likely is it that 3 or more of them have a loose knot? (a) Explain how to use random digits to simulate one purchase of 10 boards, and carry out one trial with the digits 47098 36115 92035 87634. How many knotty boards were in this trial? (b) A computer ran 50 trials; the dotplot shows the number of knotty boards in each trial. Estimate the probability that 3 or more of the 10 boards have a loose knot.



1. Single digits will not work: each digit has probability 0.1, and 0.12 is not a multiple of 0.1. Use two-digit pairs 00 to 99; each of the 100 pairs has probability 0.01.
2. Assign digits: the 12 pairs 00–11 mean a knotty board and the 88 pairs 12–99 mean a good board, so a knotty board has probability $\frac{12}{100} = 0.12$.
3. One trial is 10 pairs (one per board); record the number of knotty boards.
4. The pairs are 47, 09, 83, 61, 15, 92, 03, 58, 76, 34. Only 09 and 03 are in 00–11, so this trial had 2 knotty boards.
5. From the dotplot: 14 trials had 0 knotty boards, 19 had 1, 12 had 2, 4 had 3, and 1 had 4; $14 + 19 + 12 + 4 + 1 = 50$ trials.
6. Trials with 3 or more knotty boards: $4 + 1 = 5$. Estimate: $\frac{5}{50} = 0.10$.
7. Conclude: about 1 purchase in 10 would include 3 or more boards with a loose knot.

Answer: (a) Use pairs 00–11 for a knotty board and 12–99 for a good board; 10 pairs make one trial. The trial 47, 09, 83, 61, 15, 92, 03, 58, 76, 34 has 2 knotty boards (09 and 03). (b) $\frac{5}{50} = 0.10$: about 1 time in 10, a carpenter buying 10 boards would get 3 or more with a loose knot.

PRACTICE

1. EASY A seed company states that the probability a carrot seed sprouts is 0.85. Explain what this probability means for a gardener planting carrot seeds. Does it mean that exactly 85 of every 100 seeds will sprout?
2. EASY Ruth wants to simulate a free throw that is made with probability 0.4 using single random digits. How many of the ten digits 0–9 should stand for a made shot? Give one correct choice of digits.
3. EASY Grace tosses a coin 10 times and gets 8 heads. (a) What proportion of the tosses were heads? (b) Does this show that the coin is unfair? Explain.
4. EASY A coin has landed tails on each of its last 5 tosses. Henry says heads is now "due," so the next toss is more likely to be heads. Is Henry right? Explain.
5. MEDIUM A rancher's cows each have one calf, and each calf is equally likely to be a heifer (female) or a bull (male). To simulate 4 calves, let the digits 0–4 stand for a heifer and 5–9 for a bull. Use line 104 of the table, in groups of 4 digits, to carry out 5 trials. (a) In how many of the 5 trials were there at least 3 heifers? (b) What is the estimated probability of at least 3 heifers among 4 calves?

Random digits

Line	Random digits
104	07215 68342 91570 43386
105	73091 48265 30917 52448
106	52204 92171 38665 01937

6. MEDIUM At a county fair, 37% of the visitors stop at the pie-tasting tent. (a) If you simulate visitors with two-digit random numbers 00–99, how many of the 100 pairs should stand for a visitor who stops at the tent? Give one correct set of pairs. (b) Explain why single digits cannot be used for this simulation.

7. MEDIUM Samuel tosses a thumbtack many times and records how often it lands point up. (a) What is the best estimate from his results of the probability that the tack lands point up? (b) Explain why the proportion after 2000 tosses is a better estimate than the proportion after 20 tosses.

Tossing a thumbtack

Number of tosses	Times point up	Proportion point up
20	13	0.65
100	58	0.58
400	247	0.6175
1000	611	0.611
2000	1224	0.612

8. MEDIUM A coin is tossed 20 times and lands heads 14 times. (a) What is the proportion of heads? (b) Suppose the next 980 tosses come out exactly half heads. What is the proportion of heads after all 1000 tosses? (c) Were the extra heads from the first 20 tosses "made up for"? Explain.

9. MEDIUM To simulate rolling a six-sided die, let the digits 1–6 stand for the faces and skip the digits 0, 7, 8 and 9. Use line 105 of the table to simulate rolling the die until a 6 appears. How many rolls did it take?

Random digits

Line	Random digits
104	07215 68342 91570 43386
105	73091 48265 30917 52448
106	52204 92171 38665 01937

10. HARD A late spring frost kills the blossoms on the Miller family's cherry tree in about 15% of years, and one year's weather does not affect the next. Describe how to use random digits to estimate the probability that the tree loses its blossoms at least once in the next 5 years. Include the model, the digit assignment, what one trial is, what you record, and how you find the estimate.

11. HARD Ruth says she makes 80% of her free throws, but in one game she made only 5 of 10. To see whether that is surprising, a computer simulated 100 games of 10 shots for a shooter who makes each shot with probability 0.8. The dotplot shows the number made in each simulated game (11 games with 10 made, 27 with 9, 30 with 8, 20 with 7, 9 with 6, 2 with 5, 1 with 4). (a) Estimate the probability that an 80% shooter makes 5 or fewer of 10 shots. (b) Does Ruth's game give some reason to doubt her claim? Explain.

12. HARD Samuel and Grace simulate the same chance process to estimate a probability. Samuel runs 20 trials and gets an estimate of 0.35; Grace runs 500 trials and gets 0.27. Which estimate should you trust more? Explain, using the law of large numbers.

13. **CHALLENGE** A feed store puts one of 4 farm-animal cards (cow, horse, pig, sheep) in each bag of chicken feed, each card equally likely. Henry wants all 4 cards. Let the digits 1, 2, 3, 4 stand for the four cards and skip 0 and 5–9. (a) Use line 106 of the table to simulate buying bags until Henry has all four cards. How many bags did this trial take? (b) A computer ran 1000 such trials and found an average of about 8.3 bags. Explain why it usually takes many more than 4 bags.

Random digits

Line	Random digits
104	07215 68342 91570 43386
105	73091 48265 30917 52448
106	52204 92171 38665 01937

COMMON MISTAKES

- Believing the "law of averages": thinking that after a run of tails, heads is more likely. Each toss is new; only the long-run proportion settles down, and it does so by swamping, not by correcting.
- Assigning the wrong number of digits, such as 0–7 for a probability of 0.7 (that is eight digits, or 0.8). Count the digits or pairs you assign and check that they give the stated probability.
- Trusting a simulation with very few trials. Ten trials give a rough guess; run many trials (a computer helps) before drawing a firm conclusion.
- Not saying what one trial is or when it stops. Write down what a trial is, what you record, and which digits to skip before you start reading digits.

ANSWERS TO PRACTICE

1. If a very large number of these carrot seeds were planted, about 85% of them would sprout; 0.85 is the long-run proportion. It does not mean exactly 85 of every 100: one packet of 100 might give 81 or 89 sprouts, because a small batch varies by chance.
2. 4 digits, for example 0, 1, 2, 3 for a made shot and 4–9 for a miss.
3. (a) $\frac{8}{10} = 0.8$. (b) No. With only 10 tosses, the proportion of heads varies a lot by chance, and a fair coin gives 8 heads in 10 tosses now and then. The law of large numbers only says the proportion settles near 0.5 after many tosses; to judge the coin she would need hundreds of tosses.
4. No. The coin has no memory: the probability of heads on the next toss is still 0.5, whatever came before. The law of large numbers works by swamping early runs with many new tosses, not by making up for them.
5. Groups: 0721 (3 heifers), 5683 (1), 4291 (3), 5704 (2), 3386 (2). (a) 2 trials. (b) $\frac{2}{5} = 0.4$, a rough estimate because there are only 5 trials.

6. (a) 37 pairs, for example 00–36. (b) Each single digit has probability 0.1, so single digits can only make probabilities that are multiples of 0.1 (such as 0.3 or 0.4), not 0.37. Each pair has probability 0.01, so 37 pairs give exactly 0.37.
7. (a) About 0.612 (1224 of 2000 tosses). (b) By the law of large numbers, the proportion gets closer to the true probability as the number of tosses grows. After 20 tosses it was still bouncing around (0.65, then 0.58 after 100), but from 400 to 2000 tosses it changed only from 0.6175 to 0.612.
8. (a) $\frac{14}{20} = 0.70$. (b) Heads $= 14 + 490 = 504$, so the proportion is $\frac{504}{1000} = 0.504$. (c) No. There are still 4 more heads than half of 1000 (504 versus 500), the same excess as after 20 tosses. The early excess was not cancelled; it was swamped by many new tosses, so the proportion moved toward 0.5.
9. The digits are 7, 3, 0, 9, 1, 4, 8, 2, 6. Skipping 7, 0, 9 and 8 leaves the rolls 3, 1, 4, 2, 6, so it took 5 rolls.
10. State the model: each year has probability 0.15 of a killing frost, and the years do not affect one another. Assign digits: use two-digit pairs, with 00–14 for a frost year and 15–99 for a year without one. One trial is 5 pairs (5 years); record whether at least one pair is in 00–14. Simulate many trials (100 or more, by computer if possible). Conclude: the estimate is the number of trials with at least one frost year divided by the number of trials.
11. (a) $2 + 1 = 3$ of the 100 games, so about $\frac{3}{100} = 0.03$. (b) Yes, some reason. An 80% shooter makes 5 or fewer only about 3 times in 100 games, so a game this poor is unusual for her. It does not prove she is not an 80% shooter, since unusual games do happen, but it gives some reason to doubt the claim.
12. Grace's estimate of 0.27. By the law of large numbers, the proportion of trials with the event gets closer to the true probability as the number of trials grows. With only 20 trials, chance variation can easily move the proportion far from the truth; 500 trials smooth out most of that variation.
13. (a) The digits are 5, 2, 2, 0, 4, 9, 2, 1, 7, 1, 3. Skipping 5, 0, 9 and 7 leaves the bags 2, 2, 4, 2, 1, 1, 3; the fourth different card (3) comes in bag 7, so this trial took 7 bags. (b) The first bag always gives a new card, but each later bag may repeat a card he already has. Once he has 3 cards, only 1 bag in 4 gives the missing card, so he usually buys several repeats while waiting for the last one.

LESSON 6.2 · 2 SESSIONS OF 45 MINUTES

Probability Rules

OBJECTIVES

- List the sample space of a chance process and find probabilities when outcomes are equally likely.
- Use the basic rules of probability, including the complement rule, to check a probability model and find missing probabilities.
- Apply the addition rule for mutually exclusive events and the general addition rule, and explain why the overlap is subtracted.
- Find probabilities of events from a two-way table and interpret them in context.

VIDEO

WATCH

Probability Fundamentals (2 of 2: The Addition Rule, Complements)

Eddie Woo · 5:11

When to watch: Before Example 1: the addition rule and complements.

https://www.youtube.com/watch?v=aCvtSuFQYnY

BACKUP

Basics of Probability: Unions, Intersections, and Complements

jbstatistics · 15:15

When to watch: Before Example 2: unions, intersections and complements.

https://www.youtube.com/watch?v=B1v9OeCTlu0

KEY TERMS

sample space	The list of all possible outcomes of a chance process, written S.
event	Any outcome or group of outcomes from the sample space, named with a capital letter such as A.
probability model	A description of a chance process that lists the sample space and gives a probability for each outcome.
complement	The complement of event A, written A^c, is the event that A does not happen: all the outcomes not in A.
mutually exclusive	Two events are mutually exclusive (or disjoint) if they have no outcomes in common, so they cannot both happen at once.

Venn diagram A picture of a sample space as a rectangle with each event drawn as a circle inside it; overlapping circles show outcomes that are in both events.

CORE IDEAS

To talk about probability carefully, first list what can happen. The **sample space** S of a chance process is the list of all possible outcomes. If you toss a coin twice, S = {HH, HT, TH, TT}. An **event** is any group of outcomes, such as A = "exactly one head" = {HT, TH}. A **probability model** gives the sample space and a probability for each outcome.

When all outcomes are **equally likely**, probability is counting: $P(A) = \dfrac{\text{number of outcomes in } A}{\text{number of outcomes in } S}$. Roll two dice, say a red one and a blue one. Each die has 6 faces, so there are $6 \times 6 = 36$ equally likely outcomes. The table shows the sum for each of them.

Sum of two dice (red die down the side, blue die across the top)

Red die	Blue 1	Blue 2	Blue 3	Blue 4	Blue 5	Blue 6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

The sum 7 appears in 6 of the 36 cells, so $P(\text{sum} = 7) = \dfrac{6}{36} = 0.1667$. The basic rules follow from the meaning of probability as a long-run proportion. A proportion can never be below 0 or above 1, so $0 \le P(A) \le 1$. Some outcome in S must happen every time, so $P(S) = 1$: the probabilities in a model must add to exactly 1.

Every outcome is either in A or not in A, never both. So A and its **complement** A^c split the whole sample space, and their probabilities add to 1. That gives the **complement rule** $P(A^c) = 1 - P(A)$. The probability that the sum of two dice is not 7 is $1 - \dfrac{6}{36} = \dfrac{30}{36} = 0.8333$, much quicker than counting 30 cells. When an event is messy and its complement is simple, use the complement.

Two events are **mutually exclusive** if they cannot happen together. A sum of 7 and a sum of 11 are mutually exclusive, so their outcomes can simply be added: $P(7 \text{ or } 11) = \dfrac{6}{36} + \dfrac{2}{36} = \dfrac{8}{36} = 0.2222$. In probability, "A or B" means A happens, B happens, or both happen.

If the events can happen together, simple adding goes wrong. Picture a **Venn diagram**: a rectangle for S with two overlapping circles, A and B. Adding P(A) and P(B) counts the outcomes in the overlap twice, once in each circle, so we subtract the overlap once. This is the **general addition rule**. For example, let A = "the red die shows 6" (a row of 6 cells) and B = "the blue die shows 6" (a column of 6 cells). The cell (6, 6) is in both. So $P(A \text{ or } B) = \frac{6}{36} + \frac{6}{36} - \frac{1}{36} = \frac{11}{36} = 0.3056$, and counting the cells in that row and column together gives 11, as it should.

Two-way tables work the same way. Each cell is an "and" (a row category and a column category), the row and column totals give the probabilities of single events, and an "or" probability adds a row total and a column total and subtracts the cell where they meet.

KEY FORMULAS

$$0 \le P(A) \le 1 \qquad P(S) = 1$$

$$P(A) = \frac{\text{number of outcomes in } A}{\text{number of outcomes in } S} \quad \text{(equally likely outcomes)}$$

$$P(A^c) = 1 - P(A)$$

$$P(A \text{ or } B) = P(A) + P(B) \quad \text{(A and B mutually exclusive)}$$

$$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$$

Practice habit. Before you add two probabilities, ask: "Can both of these happen at once?" If they can, subtract the overlap. And whenever a model is given, check that its probabilities add to 1.

WORKED EXAMPLES

Example 1. At the Miller farm, eggs are graded by size. The table gives the probability model for the size of a randomly chosen egg, but the probability for Jumbo is missing. (a) Find P(Jumbo). (b) Find the probability that an egg is Large or bigger. (c) Find the probability that an egg is not Small.

Size of a randomly chosen egg

Egg size	Small	Medium	Large	Extra large	Jumbo
Probability	0.08	0.22	0.45	0.2	?

1. The probabilities must add to 1. The known ones add to $0.08 + 0.22 + 0.45 + 0.20 = 0.95$.
2. (a) $P(\text{Jumbo}) = 1 - 0.95 = 0.05$.
3. (b) An egg has only one size, so the sizes are mutually exclusive and we may add: $P(\text{Large or bigger}) = 0.45 + 0.20 + 0.05 = 0.70$.
4. (c) Complement rule: $P(\text{not Small}) = 1 - 0.08 = 0.92$.
5. In context: 5% of the eggs are Jumbo, 70% are Large or bigger, and 92% are bigger than Small.

Answer: (a) 0.05 (b) 0.70 (c) 0.92

Example 2. The two-way table describes the 120 students at a homeschool co-op. One student is chosen at random to lead the opening of the spring program. (a) Find P(plays a sport). (b) Find P(plays a sport and plays an instrument). (c) Find P(plays a sport or plays an instrument).

Students at a homeschool co-op

	Plays an instrument	No instrument	Total
Plays a sport	18	30	48
No sport	27	45	72
Total	45	75	120

1. (a) The "Plays a sport" row total is 48: $P(\text{sport}) = \dfrac{48}{120} = 0.40$.
2. (b) The cell where the "Plays a sport" row meets the "Plays an instrument" column is 18: $P(\text{sport and instrument}) = \dfrac{18}{120} = 0.15$.
3. (c) The instrument column total is 45: $P(\text{instrument}) = \dfrac{45}{120} = 0.375$.
4. These events can happen together, so use the general addition rule: $P(\text{sport or instrument}) = 0.40 + 0.375 - 0.15 = 0.625$.
5. The 18 students who do both were counted in the sport row and again in the instrument column, which is why 0.15 is subtracted once.
6. Check with the complement: 45 students do neither, so $1 - \dfrac{45}{120} = \dfrac{75}{120} = 0.625$.
7. In context: 62.5% of the co-op students play a sport, an instrument, or both.

Answer: (a) 0.40 (b) 0.15 (c) $0.40 + 0.375 - 0.15 = 0.625$

PRACTICE

1. EASY Henry tosses a coin and then spins a spinner with three equal sections colored red, blue and green. List the sample space. How many outcomes does it have?
2. EASY A box at the garden club holds 40 seed packets: 12 tomato, 10 bean, 8 squash and 10 pepper. Grace picks one packet at random. What is the probability that it is bean or squash?
3. EASY The forecast gives a 35% chance of rain tomorrow. What is the probability that it does not rain tomorrow?

4. EASY A rancher proposes the probability model in the table for the coat color of the next calf born in the herd. Is this a valid probability model? Explain.

Coat color of a calf (proposed model)

Coat color	Black	Red	White
Probability	0.3	0.5	0.25

5. MEDIUM Ruth recorded the first bird to visit her feeder each morning for a whole winter. The table gives the probability model, but the sparrow probability is missing. (a) Find P(sparrow). (b) Find the probability that the first bird is not a chickadee.

First bird to visit the feeder each morning

Bird	Cardinal	Chickadee	Blue jay	Sparrow	Other
Probability	0.22	0.31	0.14	?	0.08

6. MEDIUM Two fair dice are rolled. Use the table of sums. Give each answer to 4 decimal places. (a) Find the probability that the sum is 8. (b) Find the probability that the sum is at least 4.

Sum of two dice (red die down the side, blue die across the top)

Red die	Blue 1	Blue 2	Blue 3	Blue 4	Blue 5	Blue 6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

7. MEDIUM At a county fair, 60% of visitors buy lemonade, 45% buy kettle corn, and 30% buy both. (a) What is the probability that a randomly chosen visitor buys lemonade or kettle corn? (b) What is the probability that the visitor buys neither?

8. MEDIUM The table describes the 200 books in a parish library. One book is chosen at random. (a) Find P(hardcover). (b) Find P(fiction or hardcover).

Books in a parish library

	Hardcover	Paperback	Total
Fiction	38	72	110
Nonfiction	54	36	90
Total	92	108	200

9. MEDIUM One card is drawn at random from a standard deck of 52 cards (13 hearts; 4 kings, one of which is the king of hearts). (a) Are the events "heart" and "king" mutually exclusive? (b) Find P(heart or king) to 4 decimal places.

10. HARD At a farm stand, 70% of customers buy sweet corn and 50% buy tomatoes. Henry says the probability that a customer buys sweet corn or tomatoes is $0.70 + 0.50 = 1.20$. (a) Explain what is wrong with Henry's reasoning. (b) If 35% of customers buy both, find the correct probability.

11. HARD At a Saturday market, 55% of shoppers buy eggs, 40% buy honey, and 70% buy eggs or honey. (a) What is the probability that a shopper buys both eggs and honey? (b) What is the probability that a shopper buys eggs but not honey?

12. HARD Two fair dice are rolled. Find the probability that the roll is doubles (both dice the same) or the sum is 8. Give your answer to 4 decimal places.

Sum of two dice (red die down the side, blue die across the top)

Red die	Blue 1	Blue 2	Blue 3	Blue 4	Blue 5	Blue 6
1	2	3	4	5	6	7
2	3	4	5	6	7	8
3	4	5	6	7	8	9
4	5	6	7	8	9	10
5	6	7	8	9	10	11
6	7	8	9	10	11	12

13. CHALLENGE For a church picnic, 60% of families bring a dessert and 70% bring a salad. (a) Can "brings a dessert" and "brings a salad" be mutually exclusive? Explain. (b) What is the smallest possible probability that a family brings both?

COMMON MISTAKES

- Adding probabilities of events that can happen together. Before adding, ask whether both can happen at once; if so, subtract P(A and B).
- Accepting a probability model whose probabilities do not add to 1. Always add them up first.
- Mixing up "and" with "or." In a two-way table, "and" is a single cell; "or" covers a whole row and a whole column, with the shared cell counted once.
- Counting outcomes that are not equally likely. With two dice, a sum of 7 is not 1 of 11 possible sums; count the 36 equally likely pairs instead.

ANSWERS TO PRACTICE

1. S = {heads-red, heads-blue, heads-green, tails-red, tails-blue, tails-green}: 6 outcomes ($2 \times 3 = 6$).
2. A packet cannot be both, so add: $\frac{10 + 8}{40} = \frac{18}{40} = 0.45$.
3. Complement rule: $1 - 0.35 = 0.65$.
4. No. The probabilities add to $0.30 + 0.50 + 0.25 = 1.05$, but the probabilities of all the outcomes must add to exactly 1.
5. (a) $1 - (0.22 + 0.31 + 0.14 + 0.08) = 1 - 0.75 = 0.25$. (b) $1 - 0.31 = 0.69$.
6. (a) The sum 8 appears in 5 of 36 cells: $\frac{5}{36} = 0.1389$. (b) The complement is a sum of 2 or 3, which appears in $1 + 2 = 3$ cells, so $1 - \frac{3}{36} = \frac{33}{36} = 0.9167$.
7. (a) $0.60 + 0.45 - 0.30 = 0.75$. (b) $1 - 0.75 = 0.25$.
8. (a) $\frac{92}{200} = 0.46$. (b) $\frac{110}{200} + \frac{92}{200} - \frac{38}{200} = \frac{164}{200} = 0.82$. Check: only the 36 nonfiction paperbacks are neither, and $1 - \frac{36}{200} = 0.82$.
9. (a) No: the king of hearts is both. (b) $\frac{13}{52} + \frac{4}{52} - \frac{1}{52} = \frac{16}{52} = 0.3077$.
10. (a) A probability can never be more than 1. Henry added as if the events were mutually exclusive, but a customer can buy both, so the customers who buy both were counted twice. (b) $0.70 + 0.50 - 0.35 = 0.85$.
11. (a) From the general addition rule, $P(\text{both}) = 0.55 + 0.40 - 0.70 = 0.25$. (b) The eggs circle of the Venn diagram is 0.55, and 0.25 of it is the overlap, so $0.55 - 0.25 = 0.30$.
12. Doubles: 6 cells. Sum 8: 5 cells. Both: (4, 4), 1 cell. $\frac{6}{36} + \frac{5}{36} - \frac{1}{36} = \frac{10}{36} = 0.2778$.
13. (a) No. If they were mutually exclusive, P(dessert or salad) would be $0.60 + 0.70 = 1.30$, which is more than 1. (b) $P(\text{both}) = 0.60 + 0.70 - P(\text{dessert or salad})$, and P(dessert or salad) is at most 1, so P(both) is at least $1.30 - 1 = 0.30$.

LESSON 6.3 · 2 SESSIONS OF 45 MINUTES

Conditional Probability

OBJECTIVES

- Compute a conditional probability $P(A \mid B)$ from a two-way table by restricting to the row or column for B.
- Use the formula $P(A \mid B) = \dfrac{P(A \text{ and } B)}{P(B)}$ and explain why it works.
- Apply the general multiplication rule $P(A \text{ and } B) = P(A) \cdot P(B \mid A)$, including drawing without replacement.
- Explain in context why $P(A \mid B)$ is usually not the same as $P(B \mid A)$.

VIDEO

WATCH

Conditional Probability (1 of 4: Establishing foundations)

Eddie Woo · 9:56

When to watch: Before Example 1: what "given" means in P(A | B).

https://www.youtube.com/watch?v=Kf6vht0-Boc

BACKUP

Conditional Probability (5 of 7: Analysis with a table)

Eddie Woo · 7:34

When to watch: After Example 1: conditional probability from a two-way table.

https://www.youtube.com/watch?v=1ClpywTKVTg

KEY TERMS

conditional probability	The probability that event A happens when we know that event B has happened, written $P(A \mid B)$ and read "the probability of A given B."
given	The word for the information we already know; the "given" event is the group we restrict to, and it goes in the denominator.
general multiplication rule	The probability that A and B both happen is the probability of A times the probability of B given that A has happened: $P(A \text{ and } B) = P(A) \cdot P(B \mid A)$.
without replacement	Choosing items one after another without putting them back, so each choice changes what is left for the next one.

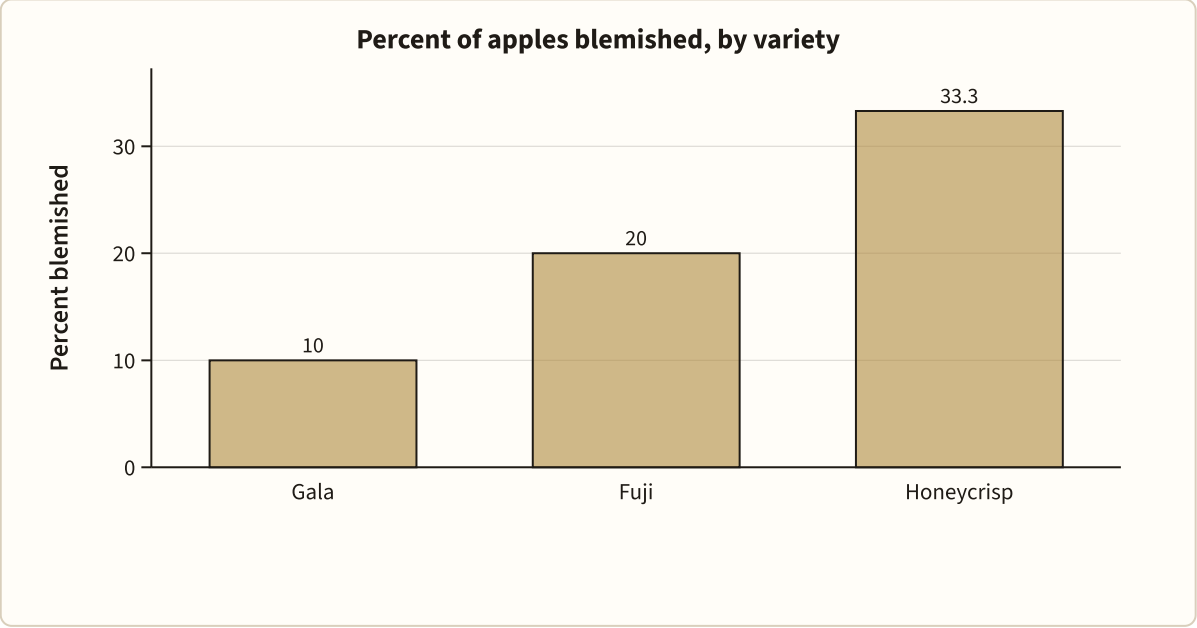
CORE IDEAS

New information can change a probability. If you pick an apple at random from an orchard's harvest, the chance it is blemished is one number; if you are told it is a Honeycrisp, the chance may be different. A **conditional probability** $P(A \mid B)$ is the probability of A when we know B has happened. The bar "|" is read "given."

300 apples from the Miller orchard

Variety	Blemished	Clean	Total
Gala	12	108	120
Fuji	18	72	90
Honeycrisp	30	60	90
Total	60	240	300

The table describes 300 apples sampled from the Miller orchard. Overall, $P(\text{blemished}) = \frac{60}{300} = 0.20$. Now suppose we know the apple is a Honeycrisp. Only the 90 Honeycrisp apples are possible, so we **restrict to the Honeycrisp row**: $P(\text{blemished} \mid \text{Honeycrisp}) = \frac{30}{90} = 0.3333$. In the same way, $P(\text{blemished} \mid \text{Gala}) = \frac{12}{120} = 0.10$ and $P(\text{blemished} \mid \text{Fuji}) = \frac{18}{90} = 0.20$. The bar graph compares these conditional probabilities as percents.



Why does the formula $P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$ work? It does with proportions exactly what we just did with counts. $P(\text{blemished and Honeycrisp}) = \frac{30}{300}$ and $P(\text{Honeycrisp}) = \frac{90}{300}$. Dividing, $\frac{30/300}{90/300} = \frac{30}{90}$

, because the 300s cancel. Dividing by P(B) rescales the B group so that it becomes the whole world, and the formula works even when we have only probabilities and no table.

Order matters. $P(\text{Honeycrisp} \mid \text{blemished})$ restricts to the blemished column: $\frac{30}{60} = 0.50$. Half of the blemished apples are Honeycrisp, but only a third of the Honeycrisp apples are blemished. The two questions use different groups as their denominators, so $P(A \mid B)$ and $P(B \mid A)$ are usually different. Always ask which event is "given."

Multiplying both sides of the formula by the given probability gives the **general multiplication rule**: $P(A \text{ and } B) = P(A) \cdot P(B \mid A)$. In words: for both to happen, A must happen first, and then, out of those times, B must happen too. Suppose a carton holds 10 eggs, 3 of them brown, and Ruth takes two eggs at random **without replacement**. $P(\text{first brown}) = \frac{3}{10}$. Given that the first was brown, 2 brown eggs remain among 9, so $P(\text{second brown} \mid \text{first brown}) = \frac{2}{9}$. Therefore $P(\text{both brown}) = \frac{3}{10} \cdot \frac{2}{9} = \frac{6}{90} = 0.0667$.

KEY FORMULAS

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}$$

$$P(A \text{ and } B) = P(A) \cdot P(B \mid A)$$

Practice habit. Before you compute a conditional probability, say which group you are restricting to. The "given" event is always the denominator.

WORKED EXAMPLES

Example 1. One of the 300 apples in the table is chosen at random. (a) Find the probability that it is clean, given that it is a Gala. (b) Find the probability that it is a Gala, given that it is clean. (c) Explain in context why the answers differ.

300 apples from the Miller orchard

Variety	Blemished	Clean	Total
Gala	12	108	120
Fuji	18	72	90
Honeycrisp	30	60	90
Total	60	240	300

1. (a) Given Gala: restrict to the Gala row, which has 120 apples, 108 of them clean.
2. $P(\text{clean} \mid \text{Gala}) = \dfrac{108}{120} = 0.90$.
3. (b) Given clean: restrict to the Clean column, which has 240 apples, 108 of them Gala.
4. $P(\text{Gala} \mid \text{clean}) = \dfrac{108}{240} = 0.45$.
5. (c) Both use the same 108 apples, but different groups: (a) asks about Gala apples, and (b) asks about clean apples. Gala apples are mostly clean, but clean apples also include many Fuji and Honeycrisp apples.

Answer: (a) 0.90: 90% of Gala apples are clean. (b) 0.45: 45% of clean apples are Galas. (c) The denominators are different groups (the 120 Galas versus the 240 clean apples).

Example 2. The table shows the sections of a church choir. The director draws two names at random, one after the other without replacement, to choose the two Scripture readers for Sunday. Give answers to 4 decimal places. (a) Find the probability that both readers are sopranos. (b) Find the probability that both readers come from the same section.

Members of a church choir

Section	Sopranos	Altos	Tenors	Basses	Total
Number of singers	8	6	4	2	20

1. (a) $P(\text{first soprano}) = \dfrac{8}{20}$. Given that, 7 sopranos remain among 19 singers: $P(\text{second soprano} \mid \text{first soprano}) = \dfrac{7}{19}$.
2. General multiplication rule: $P(\text{both sopranos}) = \dfrac{8}{20} \cdot \dfrac{7}{19} = \dfrac{56}{380} = 0.1474$.
3. (b) "Same section" means both sopranos, both altos, both tenors, or both basses. These are mutually exclusive, so we add them.
4. Altos: $\dfrac{6}{20} \cdot \dfrac{5}{19} = \dfrac{30}{380}$. Tenors: $\dfrac{4}{20} \cdot \dfrac{3}{19} = \dfrac{12}{380}$. Basses: $\dfrac{2}{20} \cdot \dfrac{1}{19} = \dfrac{2}{380}$.
5. $P(\text{same section}) = \dfrac{56 + 30 + 12 + 2}{380} = \dfrac{100}{380} = 0.2632$.
6. In context: about 26% of the time, both readers will come from the same section of the choir.

Answer: (a) $\dfrac{8}{20} \cdot \dfrac{7}{19} = 0.1474$ (b) $\dfrac{100}{380} = 0.2632$

PRACTICE

1. EASY Use the table of 200 spring days. Find the probability of rain later in the day, given that the morning is cloudy.

200 spring days at a weather station

Morning sky	Rain later	No rain	Total
Cloudy	48	32	80
Clear	12	108	120
Total	60	140	200

2. EASY Use the table of 200 spring days. (a) Find P(rain later | clear morning). (b) Find P(rain later) for a randomly chosen day.

200 spring days at a weather station

Morning sky	Rain later	No rain	Total
Cloudy	48	32	80
Clear	12	108	120
Total	60	140	200

3. EASY Use the table of 200 spring days. Find the probability that the morning was cloudy, given that it rained later in the day.

200 spring days at a weather station

Morning sky	Rain later	No rain	Total
Cloudy	48	32	80
Clear	12	108	120
Total	60	140	200

4. EASY On a poultry farm, 30% of the hens are Rhode Island Reds, and 12% of all the hens are Rhode Island Reds that lay more than 5 eggs a week. What is the probability that a hen lays more than 5 eggs a week, given that it is a Rhode Island Red?

5. MEDIUM In Problems 1 and 3 you found P(rain later | cloudy) = 0.6 and P(cloudy | rain later) = 0.8. Explain in context why these two probabilities are different.

200 spring days at a weather station

Morning sky	Rain later	No rain	Total
Cloudy	48	32	80
Clear	12	108	120
Total	60	140	200

6. MEDIUM A basket holds 12 apples, 4 of them bruised. Samuel takes two apples at random without replacement. What is the probability that both are bruised? Give your answer to 4 decimal places.

7. MEDIUM One member of the 4-H club in the table is chosen at random. Give answers to 4 decimal places. (a) Find P(raises rabbits | junior member). (b) Find P(junior member | raises rabbits).

Members of a county 4-H club

Project animal	Junior members	Senior members	Total
Sheep	14	26	40
Rabbits	30	10	40
Poultry	21	19	40
Total	65	55	120

8. MEDIUM 70% of Henry's golf drives land in the fairway. When a drive lands in the fairway, he reaches the green with his next shot 45% of the time. What is the probability that a drive lands in the fairway and he then reaches the green with his next shot?

9. MEDIUM At a bakery, 40% of customers buy bread, and 10% of customers buy both bread and a pie. What is the probability that a customer buys a pie, given that the customer buys bread? Interpret your answer.

10. HARD A shelf in the parish library holds 10 books: 6 fiction and 4 nonfiction. Grace picks 3 books at random without replacement. What is the probability that all three are fiction? Give your answer to 4 decimal places.

11. HARD Use the table of 200 spring days. (a) Find P(no rain | cloudy morning). (b) Find P(no rain) for a randomly chosen day. (c) What does comparing these two numbers tell a farmer deciding whether to cut hay on a cloudy morning?

200 spring days at a weather station

Morning sky	Rain later	No rain	Total
Cloudy	48	32	80
Clear	12	108	120
Total	60	140	200

12. HARD Nearly all professional basketball players are taller than 6 feet. Samuel concludes that a man taller than 6 feet is very likely to be a professional basketball player. Explain his mistake using conditional probability.

13. CHALLENGE Two cards are dealt without replacement from a standard 52-card deck, which has 4 aces. What is the probability that the second card is an ace? (Hint: the first card is either an ace or not; find the probability of each path and add.) Give your answer to 4 decimal places, and compare it with $\frac{4}{52}$.

COMMON MISTAKES

- Dividing by the wrong total. For $P(A \mid B)$ divide by the count (or probability) of B, the given group, not by the grand total or by the count of A.
- Treating $P(A \mid B)$ and $P(B \mid A)$ as the same. Say the given group in words before computing.
- Forgetting that the second draw changes without replacement. After one brown egg is taken, both the number of brown eggs and the total go down by 1.
- Multiplying $P(A) \cdot P(B)$ when B depends on A. Use $P(A) \cdot P(B \mid A)$; the simple product is only correct for independent events (Lesson 6.4).

ANSWERS TO PRACTICE

1. Restrict to the Cloudy row: $\frac{48}{80} = 0.6$.
2. (a) $\frac{12}{120} = 0.1$. (b) $\frac{60}{200} = 0.3$.
3. Restrict to the Rain later column: $\frac{48}{60} = 0.8$.
4. $\frac{0.12}{0.30} = 0.4$
5. They describe different groups of days. The first looks only at the 80 cloudy mornings and asks how many brought rain (48, or 60%). The second looks only at the 60 rainy days and asks how many began cloudy (48, or 80%). Most rainy days start cloudy, but not every cloudy morning brings rain.
6. $\frac{4}{12} \cdot \frac{3}{11} = \frac{12}{132} = 0.0909$
7. (a) $\frac{30}{65} = 0.4615$. (b) $\frac{30}{40} = 0.7500$.
8. $0.70 \times 0.45 = 0.315$
9. $\frac{0.10}{0.40} = 0.25$: one in four bread customers also buys a pie.
10. $\frac{6}{10} \cdot \frac{5}{9} \cdot \frac{4}{8} = \frac{120}{720} = 0.1667$
11. (a) $\frac{32}{80} = 0.4$. (b) $\frac{140}{200} = 0.7$. (c) On a typical day, 70% stay dry, but after a cloudy morning only 40% do. A cloudy morning makes a dry afternoon much less likely, so cutting hay that day is riskier than usual.
12. Samuel has confused P(taller than 6 feet | professional player), which is close to 1, with P(professional player | taller than 6 feet). The second restricts to all men taller than 6 feet, a group of millions, of whom only a few hundred play professionally, so it is tiny. Switching the given event changes the denominator.
13. Ace then ace: $\frac{4}{52} \cdot \frac{3}{51} = \frac{12}{2652}$. Non-ace then ace: $\frac{48}{52} \cdot \frac{4}{51} = \frac{192}{2652}$. Total $\frac{204}{2652} = 0.0769$, exactly $\frac{4}{52} = \frac{1}{13}$. This makes sense: if you do not look at the first card, the second card is equally likely to be any of the 52 cards.

LESSON 6.4 · 2 SESSIONS OF 45 MINUTES

Independence and the Multiplication Rule

OBJECTIVES

- Explain what it means for two events to be independent and check independence from a two-way table.
- Use the multiplication rule $P(A \text{ and } B) = P(A) \cdot P(B)$ for independent events.
- Solve "at least one" problems with the complement, $P(\text{at least one}) = 1 - P(\text{none})$, and explain why this works.
- Explain why mutually exclusive events are not independent.

VIDEO

WATCH

Basic Probability: The Multiplication Rule

jbstatistics · 11:27

When to watch: Before Example 2: the multiplication rule for independent events.

https://www.youtube.com/watch?v=k-rvcvtsV2k

BACKUP

Are mutually exclusive events independent?

jbstatistics · 5:05

When to watch: After Example 1: why mutually exclusive events are not independent.

https://www.youtube.com/watch?v=UOfsad9WWwk

KEY TERMS

independent events	Two events are independent if knowing that one has happened does not change the probability of the other: $P(B \mid A) = P(B)$.
multiplication rule for independent events	If A and B are independent, the probability that both happen is $P(A \text{ and } B) = P(A) \cdot P(B)$.
at least one	One or more of the events happen; its complement is "none of them happen."
dependent events	Events that are not independent: knowing one has happened changes the probability of the other.

CORE IDEAS

Sometimes knowing one event tells you nothing about another. Toss a coin twice: knowing the first toss landed heads does not change the chance that the second lands heads, because the coin has no memory. Two events are **independent** if knowing that one has happened does not change the probability of the other: $P(B \mid A) = P(B)$.

To check independence from data, compare a conditional probability with the overall probability. The table describes 200 students at a music and sports camp.

200 students at a summer camp

	Plays piano	Does not play piano	Total
Left-handed	5	15	20
Right-handed	45	135	180
Total	50	150	200

Overall, $P(\text{piano}) = \dfrac{50}{200} = 0.25$. Among left-handed students, $P(\text{piano} \mid \text{left}) = \dfrac{5}{20} = 0.25$, and among right-handed students, $P(\text{piano} \mid \text{right}) = \dfrac{45}{180} = 0.25$. Knowing a student's handedness does not change the chance that the student plays piano, so these events are independent. Compare the apples of Lesson 6.3: $P(\text{blemished}) = 0.20$ but $P(\text{blemished} \mid \text{Honeycrisp}) = 0.3333$, so variety and blemishes are not independent. (With real sample data, conditional percents are rarely exactly equal; in Unit 13 you will learn how to decide whether a difference is too large to be chance.)

For independent events the general multiplication rule becomes simpler. Since $P(B \mid A) = P(B)$, the rule $P(A \text{ and } B) = P(A) \cdot P(B \mid A)$ turns into $P(A \text{ and } B) = P(A) \cdot P(B)$. Check it in the camp table: $P(\text{left and piano}) = 0.10 \times 0.25 = 0.025$, and indeed $\dfrac{5}{200} = 0.025$. This rule finishes Henry's free-throw problem from Lesson 6.1: if each shot is made with probability 0.7, independently, then $P(\text{all 4 made}) = 0.7 \times 0.7 \times 0.7 \times 0.7 = 0.2401$, very close to the simulation's 0.243.

“At least one” questions are where the complement rule shines. Suppose a gardener plants 3 seeds that each sprout with probability 0.8, independently. “At least one sprouts” includes many cases: exactly one, exactly two, or all three. Its complement, “none sprout,” is a single case, and independence lets us multiply: $P(\text{none}) = 0.2 \times 0.2 \times 0.2 = 0.008$. So $P(\text{at least one}) = 1 - 0.008 = 0.992$. Do not add $0.8 + 0.8 + 0.8 = 2.4$: the events overlap, and a probability cannot exceed 1.

Mutually exclusive and independent sound alike but are almost opposites. Mutually exclusive events cannot happen together. So if one of them happens, the other's probability drops to 0, and knowing one changes the other completely. Two mutually exclusive events (each with positive probability) are therefore never independent. Independence is about information: knowing A tells you nothing about B.

Independence is often a reasonable model for separate physical processes, such as two coins, two different seeds, or two wells on different land. It is not automatic. Rain on Saturday and rain on Sunday are not independent, because weather systems last for days, and drawing without replacement makes later draws depend on earlier ones. Always ask whether one event could affect the other before you multiply.

KEY FORMULAS

$$A \text{ and } B \text{ are independent if } P(B \mid A) = P(B)$$

$$P(A \text{ and } B) = P(A) \cdot P(B) \quad \text{(independent events)}$$

$$P(\text{at least one}) = 1 - P(\text{none})$$

Practice habit. For any "at least one" question, write $P(\text{at least one}) = 1 - P(\text{none})$ first, then find P(none). And before multiplying probabilities, ask whether the events are really independent.

WORKED EXAMPLES

Example 1. The table shows 200 tomato plants in a community garden, by watering method and whether the plant developed blight. One plant is chosen at random. (a) Find P(blight | drip line). (b) Find P(blight | sprinkler). (c) Are "gets blight" and "watering method" independent? Explain.

Tomato plants in a community garden

Watering	Blight	No blight	Total
Drip line	9	81	90
Sprinkler	33	77	110
Total	42	158	200

1. (a) Restrict to the drip-line row: $P(\text{blight} \mid \text{drip}) = \dfrac{9}{90} = 0.10$.
2. (b) Restrict to the sprinkler row: $P(\text{blight} \mid \text{sprinkler}) = \dfrac{33}{110} = 0.30$.
3. Overall, $P(\text{blight}) = \dfrac{42}{200} = 0.21$.
4. (c) Knowing the watering method changes the probability of blight (0.10 or 0.30 instead of 0.21), so the events are not independent.
5. In context: in this garden, sprinkler-watered plants were three times as likely to get blight as drip-watered plants. This is an observational comparison, so it shows an association; an experiment would be needed to show that sprinklers cause blight.

Answer: (a) 0.10 (b) 0.30 (c) Not independent: the chance of blight depends on the watering method (10% versus 30%, compared with 21% overall).

Example 2. A packet says 85% of its bean seeds sprout. Ruth plants 4 seeds in a pot, and the seeds sprout independently. Give answers to 4 decimal places. (a) Find the probability that all 4 sprout. (b) Find the probability that at least one sprouts. (c) Find the probability that at least one fails to sprout.

1. (a) Multiplication rule for independent events: $P(\text{all } 4) = 0.85^4 = 0.85 \times 0.85 \times 0.85 \times 0.85 = 0.52200625 \approx 0.5220$.
2. (b) The complement of "at least one sprouts" is "none sprout." Each seed fails with probability $1 - 0.85 = 0.15$.
3. $P(\text{none sprout}) = 0.15^4 = 0.00050625$.
4. $P(\text{at least one sprouts}) = 1 - 0.00050625 = 0.99949375 \approx 0.9995$.
5. (c) The complement of "at least one fails" is "all 4 sprout," found in (a): $1 - 0.52200625 = 0.47799375 \approx 0.4780$.
6. In context: Ruth is almost certain to get at least one bean plant, but nearly half the time at least one of her 4 seeds will not come up.

Answer: (a) $0.85^4 = 0.5220$ (b) $1 - 0.15^4 = 0.9995$ (c) $1 - 0.85^4 = 0.4780$

PRACTICE

1. EASY Samuel tosses a fair coin and rolls a fair die. What is the probability of heads and a 6? Give your answer to 4 decimal places.
2. EASY A farm has two wells on opposite sides of the property. On a given day, each pump fails with probability 0.05, independently of the other. What is the probability that both pumps fail on the same day?
3. EASY A calf is born. Let A = "the calf is a heifer" and B = "the calf is a bull." These events are mutually exclusive. Are they independent? Explain.
4. EASY Grace makes 70% of her free throws, and her shots are independent. What is the probability that she makes all 3 of her next free throws?
5. MEDIUM The table shows 200 bread orders at a bakery. (a) Find P(large loaf). (b) Find P(large loaf | white bread). (c) Are loaf size and bread type independent?

Bread orders at a bakery

Bread	Small loaf	Large loaf	Total
White	48	72	120
Wheat	32	48	80
Total	80	120	200

6. MEDIUM The table shows Henry's last 60 rounds of golf. (a) Find P(score under 80). (b) Find P(score under 80 | windy). (c) Are “score under 80” and “windy” independent? Explain in context.

Henry's last 60 rounds of golf

Weather	Score under 80	Score 80 or more	Total
Windy	6	18	24
Calm	21	15	36
Total	27	33	60

7. MEDIUM A weather station has 3 separate rain gauges. On any day, each gauge fails with probability 0.02, independently of the others. What is the probability that at least one gauge fails on a given day? Give your answer to 4 decimal places.

8. MEDIUM Samuel plants 5 squash seeds, each of which sprouts with probability 0.9, independently. What is the probability that at least one seed fails to sprout? Give your answer to 4 decimal places.

9. MEDIUM Ruth says, “If I toss a coin twice, the probability of at least one head is $\frac{1}{2} + \frac{1}{2} = 1$, so I am sure to get a head.” (a) Explain her mistake. (b) Find the correct probability.

10. HARD The forecast gives a 40% chance of rain on Saturday and a 40% chance of rain on Sunday. Henry says the probability of rain on both days is $0.4 \times 0.4 = 0.16$. Explain why Henry's calculation may be wrong.

11. HARD Grace has two hens. On a given day, the first lays an egg with probability 0.8 and the second with probability 0.7, independently. What is the probability that exactly one of the two hens lays an egg?

12. HARD Ruth goes to the Saturday market with probability 0.4, and Henry, who lives in another town, goes with probability 0.5, independently. Find the probability that at least one of them goes, in two ways: with the general addition rule and with the complement.

13. CHALLENGE Old seed sprouts with probability 0.6, independently. Samuel will drop several seeds in each hole. What is the smallest number of seeds per hole that makes the probability that at least one sprouts at least 0.99?

COMMON MISTAKES

- Adding probabilities for “at least one” ($0.8 + 0.8 + 0.8 = 2.4$). Use $1 - P(\text{none})$ instead.
- Thinking mutually exclusive means independent. If two events (each with positive probability) cannot happen together, knowing one happened tells you the other did not, so they are dependent.
- Multiplying probabilities without checking independence. Draws without replacement and weather on nearby days are usually not independent.

- Using the wrong complement. The complement of “at least one sprouts” is “none sprout,” not “at least one fails.”

ANSWERS TO PRACTICE

1. The coin and die are independent: $\frac{1}{2} \times \frac{1}{6} = \frac{1}{12} = 0.0833$.
2. $0.05 \times 0.05 = 0.0025$
3. Not independent. If you know the calf is a heifer, the probability that it is a bull drops to 0, so knowing A changes the probability of B.
4. $0.7 \times 0.7 \times 0.7 = 0.343$
5. (a) $\frac{120}{200} = 0.6$. (b) $\frac{72}{120} = 0.6$. (c) Independent: knowing the bread is white does not change the chance of a large loaf (and for wheat, $\frac{48}{80} = 0.6$ as well).
6. (a) $\frac{27}{60} = 0.45$. (b) $\frac{6}{24} = 0.25$. (c) Not independent: Henry broke 80 in 45% of all rounds but in only 25% of windy rounds, so knowing it is windy lowers his chance of a good score.
7. $P(\text{none fail}) = 0.98^3 = 0.941192$, so $P(\text{at least one fails}) = 1 - 0.941192 = 0.0588$.
8. The complement is “all 5 sprout”: $0.9^5 = 0.59049$. So $1 - 0.59049 = 0.4095$.
9. (a) Heads on the first toss and heads on the second toss are not mutually exclusive (both can happen), so their probabilities cannot simply be added; she counted HH twice. (b) $P(\text{no heads}) = 0.5 \times 0.5 = 0.25$, so $P(\text{at least one head}) = 1 - 0.25 = 0.75$.
10. Multiplying assumes rain on Saturday and rain on Sunday are independent. They probably are not: a storm system that brings rain on Saturday often lasts into Sunday, so P(rain Sunday | rain Saturday) is likely more than 0.4, and the true chance of rain on both days is likely more than 0.16.
11. First only: $0.8 \times 0.3 = 0.24$. Second only: $0.2 \times 0.7 = 0.14$. These are mutually exclusive, so add: $0.24 + 0.14 = 0.38$.
12. Addition rule: $0.4 + 0.5 - 0.4 \times 0.5 = 0.9 - 0.2 = 0.7$. Complement: $1 - 0.6 \times 0.5 = 1 - 0.3 = 0.7$.
13. We need $1 - 0.4^n \geq 0.99$, that is, $0.4^n \leq 0.01$. $0.4^5 = 0.01024$, which is too large; $0.4^6 = 0.004096$, which works. So 6 seeds per hole.

LESSON 6.5 · 2 SESSIONS OF 45 MINUTES

Tree Diagrams

OBJECTIVES

- Organize a multi-stage chance process with a tree diagram, labeling each branch with its (conditional) probability.
- Find the probability of an event by multiplying along branches and adding the paths that end in the event.
- Find a "reverse" conditional probability from a tree and interpret it in context.
- Explain why a test for a rare condition can give many false alarms.

VIDEO

WATCH

Probability Trees (1 of 2: How to construct them)

Eddie Woo · 7:36

When to watch: Before Example 1: drawing a probability tree.

https://www.youtube.com/watch?v=wZh822QhiKk

BACKUP

Probability Trees (2 of 2: How to read them)

Eddie Woo · 6:39

When to watch: Before Example 2: reading a tree, including working backward.

https://www.youtube.com/watch?v=9ofKRuCze_o

KEY TERMS

tree diagram	A diagram for a chance process that happens in stages: branches from a starting point show the first-stage outcomes, and branches from the end of each of those show the second-stage outcomes.
branch probability	The probability written on a branch. Second-stage branches carry conditional probabilities: the probability of that outcome given the branch it grows from.
path	A route through the tree from the start to the end of one branch at the last stage; each path is one complete outcome of the process.
false alarm	A test result that signals a problem (or a success) that is not really there, such as a viable seed that fails a quick test.

CORE IDEAS

Many chance processes happen in stages: first the weather, then the forecast; first which field a pumpkin came from, then whether it is blemished. A **tree diagram** organizes them. Draw a starting point with one branch for each first-stage outcome, labeled with its probability. From the end of each branch, draw branches for the second-stage outcomes, labeled with conditional probabilities: the probability of that outcome given the first-stage branch. The branches leaving any point must add to 1.

Here is an example. A farmer knows that 90% of the seeds in an old lot are viable (alive and able to sprout). A quick soaking test calls 95% of viable seeds "good," but it also calls 20% of non-viable seeds "good." The tree starts with two branches, Viable 0.90 and Not viable 0.10. From Viable grow Good 0.95 and Bad 0.05; from Not viable grow Good 0.20 and Bad 0.80. The table lists the four paths.

The four paths of the seed-test tree

Seed (first stage)	Probability	Test result (second stage)	Conditional probability	Path probability
Viable	0.9	Good	0.95	0.855
Viable	0.9	Bad	0.05	0.045
Not viable	0.1	Good	0.2	0.02
Not viable	0.1	Bad	0.8	0.08

Multiply along the branches to get the probability of a path. This is the general multiplication rule from Lesson 6.3: $P(\text{viable and good}) = P(\text{viable}) \cdot P(\text{good} \mid \text{viable}) = 0.90 \times 0.95 = 0.855$. The paths are mutually exclusive (each seed follows exactly one) and together they cover everything, so the four path probabilities add to 1. To find the probability of an event, **add the paths that end in it**: $P(\text{test says good}) = 0.855 + 0.020 = 0.875$.

Now the reverse question: if the test says good, how likely is it that the seed is really viable? Restrict to the paths that end in Good, which have total probability 0.875, and ask what part of that came through Viable: $P(\text{viable} \mid \text{good}) = \dfrac{0.855}{0.875} = 0.9771$. This is just $P(A \mid B) = \dfrac{P(A \text{ and } B)}{P(B)}$ read off the tree.

It helps to imagine 1000 seeds. The probabilities become counts, and the tree becomes a two-way table like those in Lesson 6.3. Of the 875 seeds the test calls good, 855 are viable, and $\dfrac{855}{875} = 0.9771$, as before.

What happens to 1000 seeds

	Test says good	Test says bad	Total
Viable	855	45	900
Not viable	20	80	100
Total	875	125	1000

Look at the Bad column. $P(\text{not viable} \mid \text{bad}) = \dfrac{80}{125} = 0.64$, so 36% of the seeds the test rejects are actually viable. Why so many? Viable seeds are so common (900 of 1000) that even the small 5% of them that fail the test (45 seeds) is a large group next to the 80 truly dead seeds. When one group is much larger than the other, its small error rate can produce many **false alarms**. The base rate (how common each first-stage outcome is) matters as much as the accuracy of the test.

KEY FORMULAS

$$P(\text{path}) = \text{product of the probabilities along its branches}$$

$$P(\text{event}) = \text{sum of the probabilities of the paths that end in the event}$$

$$P(\text{first-stage outcome} \mid \text{second-stage outcome}) = \frac{P(\text{the path through both})}{P(\text{all paths ending in that second-stage outcome})}$$

From history. The English minister Thomas Bayes studied how to reason backward from an observed result to its likely cause. His essay on the problem was published in 1763, after his death, by his friend Richard Price. The “reverse” tree calculations in this lesson carry out the same reasoning.

Practice habit. Check every tree two ways: the branches leaving each point add to 1, and all the path probabilities add to 1. If either check fails, find the error before going on.

WORKED EXAMPLES

Example 1. A family pumpkin patch grows 60% of its pumpkins in Field A and 40% in Field B. Field B is shadier, and the table gives the probability that a pumpkin from each field is blemished. A pumpkin is picked at random from the harvest. (a) Find the probability that it is blemished. (b) Given that it is blemished, find the probability that it came from Field B, to 4 decimal places.

Branches of the pumpkin tree

Field (first stage)	P(field)	Pumpkin (second stage)	P(second stage | field)
Field A	0.6	Blemished	0.05
Field A	0.6	Not blemished	0.95
Field B	0.4	Blemished	0.15
Field B	0.4	Not blemished	0.85

1. Tree: first branches Field A 0.60 and Field B 0.40; from each, Blemished and Not blemished with the conditional probabilities in the table.
2. Multiply along the paths that end in Blemished: Field A, $0.60 \times 0.05 = 0.03$; Field B, $0.40 \times 0.15 = 0.06$.
3. (a) Add those paths: $P(\text{blemished}) = 0.03 + 0.06 = 0.09$.
4. (b) Restrict to the blemished paths (total 0.09). The part through Field B is 0.06.
5. $P(\text{Field B} \mid \text{blemished}) = \dfrac{0.06}{0.09} = 0.6667$.
6. In context: 9% of the pumpkins are blemished, and although Field B grows only 40% of the pumpkins, it produces about two-thirds of the blemished ones.

Answer: (a) $0.60 \times 0.05 + 0.40 \times 0.15 = 0.09$ (b) $\dfrac{0.06}{0.09} = 0.6667$: about two-thirds of the blemished pumpkins come from Field B.

Example 2. In a valley town, it rains on 20% of spring days. The local forecast predicts rain on 80% of the days when rain is coming, and on 10% of the days when no rain is coming. Give answers to 4 decimal places. (a) Find the probability that the forecast predicts rain. (b) The forecast predicts rain for tomorrow. What is the probability that it actually rains? (c) The forecast predicts no rain. What is the probability that it rains anyway? Interpret (b) and (c).

Branches of the forecast tree

Weather (first stage)	P(weather)	Forecast (second stage)	P(forecast | weather)
Rain	0.2	Rain forecast	0.8
Rain	0.2	No-rain forecast	0.2
No rain	0.8	Rain forecast	0.1
No rain	0.8	No-rain forecast	0.9

1. Paths: Rain and rain forecast: $0.20 \times 0.80 = 0.16$. Rain and no-rain forecast: $0.20 \times 0.20 = 0.04$. No rain and rain forecast: $0.80 \times 0.10 = 0.08$. No rain and no-rain forecast: $0.80 \times 0.90 = 0.72$. Check: $0.16 + 0.04 + 0.08 + 0.72 = 1$.
2. (a) Add the paths ending in a rain forecast: $0.16 + 0.08 = 0.24$.
3. (b) $P(\text{rain} \mid \text{rain forecast}) = \dfrac{0.16}{0.24} = 0.6667$.
4. (c) The no-rain forecast paths total $0.04 + 0.72 = 0.76$, so $P(\text{rain} \mid \text{no-rain forecast}) = \dfrac{0.04}{0.76} = 0.0526$.
5. Check with 1000 days: 200 rainy days give 160 rain forecasts; 800 dry days give 80. Of the 240 rain forecasts, 160 are right: $\dfrac{160}{240} = 0.6667$.
6. In context: when this forecast says rain, it rains about 2 times in 3; when it says no rain, it rains only about 5 times in 100. A no-rain forecast is more trustworthy than a rain forecast, because dry days are so common that their 10% false alarms add up.

Answer: (a) 0.2400 (b) $\dfrac{0.16}{0.24} = 0.6667$ (c) $\dfrac{0.04}{0.76} = 0.0526$. A rain forecast is right about two-thirds of the time; after a no-rain forecast it rains only about 5% of the time.

PRACTICE

1. EASY At a bakery, 70% of customers order coffee. Of the customers who order coffee, 40% also order a muffin. What is the probability that a customer orders coffee and a muffin?
2. EASY Samuel's baseball team plays 40% of its games at home. It wins 70% of its home games and 45% of its away games. (a) Find P(home game and win). (b) Find P(away game and win).

3. EASY In a tree for a farmer's season, the first stage has the branches “plant early” with probability 0.35 and “plant late.” From “plant early,” the branches are “good harvest” with probability 0.8 and “poor harvest.” (a) What is the probability on the “plant late” branch? (b) What is the probability on the “poor harvest” branch that grows from “plant early”?

4. EASY In Problem 2, explain why you may add the probabilities of the paths “home and win” and “away and win” to get the probability that the team wins a game.

5. MEDIUM A carton holds 12 eggs, 4 brown and 8 white. Ruth takes two eggs at random without replacement. What is the probability that exactly one of them is brown? Give your answer to 4 decimal places.

6. MEDIUM Use the information in Problem 2. (a) Find the probability that Samuel's team wins a game. (b) Given that the team won, find the probability that the game was at home, to 4 decimal places.

7. MEDIUM In Henry's town in summer, if one day is sunny, the next day is sunny with probability 0.7; if a day is not sunny, the next day is sunny with probability 0.4. This morning's forecast gives a 60% chance that today will be sunny. What is the probability that tomorrow is sunny?

8. MEDIUM The table describes the apple harvest at a family orchard. An apple is chosen at random from the whole harvest. What is the probability that it is bruised?

Apple harvest at a family orchard

Variety	Share of the crop	Percent of that variety bruised
Gala	50%	4%
Fuji	30%	6%
Honeycrisp	20%	10%

9. MEDIUM Use the orchard table. Given that a randomly chosen apple is bruised, what is the probability that it is a Honeycrisp? Give your answer to 4 decimal places and interpret it.

Apple harvest at a family orchard

Variety	Share of the crop	Percent of that variety bruised
Gala	50%	4%
Fuji	30%	6%
Honeycrisp	20%	10%

10. HARD At a sawmill, 15% of boards are too wet to sell. A moisture meter flags 90% of the wet boards and also flags 5% of the dry boards. A board is chosen at random. Give answers to 4 decimal places. (a) Find the probability that the meter flags the board. (b) Given that the meter flags the board, find the probability that it is really wet.

11. HARD In a large orchard, 2% of the trees carry a root fungus. A soil test is positive for 95% of the trees with the fungus and for 5% of the healthy trees. (a) A tree tests positive. What is the probability that it has the fungus? Give your answer to 4 decimal places. (b) Explain, by imagining 1000 trees, why the answer is so much lower than 95%.

12. HARD Grace makes 70% of her free throws, independently. In a "one-and-one" she shoots one free throw; only if she makes it does she get a second shot. Each made shot is worth 1 point. Find the probability that she scores (a) 0 points, (b) 1 point, (c) 2 points.

13. CHALLENGE Henry's first serve in tennis goes in 60% of the time, and he wins 75% of the points when it goes in. If the first serve misses, his second serve goes in 90% of the time, and he wins 50% of the points when the second serve goes in. If both serves miss, he loses the point. (a) Find the probability that Henry wins a point on his serve. (b) Given that he won the point, find the probability that his first serve went in, to 4 decimal places.

COMMON MISTAKES

- Putting unconditional probabilities on second-stage branches. Each second-stage branch is conditional on the branch it grows from: P(good | viable), not P(good).
- Confusing a test's accuracy with the reverse probability. P(good | viable) = 0.95 is not P(viable | good); find the reverse by dividing one path by the total of the paths that end in "good."
- Forgetting a path. For P(blemished), add every path that ends in blemished; for P(win) in tennis, include the second-serve path.
- Ignoring how common each first-stage outcome is. A small error rate in a large group can produce more false alarms than true results; imagine 1000 cases to see it.

ANSWERS TO PRACTICE

1. Multiply along the branches: $0.70 \times 0.40 = 0.28$.
2. (a) $0.40 \times 0.70 = 0.28$. (b) $0.60 \times 0.45 = 0.27$.
3. The branches leaving any point add to 1. (a) $1 - 0.35 = 0.65$. (b) $1 - 0.8 = 0.2$.
4. Each game is either a home game or an away game, never both, so the two paths are mutually exclusive: no win is counted twice. Together they include every way the team can win, so their sum is P(win).
5. Brown then white: $\frac{4}{12} \cdot \frac{8}{11} = \frac{32}{132}$. White then brown: $\frac{8}{12} \cdot \frac{4}{11} = \frac{32}{132}$. Total $\frac{64}{132} = 0.4848$.
6. (a) $0.28 + 0.27 = 0.55$. (b) $\frac{0.28}{0.55} = 0.5091$.
7. $0.6 \times 0.7 + 0.4 \times 0.4 = 0.42 + 0.16 = 0.58$

8. $0.50 \times 0.04 + 0.30 \times 0.06 + 0.20 \times 0.10 = 0.020 + 0.018 + 0.020 = 0.058$
9. $\dfrac{0.020}{0.058} = 0.3448$. Honeycrisp apples are only 20% of the crop but about 34% of the bruised apples.
10. (a) $0.15 \times 0.90 + 0.85 \times 0.05 = 0.135 + 0.0425 = 0.1775$. (b) $\dfrac{0.135}{0.1775} = 0.7606$.
11. (a) $P(\text{positive}) = 0.02 \times 0.95 + 0.98 \times 0.05 = 0.019 + 0.049 = 0.068$, so $P(\text{fungus} \mid \text{positive}) = \dfrac{0.019}{0.068} = 0.2794$. (b) Of 1000 trees, only 20 have the fungus, and 19 of them test positive. But 980 are healthy, and 5% of them, 49 trees, also test positive. So of the 68 positive trees, only 19 really have the fungus: false alarms outnumber true results because healthy trees are so common.
12. (a) She misses the first: 0.3. (b) She makes the first and misses the second: $0.7 \times 0.3 = 0.21$. (c) She makes both: $0.7 \times 0.7 = 0.49$. Check: $0.3 + 0.21 + 0.49 = 1$.
13. (a) First serve in and win: $0.6 \times 0.75 = 0.45$. First miss, second in, win: $0.4 \times 0.9 \times 0.5 = 0.18$. $P(\text{win}) = 0.45 + 0.18 = 0.63$. (b) $\dfrac{0.45}{0.63} = 0.7143$.

LESSON 6.6 · 2 SESSIONS OF 45 MINUTES

Counting: Permutations and Combinations

OBJECTIVES

- Use the multiplication counting principle and factorials to count outcomes.
- Decide whether order matters, and count with permutations ${}_nP_r$ or combinations $\binom{n}{r}$.
- Explain why $\binom{n}{r} = \frac{{}_nP_r}{r!}$ and why each entry of Pascal's triangle is the sum of two entries above it.
- Compute probabilities of random selections by counting equally likely outcomes.

VIDEO

WATCH

How to Use Permutations and Combinations

Mario's Math Tutoring · 6:58

When to watch: Before Example 1: permutations versus combinations.

https://www.youtube.com/watch?v=NEGxh_D7yKU

BACKUP

Permutations and Combinations (1 of 2: Outline of the various expressions)

Eddie Woo · 7:14

When to watch: Review: the notation nPr and nCr.

https://www.youtube.com/watch?v=z4ykeFZSpwg

KEY TERMS

multiplication counting principle	If one choice can be made in m ways and, for each of them, a second choice can be made in n ways, then the two choices together can be made in $m \times n$ ways (and so on for more choices).
factorial	For a whole number n, $n! = n \times (n-1) \times \cdots \times 2 \times 1$, the number of ways to arrange n different items in order; by agreement $0! = 1$.
permutation	An arrangement of r items chosen from n in which order matters; the number of them is written ${}_nP_r$.
combination	A selection of r items chosen from n in which order does not matter; the number of them is written $\binom{n}{r}$ or ${}_nC_r$ and read “n choose r.”

Pascal's triangle	A triangle of numbers with 1s along the edges in which each entry is the sum of the two entries above it; row n lists $\binom{n}{0}, \binom{n}{1}, \ldots, \binom{n}{n}$.

CORE IDEAS

When outcomes are equally likely, probability is counting (Lesson 6.2), and some counts are too large to list. The **multiplication counting principle** handles choices made in steps: multiply the number of options at each step. A church supper offers 3 soups and 4 breads, so there are $3 \times 4 = 12$ soup-and-bread meals. Why multiply? Each soup can be paired with each of the 4 breads, so the choices form a 3-by-4 grid of meals, just as two dice form a 6-by-6 grid of 36 outcomes.

How many orders are there for singing 5 hymns? Any of the 5 can go first, then any of the 4 left, and so on: $5 \times 4 \times 3 \times 2 \times 1 = 120$. This product is **5 factorial**, written $5!$. In general, $n!$ counts the ways to arrange n different items in a row. We agree that $0! = 1$, which keeps the formulas below working.

Often we arrange only some of the items. A choir of 8 singers needs one soloist for each of verses 1, 2 and 3. There are 8 choices for verse 1, then 7, then 6: $8 \times 7 \times 6 = 336$. Here **order matters**, since Grace on verse 1 and Henry on verse 2 is different from the reverse. Such an arrangement is a **permutation**, and ${}_nP_r = n(n-1)\cdots(n-r+1) = \dfrac{n!}{(n-r)!}$; here ${}_8P_3 = \dfrac{8!}{5!} = 336$.

Now suppose the director just wants a committee of 3 to plan the Christmas concert. Order no longer matters: Grace, Henry and Ruth make the same committee in any order. The 336 ordered lists count each committee several times, and the table shows exactly how many: the 3 people of each committee can be ordered in $3! = 6$ ways.

Each committee of 3 appears 3! = 6 times among the ordered lists

Committee (order does not matter)	Ordered lists of the same 3 people (order matters)
Grace, Henry, Ruth	GHR, GRH, HGR, HRG, RGH, RHG
Grace, Henry, Samuel	GHS, GSH, HGS, HSG, SGH, SHG

So the number of committees is $\dfrac{336}{6} = 56$. Dividing by $r!$ removes the orderings we do not care about. That is why the number of **combinations** is $\binom{n}{r} = \dfrac{{}_nP_r}{r!} = \dfrac{n!}{r!\,(n-r)!}$. Counting also gives probabilities: if 3 of the 8 singers (5 sopranos and 3 altos) are chosen at random, every committee is equally likely, so $P(\text{all sopranos}) = \dfrac{\binom{5}{3}}{\binom{8}{3}} = \dfrac{10}{56} = 0.1786$.

Pascal's triangle: the entry in row n, column r is n choose r

n	r = 0	r = 1	r = 2	r = 3	r = 4	r = 5	r = 6
0	1						
1	1	1					
2	1	2	1				
3	1	3	3	1			
4	1	4	6	4	1		
5	1	5	10	10	5	1	
6	1	6	15	20	15	6	1

The values of $\binom{n}{r}$ form **Pascal's triangle**. In the table, each entry is the sum of the entry just above it and the entry above and to the left; for example $\binom{5}{2} = 4 + 6 = 10$. Why? To choose 2 of 5 people, fix attention on one of them, say Grace. Either Grace is chosen, and we pick 1 more from the other 4 ($\binom{4}{1} = 4$ ways), or she is not, and we pick both from the other 4 ($\binom{4}{2} = 6$ ways). The rows are also symmetric, $\binom{n}{r} = \binom{n}{n-r}$, because choosing who is on a committee is the same as choosing who is left off. These numbers return in Unit 8, where they count the arrangements of successes and failures.

KEY FORMULAS

$$n! = n \times (n-1) \times \cdots \times 2 \times 1, \qquad 0! = 1$$

$${}_nP_r = n(n-1)\cdots(n-r+1) = \frac{n!}{(n-r)!} \quad \text{(order matters)}$$

$$\binom{n}{r} = {}_nC_r = \frac{{}_nP_r}{r!} = \frac{n!}{r!\,(n-r)!} \quad \text{(order does not matter)}$$

$$\binom{n}{r} = \binom{n-1}{r-1} + \binom{n-1}{r}$$

From history. The triangle of numbers is named for Blaise Pascal, who wrote a treatise on the “arithmetic triangle” in 1654 and used it in his work on probability. The triangle itself is much older: mathematicians in China, Persia and India knew it centuries before Pascal.

Practice habit. Before you count, ask: “Does order matter?” If swapping two chosen people gives a different result (first and second place, verse 1 and verse 2), use permutations; if it does not (a committee, a handful of seeds), use combinations.

WORKED EXAMPLES

Example 1. A church choir has 10 singers. (a) In how many ways can the director choose 3 singers to sing the solos in verses 1, 2 and 3 of a hymn? (b) In how many ways can the director choose a committee of 3 singers to plan the spring concert?

1. (a) Order matters: a different singer on verse 1 gives a different arrangement. Use a permutation.
2. ${}_{10}P_3 = 10 \times 9 \times 8 = 720$.
3. (b) Order does not matter: the same 3 people form the same committee in any order. Use a combination.
4. Each committee of 3 appears $3! = 6$ times among the 720 ordered lists, so $\binom{10}{3} = \dfrac{720}{6} = 120$.
5. Check with the formula: $\binom{10}{3} = \dfrac{10!}{3!\,7!} = \dfrac{10 \times 9 \times 8}{3 \times 2 \times 1} = 120$.
6. In context: there are 720 ways to assign the three solos but only 120 different planning committees.

Answer: (a) ${}_{10}P_3 = 720$ ways (b) $\binom{10}{3} = 120$ committees

Example 2. A 4-H club has 12 members: 7 juniors and 5 seniors. A committee of 4 is chosen at random to plan the fair booth. Give answers to 4 decimal places. (a) Find the probability that the committee has exactly 2 juniors and 2 seniors. (b) Find the probability that all 4 members are juniors.

1. Order does not matter, and every committee of 4 is equally likely. Total number of committees: $\binom{12}{4} = \dfrac{12 \times 11 \times 10 \times 9}{4 \times 3 \times 2 \times 1} = \dfrac{11880}{24} = 495$.
2. (a) Ways to choose 2 of the 7 juniors: $\binom{7}{2} = \dfrac{7 \times 6}{2} = 21$. Ways to choose 2 of the 5 seniors: $\binom{5}{2} = \dfrac{5 \times 4}{2} = 10$.
3. By the multiplication counting principle, each pair of juniors can go with each pair of seniors: $21 \times 10 = 210$ committees.
4. $P(\text{2 juniors and 2 seniors}) = \dfrac{210}{495} = 0.4242$.
5. (b) Ways to choose 4 of the 7 juniors: $\binom{7}{4} = \dfrac{7 \times 6 \times 5 \times 4}{24} = 35$.
6. $P(\text{all juniors}) = \dfrac{35}{495} = 0.0707$.
7. In context: a random committee is split evenly between juniors and seniors about 42% of the time, and is all juniors only about 7% of the time.

Answer: (a) $\dfrac{21 \times 10}{495} = 0.4242$ (b) $\dfrac{35}{495} = 0.0707$

PRACTICE

1. EASY The ice cream stand at the county fair offers the flavors, holders and toppings in the table. A treat is one flavor in one holder with one topping. How many different treats are possible?

Ice cream stand at the county fair

Flavors	Holders	Toppings
Vanilla	Sugar cone	Sprinkles
Chocolate	Waffle cone	Chopped nuts
Strawberry	Cup	
Peach		
Butter pecan		

2. EASY Compute (a) 6! and (b) $\dfrac{7!}{5!}$.
3. EASY The pastor has chosen 5 hymns for Sunday. In how many different orders can they be sung?

4. EASY For each situation, say whether order matters and whether you would count with permutations or combinations. (a) Choosing 3 apples from a basket to bake in a pie. (b) Awarding first, second and third place ribbons to 3 of the pies at the fair.

5. MEDIUM Nine pies are entered in the county fair contest. In how many ways can the judges award first, second and third place?

6. MEDIUM Ruth has 9 books she would like to read and room in her bag for 4 on a family trip. How many different sets of 4 books can she pack?

7. MEDIUM Use Pascal's triangle in the table. (a) What is $\binom{6}{2}$? (b) Use row 6 to write row 7, and give $\binom{7}{3}$.

Pascal's triangle: the entry in row n, column r is n choose r

n	r = 0	r = 1	r = 2	r = 3	r = 4	r = 5	r = 6
0	1						
1	1	1					
2	1	2	1				
3	1	3	3	1			
4	1	4	6	4	1		
5	1	5	10	10	5	1	
6	1	6	15	20	15	6	1

8. MEDIUM The lock on the Millers' barn uses a 4-digit code, each digit 0 to 9. (a) How many codes are possible if digits may repeat? (b) How many are possible if no digit may repeat?

9. MEDIUM Five ushers serve at a church. (a) In how many ways can a head usher and an assistant usher be chosen? (b) In how many ways can 2 ushers be chosen to count the offering together? (c) Explain why the answer to (b) is exactly half the answer to (a).

10. HARD A drawer holds 10 seed packets: 4 flower and 6 vegetable. Grace grabs 3 packets at random. Give answers to 4 decimal places. (a) Find the probability that all 3 are vegetable packets. (b) Find the probability that at least one is a flower packet.

11. HARD A basketball coach chooses 5 starters at random from 10 players, who include the twins Samuel and Henry. What is the probability that both twins start? Give your answer to 4 decimal places.

12. HARD Explain why $\binom{6}{2} = \binom{5}{1} + \binom{5}{2}$ by thinking about choosing 2 of 6 cousins, one of whom is Grace, to carry the cake to the family picnic. Check that the numbers agree.

13. **CHALLENGE** A choir has 12 members. The director first chooses a 3-person committee to plan a concert, and then, from the 9 singers not on the committee, chooses a soloist for the first song and a different soloist for the second song. In how many ways can all of this be done?

COMMON MISTAKES

- Using permutations when order does not matter (or combinations when it does). Ask whether swapping two of the chosen items changes the result.
- Adding the choices at each step instead of multiplying. For a soup and a bread, each soup pairs with every bread: 3×4, not $3 + 4$.
- Multiplying out huge factorials. Cancel first: $\dfrac{10!}{7!} = 10 \times 9 \times 8$.
- Counting the favorable outcomes one way (in order) and the total another way (not in order) in a probability. Use combinations for both, or permutations for both.

ANSWERS TO PRACTICE

1. $5 \times 3 \times 2 = 30$ treats.
2. (a) $6 \times 5 \times 4 \times 3 \times 2 \times 1 = 720$. (b) $\dfrac{7!}{5!} = 7 \times 6 = 42$ (the rest cancels).
3. $5! = 120$ orders.
4. (a) Order does not matter (the same 3 apples make the same pie): combinations. (b) Order matters (first place differs from second): permutations.
5. ${}_9P_3 = 9 \times 8 \times 7 = 504$ ways.
6. $\binom{9}{4} = \dfrac{9 \times 8 \times 7 \times 6}{4 \times 3 \times 2 \times 1} = \dfrac{3024}{24} = 126$ sets.
7. (a) 15. (b) Row 7 is 1, 7, 21, 35, 35, 21, 7, 1 (for example $\binom{7}{3} = \binom{6}{2} + \binom{6}{3} = 15 + 20$), so $\binom{7}{3} = 35$.
8. (a) $10 \times 10 \times 10 \times 10 = 10000$. (b) $10 \times 9 \times 8 \times 7 = 5040$.
9. (a) ${}_5P_2 = 5 \times 4 = 20$. (b) $\binom{5}{2} = \dfrac{20}{2} = 10$. (c) Each pair of ushers appears twice in (a), once with each as head usher, but only once in (b), where the two jobs are the same. So we divide by $2! = 2$.
10. (a) $\dfrac{\binom{6}{3}}{\binom{10}{3}} = \dfrac{20}{120} = 0.1667$. (b) Complement: $1 - 0.1667 = 0.8333$.
11. Total: $\binom{10}{5} = 252$. If both twins start, the other 3 starters come from the other 8 players: $\binom{8}{3} = 56$. $P = \dfrac{56}{252} = 0.2222$.
12. Every pair either includes Grace or does not. Pairs with Grace: choose her partner from the other 5 cousins, $\binom{5}{1} = 5$ ways. Pairs without Grace: choose both from the other 5, $\binom{5}{2} = 10$ ways. No pair is counted twice and none is missed, so $\binom{6}{2} = 5 + 10 = 15$, which matches the triangle.

13. Committee: $\binom{12}{3} = \dfrac{12 \times 11 \times 10}{6} = 220$. Soloists in order from the other 9: ${}_9P_2 = 9 \times 8 = 72$. Multiply: $220 \times 72 = 15840$ ways.

Unit 6 test

You have 55 minutes. You may use a scientific calculator; no tables are needed. Show your work: write the rule you use, the numbers you substitute, and the result, and answer in a sentence in context when asked to interpret or explain. Keep full precision until the last step, and give probabilities to 4 decimal places unless the problem says otherwise.

1. **6 PTS** A pitcher throws a strike on 60% of her pitches, and one pitch does not affect the next. To simulate pitches, let the digits 0–5 stand for a strike and 6–9 for a ball. (a) Explain why this digit assignment matches the pitcher. (b) Use line 201 of the table, in groups of 5 digits, to simulate two sets of 5 pitches. How many strikes are in each set?

   Random digits

Line	Random digits
201	48170 92685

2. **6 PTS** A baseball player who gets a hit in 30% of his at-bats has gone hitless in his last 12 at-bats. The announcer says, “He's due for a hit.” Explain what is wrong with the announcer's reasoning, and state what the law of large numbers really says about the player's hits.

3. **6 PTS** The table gives a probability model for the rainfall on an April day in a farming town, but the probability for Moderate is missing. (a) Find P(Moderate). (b) Find the probability that there is some rain (Light, Moderate or Heavy).

   Rainfall on an April day in a farming town

Rainfall	None	Light	Moderate	Heavy
Probability	0.45	0.3	?	0.07

4. **6 PTS** One card is drawn at random from a standard 52-card deck, which has 12 face cards (jacks, queens and kings) and 13 clubs, including 3 face cards that are clubs. Find the probability that the card is a face card or a club.

5. **8 PTS** The table describes the 180 students in a library's summer reading program. One student is chosen at random. (a) Find P(finished the reading goal). (b) Find P(finished | middle school). (c) Find P(middle school | finished).

Students in a library summer reading program

	Finished reading goal	Did not finish	Total
Middle school	63	27	90
High school	54	36	90
Total	117	63	180

6. **6 PTS** Use the reading program table. Are the events “finished the reading goal” and “middle school student” independent? Justify your answer with probabilities and explain in context.

Students in a library summer reading program

	Finished reading goal	Did not finish	Total
Middle school	63	27	90
High school	54	36	90
Total	117	63	180

7. **6 PTS** A tray at the church bake sale holds 15 cookies, 5 of them oatmeal. Henry picks two cookies at random without replacement. What is the probability that both are oatmeal?

8. **7 PTS** A farmer runs 4 irrigation pumps that work independently. On a very hot day, each pump breaks down with probability 0.08. (a) Find the probability that no pump breaks down. (b) Find the probability that at least one pump breaks down.

9. **7 PTS** At a summer camp, each camper chooses exactly one activity for the first period. The probability that a randomly chosen camper chooses canoeing is 0.3, and the probability that the camper chooses archery is 0.4. (a) Are “chooses canoeing” and “chooses archery” mutually exclusive? (b) Are they independent? Explain. (c) Find P(canoeing or archery).

10. **8 PTS** A bakery bakes its loaves in two ovens; the table shows each oven's share of the loaves and the percent of its loaves that come out underweight. A loaf is chosen at random. (a) Find the probability that it is underweight. (b) Given that the loaf is underweight, find the probability that it came from Oven B. Interpret your answer.

Loaves from a bakery's two ovens

Oven	Share of all loaves	Percent of that oven's loaves underweight
Oven A	70%	3%
Oven B	30%	8%

11. **8 PTS** On a poultry farm, 80% of the eggs set in the incubator are fertile. Candling (shining a light through the shell) marks 90% of fertile eggs as fertile, but it also marks 15% of infertile eggs as fertile. An egg is chosen at random. (a) Find the probability that candling marks it as fertile. (b) Given that candling marks the egg as fertile, find the probability that it really is fertile. (c) Given that candling marks the egg as not fertile, find the probability that it really is infertile.

12. **6 PTS** Eight runners are in the final of a cross-country race. (a) In how many ways can first, second and third place be awarded? (b) The top 3 runners will go on to the state meet together. How many different groups of 3 could go?

13. **7 PTS** A beekeeper has 12 jars of honey on a shelf, and 3 of them are slightly underfilled. An inspector chooses 4 jars at random. (a) Find the probability that none of the 4 jars is underfilled. (b) Find the probability that at least one is underfilled.

14. **7 PTS** A service-project team has 7 volunteers. (a) In how many ways can the team choose a leader, a treasurer and a secretary (three different people)? (b) In how many ways can it choose a 3-person crew to paint the fence? (c) Explain why the answer to (b) is the answer to (a) divided by 6.

15. **6 PTS** Each hive in an apiary has probability 0.1 of swarming in May, independently of the other hives. What is the largest number of hives the beekeeper can keep and still have a probability of at least 0.5 that no hive swarms in May? Show your reasoning.

UNIT 6 TEST: ANSWER KEY

1. (a) Six of the ten equally likely digits (0, 1, 2, 3, 4, 5) stand for a strike, so a strike has probability $\frac{6}{10} = 0.6$, and the digits are independent like the pitches. (b) First set 4, 8, 1, 7, 0: strikes 4, 1, 0, so 3 strikes. Second set 9, 2, 6, 8, 5: strikes 2, 5, so 2 strikes.
2. Each at-bat is a new chance process; a hitless streak does not raise the probability of a hit on the next at-bat (it is still about 0.3). The law of large numbers says only that over a very large number of at-bats, the proportion of hits gets close to 0.3. It does this by swamping streaks with many new at-bats, not by making up for them.
3. (a) $1 - (0.45 + 0.30 + 0.07) = 1 - 0.82 = 0.18$. (b) Complement of None: $1 - 0.45 = 0.55$.
4. The events overlap, so use the general addition rule: $\frac{12}{52} + \frac{13}{52} - \frac{3}{52} = \frac{22}{52} = 0.4231$.
5. (a) $\frac{117}{180} = 0.65$. (b) $\frac{63}{90} = 0.70$. (c) $\frac{63}{117} = 0.5385$.
6. Not independent. $P(\text{finished}) = 0.65$, but $P(\text{finished} \mid \text{middle school}) = 0.70$ (and $P(\text{finished} \mid \text{high school}) = \frac{54}{90} = 0.60$). Knowing a student's level changes the chance that the student finished: middle schoolers in this program finished more often.
7. $\frac{5}{15} \cdot \frac{4}{14} = \frac{20}{210} = 0.0952$
8. (a) $0.92^4 = 0.71639296 \approx 0.7164$. (b) $1 - 0.71639296 = 0.2836$.
9. (a) Yes: a camper chooses only one activity. (b) Not independent: if a camper chose canoeing, the probability of archery drops from 0.4 to 0, so knowing one event changes the other. (c) Mutually exclusive, so add: $0.3 + 0.4 = 0.7$.
10. (a) $0.70 \times 0.03 + 0.30 \times 0.08 = 0.021 + 0.024 = 0.045$. (b) $\frac{0.024}{0.045} = 0.5333$. Oven B bakes only 30% of the loaves but produces about 53% of the underweight ones, so it is the oven to check first.
11. Paths: fertile and marked fertile $0.8 \times 0.9 = 0.72$; fertile and marked not $0.8 \times 0.1 = 0.08$; infertile and marked fertile $0.2 \times 0.15 = 0.03$; infertile and marked not $0.2 \times 0.85 = 0.17$. (a) $0.72 + 0.03 = 0.75$. (b) $\frac{0.72}{0.75} = 0.96$. (c) $\frac{0.17}{0.08 + 0.17} = \frac{0.17}{0.25} = 0.68$.
12. (a) Order matters: ${}_8P_3 = 8 \times 7 \times 6 = 336$. (b) Order does not matter: $\binom{8}{3} = \frac{336}{3!} = 56$.
13. (a) $\frac{\binom{9}{4}}{\binom{12}{4}} = \frac{126}{495} = 0.2545$. (b) $1 - \frac{126}{495} = 0.7455$.
14. (a) ${}_7P_3 = 7 \times 6 \times 5 = 210$. (b) $\binom{7}{3} = \frac{210}{6} = 35$. (c) The same 3 volunteers can be given the three offices in $3! = 6$ different ways, so each painting crew appears 6 times among the 210 ordered choices; dividing by 6 counts each crew once.
15. We need $0.9^n \geq 0.5$. $0.9^6 = 0.531441$, which works; $0.9^7 = 0.4782969$, which is too small. So at most 6 hives.

UNIT 7

Random Variables

In Unit 6 you learned to find the probability of an event. In this unit you attach numbers to chance: the number of eggs in a nest box, the profit of a fundraiser booth, the total weight of a crate of melons. A random variable is a numerical outcome of a chance process, and its probability distribution tells how likely each value is. You will learn to find a random variable's mean (its long-run average, called the expected value) and its standard deviation (how much it typically varies), and to use them to make sensible decisions about insurance, crops and games.

The unit then asks what happens when random variables are changed or combined. Converting units or turning a count into a profit changes the mean and spread in simple ways. When independent quantities are added or subtracted, the means add or subtract, but it is the variances, not the standard deviations, that add. Finally, sums and differences of independent normal random variables are themselves normal, which lets you answer practical questions such as whether a board will fit a gap.

These ideas are the foundation for the rest of the course. Unit 8 studies two important random variables, the binomial and the geometric. Units 9 through 13 treat a sample mean or sample proportion as a random variable, and the rule that variances add is exactly why averages vary less than single values and why a difference of two sample means has the spread it does.

PREREQUISITES

- Basic probability rules, complements and the addition rule (Lesson 6.2)
- Independence and the multiplication rule (Lesson 6.4)
- Mean and standard deviation of data, including why deviations are squared (Lessons 2.1 and 2.4)
- Effects of adding a constant and multiplying by a constant on center and spread (Lesson 2.6)
- Density curves and finding normal proportions with z-scores (Lessons 3.2 and 3.4)
- Finding values from normal proportions (Lesson 3.5)

MASTERY CHECKLIST

☐ I can identify a random variable and tell whether it is discrete or continuous.
☐ I can write a probability distribution as a table, check that it is valid, and draw a probability histogram.
☐ I can find probabilities such as $P(X \geq 3)$ from a distribution, using the complement when it is quicker.
☐ I can compute the expected value of a discrete random variable and explain it as a long-run average.
☐ I can compute the variance and standard deviation of a discrete random variable and interpret the standard deviation.
☐ I can use expected value to decide whether a game is fair and to compare choices such as insurance or crops.

☐ I can find the mean and standard deviation of $a + bX$.
☐ I can find the mean and standard deviation of a sum or difference of independent random variables, and explain why variances add.
☐ I can explain why a single value of a continuous random variable has probability 0.
☐ I can find probabilities for sums and differences of independent normal random variables and interpret them in context.

LESSON 7.1 · 2 SESSIONS OF 45 MINUTES

Discrete Random Variables

OBJECTIVES

- Identify a random variable in a chance process and classify it as discrete or continuous.
- Build the probability distribution of a discrete random variable as a table and check that it is valid.
- Draw and read a probability histogram and describe its shape in context.
- Compute probabilities of events such as $P(X \geq 3)$ or $P(2 < X \leq 4)$ by adding probabilities or by using the complement.

VIDEO

WATCH

Discrete Random Variables (1 of 3: Introduction)

Eddie Woo · 13:06

When to watch: Before Example 1: what a discrete random variable is.

https://www.youtube.com/watch?v=Lkom2Q13SaQ

BACKUP

Discrete Random Variables (2 of 3: Notation)

Eddie Woo · 8:06

When to watch: Before Example 2: the notation P(X = x).

https://www.youtube.com/watch?v=yCTTqzot3pE

KEY TERMS

Random variable	A variable whose value is a numerical outcome of a chance process, such as the number of heads in three coin tosses. It is named with a capital letter like X; a particular value is written with a small letter like x.
Probability distribution	A list of the possible values of a random variable together with the probability of each value. For a discrete random variable it is usually written as a table.
Discrete random variable	A random variable that takes a list of separate values, often counts such as 0, 1, 2, 3, each with its own probability.
Continuous random variable	A random variable that can take any value in an interval, such as a weight or a time. Its probabilities are areas under a density curve.

Probability histogram	A graph of a discrete probability distribution with a bar for each value whose height is the probability of that value.

CORE IDEAS

Many chance processes produce a number: the number of eggs in a nest box this morning, the number of hits a batter gets in a game, the sum of two dice. A **random variable** is a variable whose value is a numerical outcome of a chance process. We name random variables with capital letters such as X and write a particular value with a small letter x. Before the chance process happens we cannot say what X will be, but we can say how likely each value is, and that is the whole point.

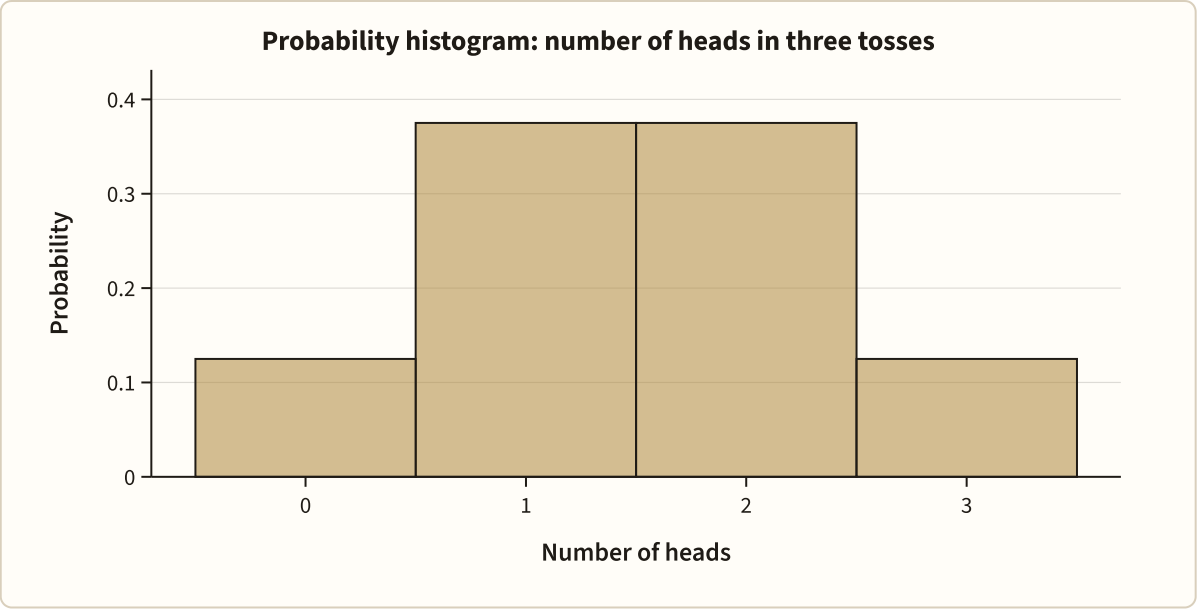
Toss three fair coins and let X = the number of heads. There are 8 equally likely outcomes: HHH, HHT, HTH, THH, HTT, THT, TTH, TTT. One outcome has 0 heads, three have 1 head, three have 2 heads and one has 3 heads, so the probabilities are $\frac{1}{8}, \frac{3}{8}, \frac{3}{8}, \frac{1}{8}$. A table of the values and their probabilities is called the **probability distribution** of X.

Number of heads in three tosses of a fair coin

Number of heads, x	P(x)
0	0.125
1	0.375
2	0.375
3	0.125

Every probability distribution of a discrete random variable obeys two rules: each probability is between 0 and 1, and the probabilities add to exactly 1. Why must they add to 1? The values in the table are mutually exclusive (X cannot equal 1 and 2 at the same time), and together they cover every possible outcome. By the addition rule from Unit 6, their probabilities add to $P(S) = 1$. If the probabilities in a table add to 0.95 or 1.10, a value is missing or a probability is wrong.

A **probability histogram** pictures a distribution: one bar for each value, with height equal to its probability. Because each bar has width 1, the area of each bar equals its probability and the total area is 1. It looks like a relative frequency histogram for a good reason: probability is long-run relative frequency. If you tossed three coins thousands of times and graphed the relative frequency of each number of heads, the graph would look almost exactly like this one.



To find the probability of an event, add the probabilities of the values that make up the event: $P(X \geq 2) = P(X = 2) + P(X = 3) = 0.375 + 0.125 = 0.5$. Read the inequality carefully. "At least 2" ($X \geq 2$) includes 2, but "more than 2" ($X > 2$) does not, so $P(X > 2) = 0.125$. For a discrete random variable these give different answers. When an event contains many values, the complement rule is often quicker: $P(X \geq 1) = 1 - P(X = 0) = 1 - 0.125 = 0.875$.

A **discrete random variable** takes a list of separate values (usually counts), and we can give each value its own probability. A **continuous random variable**, such as the weight of a randomly chosen melon or the exact time until a train arrives, can take any value in an interval. Its values cannot be listed one by one, so its probabilities are areas under a density curve, just as in Unit 3. We return to continuous random variables in Lesson 7.4.

KEY FORMULAS

$$0 \leq P(x) \leq 1 \quad \text{for every value } x, \qquad \sum P(x) = 1$$

$$P(\text{event}) = \text{sum of } P(x) \text{ for the values } x \text{ in the event}$$

$$P(X \geq 1) = 1 - P(X = 0)$$

Practice habit. Before you use any probability distribution, add up its probabilities and check that the total is exactly 1. Then underline the words "at least", "at most", "more than" and "fewer than" and write the event with the correct symbol ($\geq$, $\leq$, $>$, $<$) before you add anything.

WORKED EXAMPLES

Example 1. Records on the Miller family farm suggest that X = the number of eggs a certain hen lays in a week has the probability distribution shown. (a) Find $P(X \geq 6)$. (b) Find $P(X < 5)$. (c) Find $P(4 \leq X \leq 6)$.

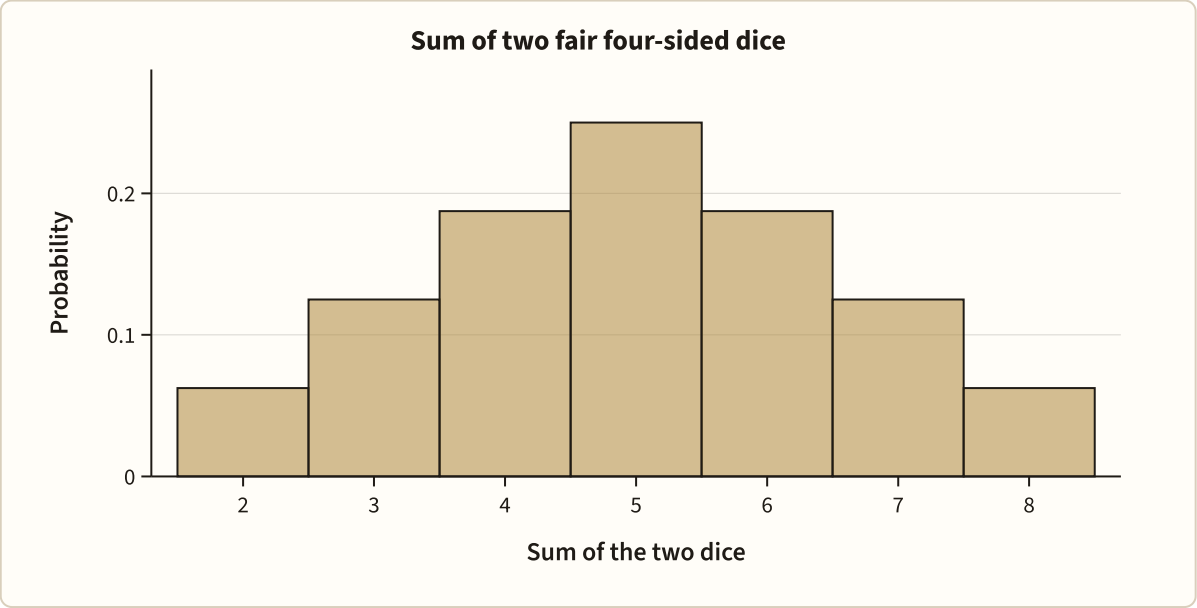
Eggs laid by one hen in a week

Eggs, x	P(x)
3	0.05
4	0.1
5	0.25
6	0.4
7	0.2

1. Check the distribution first: $0.05 + 0.10 + 0.25 + 0.40 + 0.20 = 1.00$, and every probability is between 0 and 1, so the distribution is valid.
2. (a) "At least 6" means 6 or 7: $P(X \geq 6) = P(6) + P(7) = 0.40 + 0.20 = 0.60$.
3. (b) "Fewer than 5" means 3 or 4 (5 is not included): $P(X < 5) = 0.05 + 0.10 = 0.15$.
4. (c) Both ends are included, so the values are 4, 5 and 6: $P(4 \leq X \leq 6) = 0.10 + 0.25 + 0.40 = 0.75$.
5. Conclusion: in about 60% of weeks this hen lays 6 or more eggs, in about 15% of weeks she lays fewer than 5, and in about 75% of weeks she lays between 4 and 6 eggs.

Answer: (a) 0.60 (b) 0.15 (c) 0.75

Example 2. A board game uses two fair four-sided dice, each numbered 1, 2, 3, 4. Let X = the sum of the two dice. (a) Find the probability distribution of X. (b) Find $P(X \geq 6)$. (c) Find $P(X > 6)$.



1. Each die has 4 equally likely faces, so there are $4 \times 4 = 16$ equally likely ordered outcomes (first die, second die).
2. Count the outcomes for each sum. Sum 2: (1,1), 1 outcome. Sum 3: (1,2), (2,1), 2 outcomes. Sum 4: (1,3), (2,2), (3,1), 3 outcomes. Sum 5: (1,4), (2,3), (3,2), (4,1), 4 outcomes. Sum 6: (2,4), (3,3), (4,2), 3 outcomes. Sum 7: (3,4), (4,3), 2 outcomes. Sum 8: (4,4), 1 outcome.
3. Check the count: $1 + 2 + 3 + 4 + 3 + 2 + 1 = 16$.
4. Divide each count by 16: $P(2) = 0.0625$, $P(3) = 0.125$, $P(4) = 0.1875$, $P(5) = 0.25$, $P(6) = 0.1875$, $P(7) = 0.125$, $P(8) = 0.0625$. The probability histogram is symmetric about 5.
5. (b) $P(X \geq 6) = P(6) + P(7) + P(8) = \dfrac{3 + 2 + 1}{16} = \dfrac{6}{16} = 0.375$.
6. (c) "More than 6" leaves out 6: $P(X > 6) = \dfrac{2 + 1}{16} = \dfrac{3}{16} = 0.1875$.
7. Conclusion: in the long run, a roll of the two dice gives a sum of 6 or more about 37.5% of the time and a sum greater than 6 about 18.75% of the time.

Answer: (a) x = 2, 3, 4, 5, 6, 7, 8 with P(x) = 0.0625, 0.125, 0.1875, 0.25, 0.1875, 0.125, 0.0625 (b) 0.375 (c) 0.1875

PRACTICE

1. EASY Classify each random variable as discrete or continuous. (a) X = the number of lambs born on a farm this spring. (b) Y = the exact amount of rain, in inches, that falls on the farm in April.

2. EASY Let X = the number of books a visitor checks out of the parish library on one visit. The table gives the distribution, but one probability is missing. Find $P(X = 3)$.

Books checked out from the parish library per visit

Books, x	P(x)
0	0.2
1	0.35
2	0.25
3	?
4	0.05

3. EASY Using the parish library distribution (with your answer to Problem 2), find the probability that a visitor checks out at least 2 books.

Books checked out from the parish library per visit

Books, x	P(x)
0	0.2
1	0.35
2	0.25
3	?
4	0.05

4. EASY Let X = the number of hits a softball player gets in a game, with the distribution shown. Find the probability that she gets at least one hit. Use the complement rule.

Hits by a softball player in one game

Hits, x	P(x)
0	0.3
1	0.4
2	0.2
3	0.08
4	0.02

5. MEDIUM For the softball player in Problem 4, find (a) $P(X > 1)$ and (b) $P(1 \leq X \leq 3)$.

Hits by a softball player in one game

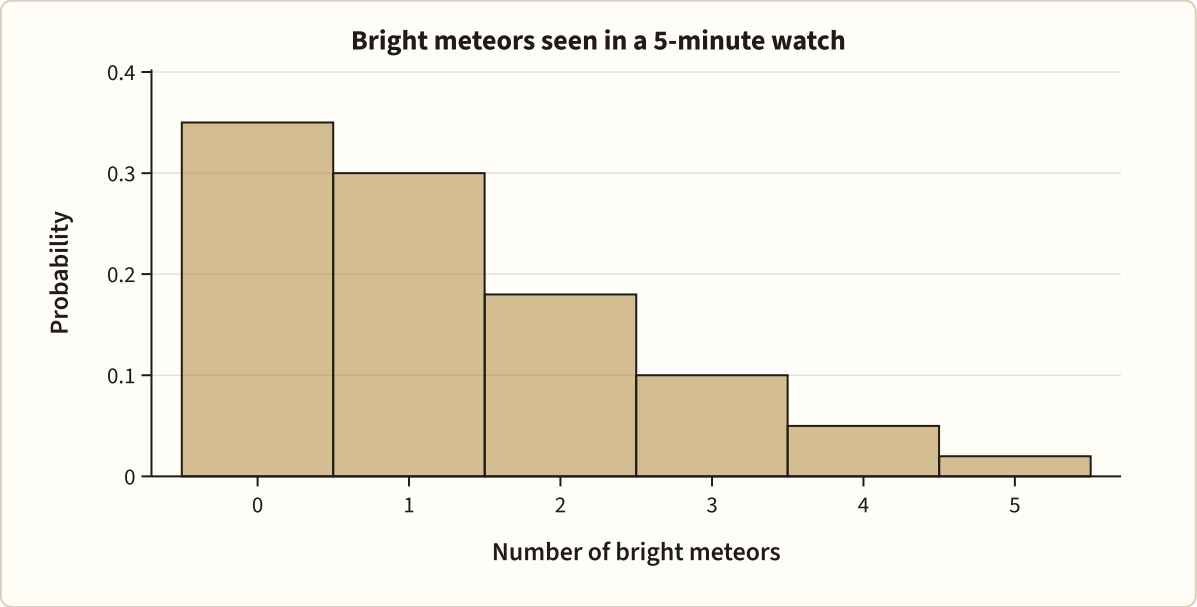
Hits, x	P(x)
0	0.3
1	0.4
2	0.2
3	0.08
4	0.02

6. MEDIUM A dairy farmer proposes the table shown as the probability distribution of X = the number of cows in the herd that will need a veterinarian's visit next month. Is this a valid probability distribution? Answer yes or no and explain.

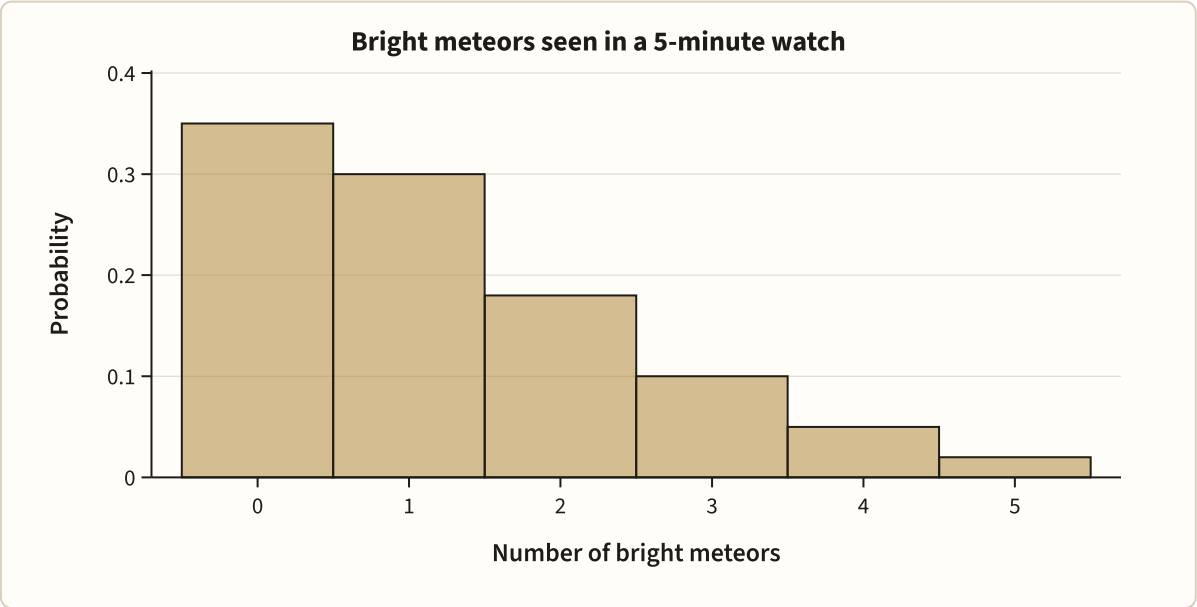
Proposed distribution: cows needing a vet visit in a month

Cows, x	P(x)
0	0.45
1	0.3
2	0.2
3	0.1

7. **MEDIUM** Henry watches the sky for 5 minutes on a clear August night. Let X = the number of bright meteors he sees; the probability histogram is shown. Describe the shape of the distribution: symmetric, skewed left or skewed right?



8. **MEDIUM** For Henry's meteor watch (the probabilities are 0.35, 0.30, 0.18, 0.10, 0.05, 0.02 for 0 through 5 meteors), find (a) the probability that he sees at least 3 meteors and (b) the probability that he sees at least one meteor.



9. **MEDIUM** In Henry's meteor distribution, $P(X = 0) = 0.35$. Explain what this probability means in the context of many 5-minute watches.

10. MEDIUM For the softball player in Problem 4, Samuel says that "at least 2 hits" and "more than 2 hits" have the same probability. (a) Find $P(X \geq 2)$. (b) Find $P(X > 2)$. (c) Explain Samuel's mistake.

Hits by a softball player in one game

Hits, x	P(x)
0	0.3
1	0.4
2	0.2
3	0.08
4	0.02

11. HARD A weather station recorded the number of thunderstorms on each of 80 summer days (table shown). Let X = the number of thunderstorms on a randomly chosen day from these records. (a) Give the probability distribution of X (the probabilities for 0, 1, 2, 3, in order). (b) Find $P(X \geq 2)$.

Thunderstorms per day over 80 summer days

Thunderstorms in a day	Number of days
0	44
1	24
2	10
3	2
Total	80

12. HARD Let X = the number of goals a youth hockey team scores in a game. The team never scores more than 4 goals, and $P(X = 3)$ is twice $P(X = 4)$, as shown in the table. Find $P(X = 3)$ and $P(X = 4)$.

Goals scored by a hockey team in a game

Goals, x	P(x)
0	0.15
1	0.25
2	0.3
3	2k
4	k

13. HARD Grace spins a fair spinner with four equal sections numbered 1, 2, 3, 4 twice. Let X = the larger of the two numbers (if they are equal, X is that number). Find (a) $P(X = 3)$ and (b) $P(X = 4)$. Give exact decimals.

14. CHALLENGE Two fair six-sided dice are rolled, and X = the difference between the larger and the smaller number (so X = 0 when they match). Find (a) $P(X = 1)$ and (b) $P(X \geq 3)$. Round to 4 decimal places.

15. CHALLENGE A beekeeper has a box of 5 jars of honey, and 2 of them are slightly underfilled. Ruth chooses 2 jars at random without replacement. Let X = the number of underfilled jars she chooses. Find the probability distribution of X: give $P(X = 0)$, $P(X = 1)$ and $P(X = 2)$.

COMMON MISTAKES

- Forgetting to check that the probabilities add to 1. Always add them before using a table; a total other than 1 means the table is incomplete or wrong.
- Treating "at least 3" and "more than 3" as the same event. For a discrete random variable, $P(X \geq 3)$ includes $P(X = 3)$ and $P(X > 3)$ does not; write the event with a symbol before adding.
- Adding the values x instead of the probabilities P(x) when finding the probability of an event. The answer to a probability question must be between 0 and 1.
- Using the complement incorrectly, such as writing $P(X \geq 2) = 1 - P(X = 0)$. The complement of $X \geq 2$ is $X \leq 1$, so $P(X \geq 2) = 1 - P(X = 0) - P(X = 1)$.

ANSWERS TO PRACTICE

1. (a) Discrete: it is a count (0, 1, 2, ...). (b) Continuous: rainfall can be any value in an interval, such as 2.37 inches.

2. The probabilities must add to 1: $P(X = 3) = 1 - (0.20 + 0.35 + 0.25 + 0.05) = 1 - 0.85 = 0.15$.
3. $P(X \geq 2) = 0.25 + 0.15 + 0.05 = 0.45$
4. $P(X \geq 1) = 1 - P(X = 0) = 1 - 0.30 = 0.70$
5. (a) $P(X > 1) = 0.20 + 0.08 + 0.02 = 0.30$ (b) $P(1 \leq X \leq 3) = 0.40 + 0.20 + 0.08 = 0.68$
6. No. The probabilities add to $0.45 + 0.30 + 0.20 + 0.10 = 1.05$, not 1, so at least one probability is too large.
7. Skewed right: most of the probability is at 0, 1 and 2 meteors, with a long tail of less likely values stretching to 5.
8. (a) $P(X \geq 3) = 0.10 + 0.05 + 0.02 = 0.17$ (b) $P(X \geq 1) = 1 - 0.35 = 0.65$
9. If Henry made a very large number of 5-minute watches on nights like this one, he would see no bright meteors in about 35% of them. It does not mean that exactly 35 of every 100 watches will be empty, only that the proportion settles near 0.35 in the long run.
10. (a) $0.20 + 0.08 + 0.02 = 0.30$ (b) $0.08 + 0.02 = 0.10$ (c) "At least 2" includes the value 2 and "more than 2" does not. For a discrete random variable the single value 2 has its own probability (0.20), so the two events differ by exactly that amount.
11. (a) Divide each count by 80: $P(0) = \frac{44}{80} = 0.55$, $P(1) = \frac{24}{80} = 0.30$, $P(2) = \frac{10}{80} = 0.125$, $P(3) = \frac{2}{80} = 0.025$ (they add to 1). (b) $P(X \geq 2) = 0.125 + 0.025 = 0.15$
12. The known probabilities add to $0.15 + 0.25 + 0.30 = 0.70$, so $2k + k = 0.30$, $3k = 0.30$, $k = 0.10$. So $P(X = 3) = 0.20$ and $P(X = 4) = 0.10$.
13. There are 16 equally likely ordered outcomes. X = 3 when both spins are at most 3 and at least one is 3: (1,3), (2,3), (3,1), (3,2), (3,3), so $P(X = 3) = \frac{5}{16} = 0.3125$. X = 4 when at least one spin is 4: $16 - 9 = 7$ outcomes (9 outcomes have both spins at most 3), so $P(X = 4) = \frac{7}{16} = 0.4375$.
14. There are 36 equally likely outcomes. The counts are X = 0: 6, X = 1: 10, X = 2: 8, X = 3: 6, X = 4: 4, X = 5: 2 (total 36). (a) $P(X = 1) = \frac{10}{36} \approx 0.2778$ (b) $P(X \geq 3) = \frac{6 + 4 + 2}{36} = \frac{12}{36} \approx 0.3333$
15. $P(X = 0) = \frac{3}{5} \times \frac{2}{4} = \frac{6}{20} = 0.3$; $P(X = 2) = \frac{2}{5} \times \frac{1}{4} = \frac{2}{20} = 0.1$; $P(X = 1) = 1 - 0.3 - 0.1 = 0.6$. (Check by counting: of the $\binom{5}{2} = 10$ possible pairs, 3 have no underfilled jar, 6 have one and 1 has two.)

LESSON 7.2 · 2 SESSIONS OF 45 MINUTES

Expected Value and Standard Deviation

OBJECTIVES

- Compute the mean (expected value) $\mu_X = \sum x\,P(x)$ of a discrete random variable and interpret it as a long-run average.
- Compute the variance and standard deviation of a discrete random variable and interpret the standard deviation as a typical distance from the mean.
- Use expected values to judge whether a game is fair and to compare decisions such as buying insurance or choosing a crop.
- Explain why a negative expected value means a game is a losing proposition in the long run.

VIDEO

WATCH

Expected Value (2 of 2: Non-uniform probability distributions)

Eddie Woo · 11:43

When to watch: Before Example 1: expected value from a probability table.

https://www.youtube.com/watch?v=hqvh7T7Y51w

BACKUP

The Expected Value and Variance of Discrete Random Variables

jbstatistics · 11:19

When to watch: After Example 1: the variance and standard deviation of a random variable.

https://www.youtube.com/watch?v=Vyk8HQOckIE

> **Note:** Eddie Woo uses a pay-to-play dice game to show that the player loses in the long run; it is a warning against gambling, not an endorsement.

KEY TERMS

Expected value (mean) of a random variable	The long-run average value of X over very many repetitions of the chance process, written μ_X or $E(X)$. It is found by multiplying each value by its probability and adding.
Variance of a random variable	The probability-weighted average of the squared deviations from the mean, written σ_X^2.

Standard deviation of a random variable	The square root of the variance, written σ_X. It measures how far the values of X typically fall from the mean, in the units of X.
Fair game	A game in which the player's expected net gain is exactly 0, so neither the player nor the one running the game comes out ahead in the long run.
Net gain	What a player receives minus what the player pays. A loss is a negative net gain.

CORE IDEAS

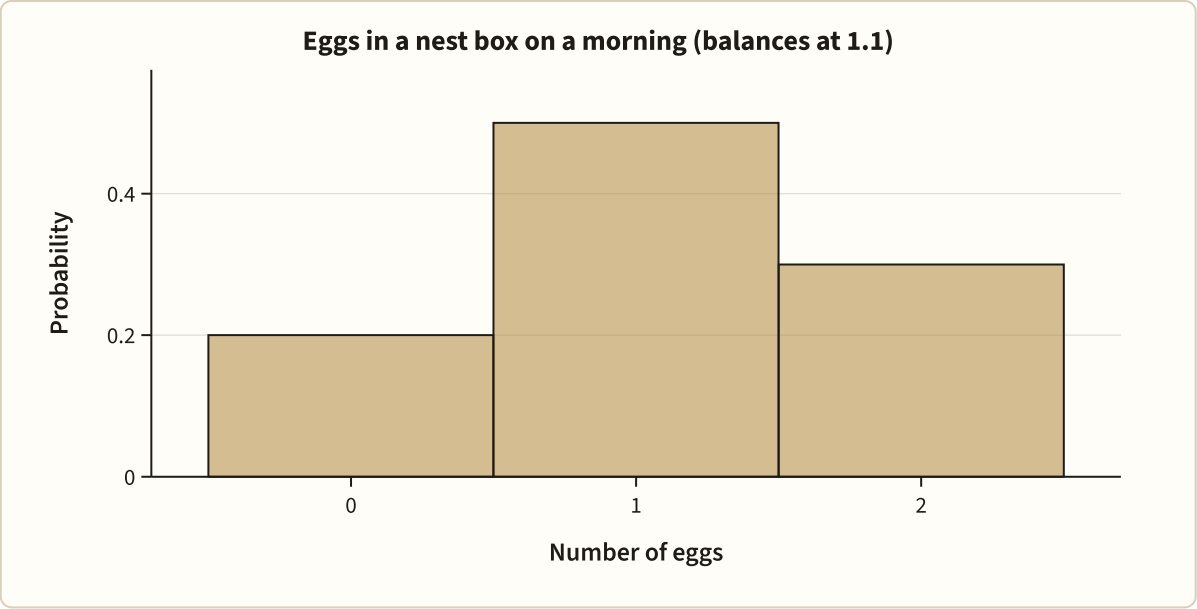
Let X = the number of eggs found in a certain nest box on a morning, with $P(0) = 0.2$, $P(1) = 0.5$ and $P(2) = 0.3$. What is the average number of eggs per morning in the long run? Imagine 1000 mornings. About 200 of them give 0 eggs, about 500 give 1 egg and about 300 give 2 eggs, for a total of $0(200) + 1(500) + 2(300) = 1100$ eggs, or $1100 \div 1000 = 1.1$ eggs per morning.

Look at what that calculation really did: $\dfrac{0(200) + 1(500) + 2(300)}{1000} = 0(0.2) + 1(0.5) + 2(0.3) = 1.1$. Each value is multiplied by its probability, and the products are added. This is the **mean** or **expected value** of X: $\mu_X = \sum x\,P(x)$. It is a weighted mean, like a grade average, where the weights are probabilities. By the law of large numbers, the average of many observed values of X gets closer and closer to μ_X.

Eggs in a nest box: computing the mean and variance

x	P(x)	x P(x)	(x - mean)^2	(x - mean)^2 P(x)
0	0.2	0	1.21	0.242
1	0.5	0.5	0.01	0.005
2	0.3	0.6	0.81	0.243
Total	1	1.1		0.49

The expected value need not be a possible value: no nest box holds 1.1 eggs. It is a long-run average, not a prediction for one morning. On the probability histogram, μ_X is the balance point: if the bars were blocks of wood on a board, the board would balance at 1.1.



To measure spread, we use the same idea as the standard deviation of data in Lesson 2.4: find each deviation $x - \mu_X$, square it so that negative and positive deviations do not cancel, and average the squares, weighting each by its probability. That gives the **variance** $\sigma_X^2 = \sum (x - \mu_X)^2 P(x)$; its square root σ_X is the **standard deviation**, back in the units of X. For the nest box, $\sigma_X^2 = 1.21(0.2) + 0.01(0.5) + 0.81(0.3) = 0.242 + 0.005 + 0.243 = 0.49$ and $\sigma_X = \sqrt{0.49} = 0.7$ eggs. Why no $n - 1$ here? The $n - 1$ for a sample corrects for measuring deviations from $\bar{x}$, an estimate pulled toward the data. A probability distribution describes the whole long run, and μ_X is the true mean, so nothing needs correcting; it is like the population σ.

Expected value tells us whether a game is fair. At a church fall festival, the youth group runs a beanbag toss: it costs \$1 to play, and from past festivals about 10% of players win a \$4 pie voucher. A player's net gain is \$3 with probability 0.10 and $-$\$1 with probability 0.90, so the expected net gain is $3(0.10) + (-1)(0.90) = -\$0.60$. Each player loses 60 cents on average, which means the booth raises about \$0.60 per play, or about \$300 from 500 plays. A negative expected value means the player is behind in the long run, and the more times they play, the more surely their average result settles near the loss. Every game of chance run for profit is designed this way, so that the one running the game wins in the long run. Here that is the purpose: the players are happy to give to the youth group.

Expected value guides decisions, but it is not the whole story. An insurance company charges more than its expected payout, so buying insurance has a negative expected value for the farmer; the farmer buys it anyway to avoid a rare loss that could be ruinous. When two choices have similar means, the standard deviation measures risk: a larger σ_X means results that swing more widely from year to year.

KEY FORMULAS

$$\mu_X = E(X) = \sum x\,P(x)$$

$$\sigma_X^2 = \sum (x - \mu_X)^2\,P(x)$$

$$\sigma_X = \sqrt{\sigma_X^2}$$

A game is fair when the expected net gain is 0.

From history. The Dutch scientist Christiaan Huygens wrote one of the first printed books on probability in 1657. In it he reasoned about the fair value of a game by weighing each payoff by its chance, the idea we now call expected value. Huygens is also remembered in astronomy for discovering Saturn's largest moon, Titan.

Practice habit. Always say what an expected value means in the long run and give its units: "over many mornings, the nest box averages 1.1 eggs per morning." When you compute a variance by hand, make a table with columns x, P(x), xP(x), $(x-\mu)^2$ and $(x-\mu)^2 P(x)$, and keep full precision until the last step.

WORKED EXAMPLES

Example 1. On the Miller farm, X = the number of eggs a certain hen lays in a week has the distribution shown. Find the mean and standard deviation of X and interpret the mean. Round the standard deviation to two decimal places.

Eggs laid by one hen in a week

Eggs, x	P(x)
3	0.05
4	0.1
5	0.25
6	0.4
7	0.2

1. Mean: $\mu_X = \sum x\,P(x) = 3(0.05) + 4(0.10) + 5(0.25) + 6(0.40) + 7(0.20)$.
2. $\mu_X = 0.15 + 0.40 + 1.25 + 2.40 + 1.40 = 5.6$ eggs.
3. Deviations from 5.6: $-2.6, -1.6, -0.6, 0.4, 1.4$. Squares: $6.76, 2.56, 0.36, 0.16, 1.96$.
4. Weight each square by its probability: $6.76(0.05) = 0.338$, $2.56(0.10) = 0.256$, $0.36(0.25) = 0.09$, $0.16(0.40) = 0.064$, $1.96(0.20) = 0.392$.
5. Variance: $\sigma_X^2 = 0.338 + 0.256 + 0.09 + 0.064 + 0.392 = 1.14$.
6. Standard deviation: $\sigma_X = \sqrt{1.14} \approx 1.07$ eggs.
7. Conclusion: over many weeks this hen averages about 5.6 eggs per week, and her weekly count typically differs from 5.6 by about 1.07 eggs.

Answer: $\mu_X = 5.6$ eggs, $\sigma_X \approx 1.07$ eggs. Over many weeks the hen averages 5.6 eggs per week.

Example 2. A farmer must choose between planting corn and planting a drought-tolerant sorghum. His profit per acre depends on the weather, as shown in the table. (a) Find the mean and standard deviation of the profit per acre for corn. (b) Do the same for sorghum. Round standard deviations to the nearest dollar. (c) Compare the two crops in context.

Profit per acre (dollars) for two crops

Weather	Probability	Corn	Sorghum
Wet year	0.3	300	180
Normal year	0.5	200	170
Dry year	0.2	-50	150

1. (a) Corn mean: $\mu = 300(0.3) + 200(0.5) + (-50)(0.2) = 90 + 100 - 10 = \$180$.
2. Corn deviations: $120, 20, -230$; squares $14400, 400, 52900$.
3. Corn variance: $14400(0.3) + 400(0.5) + 52900(0.2) = 4320 + 200 + 10580 = 15100$, so $\sigma = \sqrt{15100} \approx \$123$.
4. (b) Sorghum mean: $\mu = 180(0.3) + 170(0.5) + 150(0.2) = 54 + 85 + 30 = \$169$.
5. Sorghum deviations: $11, 1, -19$; squares $121, 1, 361$.
6. Sorghum variance: $121(0.3) + 1(0.5) + 361(0.2) = 36.3 + 0.5 + 72.2 = 109$, so $\sigma = \sqrt{109} \approx \$10$.
7. (c) Corn earns more on average ($180 versus $169 per acre, an $11 difference), but its profit varies about twelve times as much and can even be a loss in a dry year. Sorghum gives a slightly lower but very steady profit.
8. Conclusion: a farmer who can afford a bad year may prefer corn for its higher long-run average; a farmer who cannot risk a loss may sensibly choose sorghum.

Answer: (a) Corn: $\mu = \$180$, $\sigma \approx \$123$ per acre. (b) Sorghum: $\mu = \$169$, $\sigma \approx \$10$ per acre. (c) Corn has the higher expected profit but far more risk; sorghum is slightly lower on average but much more dependable.

PRACTICE

1. EASY At the end of a basketball game, Grace takes two free throws. Let X = the number she makes, with the distribution shown. Find the expected value of X.

Free throws made in two attempts

Made, x	P(x)
0	0.04
1	0.32
2	0.64

2. EASY Let X = the number of books a visitor checks out of the parish library on one visit, with the distribution shown. Find μ_X.

Books checked out from the parish library per visit

Books, x	P(x)
0	0.2
1	0.35
2	0.25
3	0.15
4	0.05

3. EASY Interpret the mean you found in Problem 2 in context. Can a visitor actually check out 1.5 books?
4. EASY A fair six-sided die is rolled. Let X = the number on top. Find the expected value of X.
5. MEDIUM For Grace's free throws in Problem 1 ($\mu_X = 1.6$), find the standard deviation of X. Round to two decimal places.

Free throws made in two attempts

Made, x	P(x)
0	0.04
1	0.32
2	0.64

6. MEDIUM A youth group runs a beanbag toss at the church fair. It costs $3 to play. From past fairs, 15% of players win a $10 prize and 10% win a $5 prize; the table shows the player's net gain. (a) Find the player's expected net gain. (b) About how much money will the booth raise from 200 plays?

Beanbag toss at the church fair: player's net gain

Result	Net gain (dollars)	Probability
Large prize	7	0.15
Small prize	2	0.1
No prize	-3	0.75

7. MEDIUM An insurance company sells a one-year policy on a tractor for $800. The table shows what the company pays out and how likely each event is. (a) Find the company's expected payout per policy. (b) Find the company's expected gain per policy.

Tractor insurance: payout in one year

Event	Payout (dollars)	Probability
Major loss	20000	0.005
Minor damage	4000	0.06
No claim	0	0.935

8. MEDIUM In Problem 7, the farmer's expected net gain from buying the policy is negative. Explain why buying the insurance can still be a wise decision for the farmer.

9. MEDIUM Let X = the number of rainy days during a 3-day church camp in June, with the distribution shown. Find the mean and standard deviation of X. Round the standard deviation to two decimal places.

Rainy days during a 3-day church camp

Rainy days, x	P(x)
0	0.3
1	0.4
2	0.2
3	0.1

10. HARD For a family game night, Samuel designs a dice game played with tokens. A player pays 2 tokens and rolls a fair die. On a 6 the player receives W tokens; otherwise the player receives nothing. (The 2 tokens are not returned.) What value of W makes the game fair?

11. HARD Ruth can sell her strawberry harvest at Saturday's market, where she earns \$480 if the day is sunny (probability 0.6) or \$150 if it rains (probability 0.4), or she can sell it all to a grocer for a certain \$350. (a) Find her expected earnings at the market. (b) Which option would you advise, and why?

Strawberry sales options

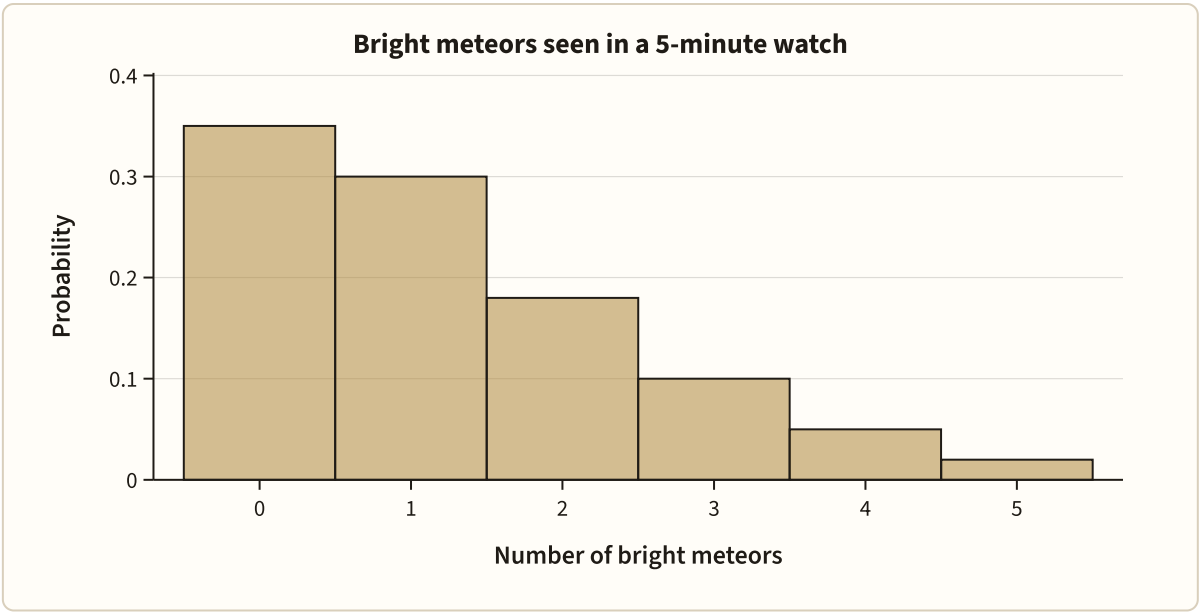
Option	Sunny Saturday (prob. 0.6)	Rainy Saturday (prob. 0.4)
Sell at the market	480	150
Sell to the grocer	350	350

12. HARD Find the standard deviation of Ruth's earnings if she sells at the market in Problem 11 ($\mu =$ \$348). Round to the nearest cent.

Strawberry sales options

Option	Sunny Saturday (prob. 0.6)	Rainy Saturday (prob. 0.4)
Sell at the market	480	150
Sell to the grocer	350	350

13. HARD For Henry's 5-minute meteor watch (distribution shown), (a) find the expected number of bright meteors. (b) The median number is 1. Explain why the mean is larger than the median.



14. **CHALLENGE** On a sheep farm, each ewe gives birth to 1, 2 or 3 lambs. The probability of twins is 0.5, and the mean number of lambs per ewe is 1.9. Find p = P(1 lamb) and q = P(3 lambs).

Lambs born to a ewe

Lambs, x	P(x)
1	p
2	0.5
3	q

COMMON MISTAKES

- Averaging the values of X without weighting them, such as $(0 + 1 + 2) \div 3 = 1$ for the nest box. Values are not equally likely; multiply each by its probability.
- Dividing by n or n − 1 when finding the variance of a random variable. The probabilities are already the weights; the variance is $\sum (x - \mu_X)^2 P(x)$ with no division.
- Forgetting to take the square root at the end, or rounding μ_X before computing the deviations. Keep the exact mean, compute the variance, then take its square root.
- Leaving out the cost to play when finding a player's expected gain. Use the net gain (prize minus cost) for each outcome, including the outcome where nothing is won.

ANSWERS TO PRACTICE

1. $\mu_X = 0(0.04) + 1(0.32) + 2(0.64) = 0 + 0.32 + 1.28 = 1.6$ free throws
2. $\mu_X = 0(0.20) + 1(0.35) + 2(0.25) + 3(0.15) + 4(0.05) = 0 + 0.35 + 0.50 + 0.45 + 0.20 = 1.5$ books
3. Over a very large number of visits, the library lends an average of about 1.5 books per visit. No single visitor checks out 1.5 books; the mean is a long-run average, not a value of one visit.
4. $\mu_X = (1 + 2 + 3 + 4 + 5 + 6)\left(\frac{1}{6}\right) = \frac{21}{6} = 3.5$
5. $\sigma_X^2 = (0 - 1.6)^2(0.04) + (1 - 1.6)^2(0.32) + (2 - 1.6)^2(0.64) = 0.1024 + 0.1152 + 0.1024 = 0.32$; $\sigma_X = \sqrt{0.32} \approx 0.57$ free throws
6. (a) $7(0.15) + 2(0.10) + (-3)(0.75) = 1.05 + 0.20 - 2.25 = -\$1.00$ per play. (b) The booth gains about \$1.00 per play, so about $200 \times \$1.00 = \$200$.
7. (a) $20000(0.005) + 4000(0.06) + 0(0.935) = 100 + 240 + 0 = \$340$ (b) $800 - 340 = \$460$ per policy

8. Expected value describes the average over many years or many farmers. One farmer faces one tractor, and a \$20,000 loss in a single year could be ruinous. Paying a certain \$800 to avoid a small chance of a huge loss trades a little expected money for a large reduction in risk; the insurance company can accept that risk because it spreads it over many policies.
9. $\mu_X = 0(0.3) + 1(0.4) + 2(0.2) + 3(0.1) = 1.1$ days. $\sigma_X^2 = 1.21(0.3) + 0.01(0.4) + 0.81(0.2) + 3.61(0.1) = 0.363 + 0.004 + 0.162 + 0.361 = 0.89$; $\sigma_X = \sqrt{0.89} \approx 0.94$ days
10. Expected net gain $= W\left(\frac{1}{6}\right) - 2$. Set it equal to 0: $\frac{W}{6} = 2$, so $W = 12$ tokens.
11. (a) $480(0.6) + 150(0.4) = 288 + 60 = \$348$ (b) The grocer pays \$350 for certain, which is slightly more than the market's expected \$348 and carries no risk of a \$150 day, so selling to the grocer is the better choice.
12. Deviations: $480 - 348 = 132$ and $150 - 348 = -198$. $\sigma^2 = 132^2(0.6) + (-198)^2(0.4) = 17424(0.6) + 39204(0.4) = 10454.4 + 15681.6 = 26136$; $\sigma = \sqrt{26136} \approx \$161.67$
13. (a) $\mu_X = 0(0.35) + 1(0.30) + 2(0.18) + 3(0.10) + 4(0.05) + 5(0.02) = 0 + 0.30 + 0.36 + 0.30 + 0.20 + 0.10 = 1.26$ meteors (b) The distribution is skewed right. The less likely large values (3, 4 and 5 meteors) pull the balance point to the right of the median, just as a long right tail pulls the mean of a data set above its median.
14. The probabilities add to 1: $p + 0.5 + q = 1$, so $p + q = 0.5$. The mean: $1p + 2(0.5) + 3q = 1.9$, so $p + 3q = 0.9$. Subtract: $2q = 0.4$, $q = 0.2$, and $p = 0.3$.

LESSON 7.3 · 2 SESSIONS OF 45 MINUTES

Transforming and Combining Random Variables

OBJECTIVES

- Find the mean and standard deviation of $Y = a + bX$ from those of X, and explain why adding a constant does not change the spread.
- Find the mean of a sum or difference of random variables, and the standard deviation when the variables are independent.
- Explain why variances (not standard deviations) add, and why a difference of independent random variables has more variability, not less.
- Apply these rules to totals and differences in context, such as the weight of a box of fruit or the profit of a bakery.

VIDEO

WATCH

Variance on a Modified Distribution (1 of 2: Worked example)

Eddie Woo · 9:53

When to watch: Before Example 1: how a linear change (a + bX) affects the mean and variance.

https://www.youtube.com/watch?v=9lwTmtQ8X8g

BACKUP

Combining Means and Variance Examples

ProfRobBob · 14:34

When to watch: Before Example 2: combining means and variances of independent random variables.

https://www.youtube.com/watch?v=RbbuQAbweLo

KEY TERMS

Linear transformation	A change of a random variable of the form $Y = a + bX$, such as converting units or turning a count into a profit.
Independent random variables	Random variables for which knowing the value of one does not change the probabilities for the other.
Sum of random variables	A new random variable $X + Y$ formed by adding the values of X and Y, such as the total weight of two items.

Difference of random variables A new random variable $X - Y$ formed by subtracting, such as how much longer one runner takes than another.

CORE IDEAS

A youth group runs a car wash. Let X = the number of cars washed in an hour, with $\mu_X = 8$ cars and $\sigma_X = 2$ cars. They charge \$6 per car and spend \$5 per hour on soap and water, so the hourly profit is $Y = 6X - 5$. Multiplying by 6 multiplies every value by 6, so the mean becomes $6(8) = 48$ and every distance from the mean is stretched by 6: the standard deviation becomes $6(2) = 12$. Subtracting 5 slides every value down by \$5; the mean drops to $48 - 5 = \$43$, but the distances between values do not change, so the spread stays \$12. In general, for $Y = a + bX$: $\mu_Y = a + b\mu_X$ and $\sigma_Y = |b|\sigma_X$, just as for data in Lesson 2.6. We use $|b|$ because a standard deviation is a distance and is never negative.

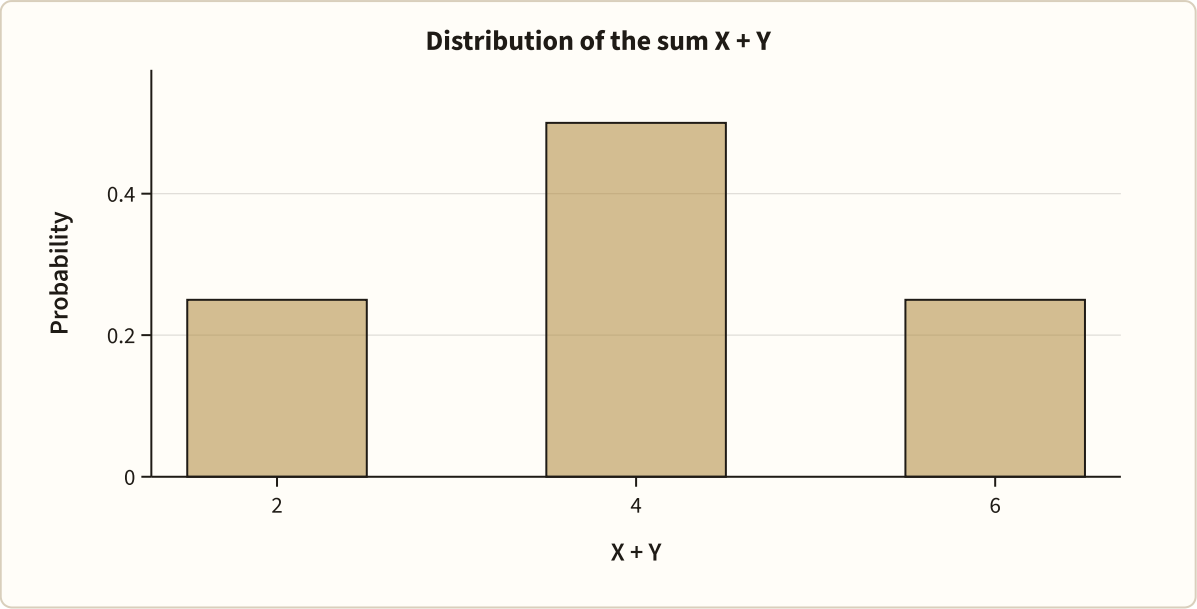
Now combine two random variables. Means always add: $\mu_{X+Y} = \mu_X + \mu_Y$ and $\mu_{X-Y} = \mu_X - \mu_Y$. This makes sense: if one hen averages 5 eggs a week and another averages 6, together they average 11. This rule holds whether or not X and Y are independent.

Spread is different. Here is a tiny example. Let X be 1 or 3 with probability 0.5 each, and let Y, independent of X, also be 1 or 3 with probability 0.5 each. Each has mean 2 and standard deviation 1 (every value is 1 away from 2). The four equally likely combinations are in the table.

Four equally likely combinations of X and Y

X	Y	X + Y	X - Y
1	1	2	0
1	3	4	-2
3	1	4	2
3	3	6	0

The sum $X + Y$ is 2, 4, 4 or 6, with mean 4. Its variance is $\frac{(2-4)^2 + 0^2 + 0^2 + (6-4)^2}{4} = \frac{8}{4} = 2$, so its standard deviation is $\sqrt{2} \approx 1.41$, not $1 + 1 = 2$. The variances added ($1 + 1 = 2$); the standard deviations did not. Why? The sum is extreme only when both variables are extreme in the same direction. Half the time one is high and the other low, and they partly cancel, so the sum is less spread out than adding the two spreads would suggest.



What about the difference $X - Y$? Its values are 0, -2, 2 and 0, with mean 0 and variance $\frac{0+4+4+0}{4} = 2$, the same as the sum. Variances add for differences too: $\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$. Subtracting does not cancel variability. The difference is largest when X is high and Y is low, and most negative when X is low and Y is high, so both sources of variation show up in it. (Subtracting Y flips the sign of its deviations, but squaring removes the sign.)

So for independent random variables, add the variances, then take the square root: $\sigma_{X\pm Y} = \sqrt{\sigma^2_X + \sigma^2_Y}$. It works like the Pythagorean theorem, with standard deviations as the legs of a right triangle. Never add or subtract standard deviations. The rule needs independence: the yields of two neighboring fields share the same weather and tend to be high or low together, so their variances do not simply add.

KEY FORMULAS

$$Y = a + bX: \quad \mu_Y = a + b\mu_X, \qquad \sigma_Y = |b|\,\sigma_X$$

$$\mu_{X+Y} = \mu_X + \mu_Y, \qquad \mu_{X-Y} = \mu_X - \mu_Y$$

$$\text{If X and Y are independent:} \quad \sigma^2_{X+Y} = \sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$$

$$\sigma_{X\pm Y} = \sqrt{\sigma^2_X + \sigma^2_Y}$$

Practice habit. Variances add, standard deviations do not. Whenever you combine independent random variables, square each standard deviation, add the variances (for a sum OR a difference), and only then take the square root.

WORKED EXAMPLES

Example 1. In a certain town, the high temperature X on a randomly chosen July day has mean 77°F and standard deviation 9°F. Find the mean and standard deviation of the high temperature in degrees Celsius, $C = \frac{5}{9}(X - 32)$.

1. Write C in the form $a + bX$: $C = \frac{5}{9}X - \frac{160}{9}$, so $b = \frac{5}{9}$.
2. Mean: apply the same conversion to the mean, $\mu_C = \frac{5}{9}(77 - 32) = \frac{5}{9}(45) = 25$°C.
3. Standard deviation: subtracting 32 does not change spread, and multiplying by $\frac{5}{9}$ shrinks it: $\sigma_C = \frac{5}{9}(9) = 5$°C.
4. Conclusion: in Celsius, July highs in this town average 25°C and typically vary from that by about 5°C.

Answer: $\mu_C = 25$°C, $\sigma_C = 5$°C

Example 2. A farm stand sells harvest boxes. Each box holds a bag of apples with weight X and a bag of pears with weight Y; the two weights are independent. The empty box weighs exactly 0.5 pound. (a) Find the mean and standard deviation of the total weight T of a filled box. (b) Find the mean and standard deviation of $D = X - Y$, how much heavier the apples are than the pears.

Harvest box contents (weights in pounds)

Part	Mean	Standard deviation
Apples, X	5	0.3
Pears, Y	3	0.4
Empty box	0.5	0

1. (a) The total is $T = X + Y + 0.5$. Means add: $\mu_T = 5.0 + 3.0 + 0.5 = 8.5$ pounds.
2. The box weight is a constant, so it adds nothing to the spread. X and Y are independent, so their variances add: $\sigma_T^2 = 0.3^2 + 0.4^2 = 0.09 + 0.16 = 0.25$.
3. $\sigma_T = \sqrt{0.25} = 0.5$ pound. (Adding the standard deviations would give 0.7, which is wrong.)
4. (b) Means subtract: $\mu_D = 5.0 - 3.0 = 2.0$ pounds.
5. Variances still add: $\sigma_D^2 = 0.09 + 0.16 = 0.25$, so $\sigma_D = 0.5$ pound.
6. Conclusion: filled boxes weigh 8.5 pounds on average, typically within about 0.5 pound of that, and the apples outweigh the pears by 2.0 pounds on average, with the difference also varying by about 0.5 pound.

Answer: (a) $\mu_T = 8.5$ lb, $\sigma_T = 0.5$ lb (b) $\mu_D = 2.0$ lb, $\sigma_D = 0.5$ lb

PRACTICE

1. EASY The time X a bakery oven takes to preheat has mean 12 minutes and standard deviation 2 minutes. The baker always spends exactly 3 more minutes setting out pans after it heats, so the total wait is $Y = X + 3$. Find the mean and standard deviation of Y.
2. EASY The number of dozens of eggs X sold at a farm stand per day has mean 12 and standard deviation 3. Eggs sell for $4 per dozen, so daily revenue is $R = 4X$. Find the mean and standard deviation of R.
3. EASY Ruth drives to church (mean 18 minutes) and then walks from the parking lot to the choir room (mean 6 minutes). What is the mean of her total travel time?

4. EASY In a two-runner relay, the leg times X and Y are independent, with the means and standard deviations shown. Find the standard deviation of the total time $X + Y$.

Relay leg times (seconds)

Runner	Mean	Standard deviation
First leg, X	60	3
Second leg, Y	62	4

5. MEDIUM For the relay in Problem 4, let $D = Y - X$, how many seconds longer the second leg takes than the first. Find the mean and standard deviation of D.

Relay leg times (seconds)

Runner	Mean	Standard deviation
First leg, X	60	3
Second leg, Y	62	4

6. MEDIUM Samuel says the standard deviation of the difference in Problem 5 should be $4 - 3 = 1$ second, because "you subtract the times, so you subtract the spreads." Explain why he is wrong.

7. MEDIUM The overnight low temperature X on a farm in early autumn has mean 50°F and standard deviation 9°F. Find the mean and standard deviation of the low in Celsius, $C = \frac{5}{9}(X - 32)$.

8. MEDIUM Each morning Samuel fills a cattle trough with 50 gallons of water. The amount X the cattle drink by evening has mean 32 gallons and standard deviation 4 gallons. Find the mean and standard deviation of the water left in the evening, $R = 50 - X$.

9. MEDIUM Henry builds birdhouses and sells them at a craft booth for $12 each. The number X he sells in a month has mean 15 and standard deviation 4. He pays a $20 booth fee each month, so his monthly profit is $P = 12X - 20$. Find the mean and standard deviation of P.

10. HARD Sacks of flour have weights with mean 50 pounds and standard deviation 0.6 pound, independently of one another. Find the mean and standard deviation of the total weight of 3 sacks. Round the standard deviation to two decimal places.

11. HARD In Problem 10, Grace says the standard deviation of the total should be $3 \times 0.6 = 1.8$ pounds. (a) For what random variable would 1.8 pounds be correct? (b) Explain why the total of 3 different sacks varies less than that.

12. **HARD** A bakery's Saturday sales are shown in the table. The numbers of loaves X and pies Y sold are independent. Revenue is $R = 5X + 12Y$. Find the mean and standard deviation of R. Round the standard deviation to the nearest cent.

Saturday bakery sales

Item	Price (dollars)	Mean number sold	Standard deviation
Loaves of bread, X	5	60	8
Pies, Y	12	20	5

13. **HARD** At a sawmill, boards X have mean length 96.0 inches and standard deviation 0.2 inch. A carpenter cuts a piece Y from each board, with mean 30.0 inches and standard deviation 0.15 inch, independently of the board's length. Find the mean and standard deviation of the leftover length $X - Y$.

14. **CHALLENGE** The Miller family's drive to their grandmother's farm has two independent parts: the highway part and the country-road part. The total drive time has standard deviation 13 minutes, and the highway part has standard deviation 5 minutes. Find the standard deviation of the country-road part.

COMMON MISTAKES

- Adding standard deviations, such as $0.3 + 0.4 = 0.7$. Square them, add the variances, then take the square root: $\sqrt{0.09 + 0.16} = 0.5$.
- Subtracting variances for a difference. For independent random variables, $\sigma^2_{X-Y} = \sigma^2_X + \sigma^2_Y$; adding or subtracting independent quantities never makes the variance smaller than the variance of either one.
- Letting a constant change the standard deviation. Adding or subtracting a fixed amount (a fee, an empty box) shifts the mean only; multiplying by b multiplies the standard deviation by $|b|$.
- Confusing $X_1 + X_2 + X_3$ (three different items) with $3X$ (one item tripled). For three independent items the variance is $3\sigma^2$; for $3X$ it is $9\sigma^2$.

ANSWERS TO PRACTICE

1. $\mu_Y = 12 + 3 = 15$ minutes; $\sigma_Y = 2$ minutes (adding a constant does not change spread)
2. $\mu_R = 4(12) = \$48$; $\sigma_R = 4(3) = \$12$
3. Means add: $18 + 6 = 24$ minutes
4. $\sigma^2 = 3^2 + 4^2 = 9 + 16 = 25$; $\sigma = \sqrt{25} = 5$ seconds
5. $\mu_D = 62 - 60 = 2$ seconds; $\sigma_D = \sqrt{3^2 + 4^2} = 5$ seconds

6. Subtracting does not cancel variability. The difference is largest when the second leg is slow and the first is fast, and smallest in the opposite case, so the variation of both runners shows up in D. Variances add for differences of independent variables: $\sigma_D = \sqrt{9 + 16} = 5$ seconds, not 1.
7. $\mu_C = \frac{5}{9}(50 - 32) = \frac{5}{9}(18) = 10$°C; $\sigma_C = \frac{5}{9}(9) = 5$°C
8. $\mu_R = 50 - 32 = 18$ gallons; $\sigma_R = |-1|(4) = 4$ gallons
9. $\mu_P = 12(15) - 20 = \$160$; $\sigma_P = 12(4) = \$48$
10. $\mu = 50 + 50 + 50 = 150$ pounds; $\sigma^2 = 0.36 + 0.36 + 0.36 = 1.08$; $\sigma = \sqrt{1.08} \approx 1.04$ pounds
11. (a) For $3X$, three times the weight of one single sack: $\sigma_{3X} = 3(0.6) = 1.8$ pounds. (b) Three different sacks vary independently, so a heavy sack is often balanced by a light one, and their errors partly cancel. Variances add to 1.08, giving $\sigma \approx 1.04$ pounds. Tripling one sack triples its error with no chance of canceling.
12. $\mu_R = 5(60) + 12(20) = 300 + 240 = \$540$. $\sigma_R^2 = 5^2(8^2) + 12^2(5^2) = 25(64) + 144(25) = 1600 + 3600 = 5200$; $\sigma_R = \sqrt{5200} \approx \$72.11$
13. $\mu = 96.0 - 30.0 = 66.0$ inches; $\sigma = \sqrt{0.2^2 + 0.15^2} = \sqrt{0.04 + 0.0225} = \sqrt{0.0625} = 0.25$ inch
14. Variances add: $13^2 = 5^2 + \sigma^2$, so $\sigma^2 = 169 - 25 = 144$ and $\sigma = 12$ minutes.

LESSON 7.4 · 2 SESSIONS OF 45 MINUTES

Normal Random Variables and Their Sums

OBJECTIVES

- Explain why a continuous random variable has probability equal to area and why any single exact value has probability 0.
- Find probabilities for a normal random variable by standardizing and using the standard normal table or technology.
- Find the distribution of a sum or difference of independent normal random variables, and use it to answer questions in context.
- Interpret probabilities about totals and differences, such as whether a board fits a gap or which runner finishes first.

VIDEO

WATCH

Combining Means and Variance in Statistics PLEASE READ DESCRIPTION

ProfRobBob · 14:59

When to watch: Before Example 1: rules for means and variances, finishing with a sum of normals.

https://www.youtube.com/watch?v=P8DdQ9ZK4y0

BACKUP

A pretty reason why Gaussian + Gaussian = Gaussian

3Blue1Brown · 13:15

When to watch: Enrichment: why the sum of two normal variables is normal (beyond this course).

https://www.youtube.com/watch?v=d_qvLDhkg00

KEY TERMS

Continuous random variable	A random variable that can take any value in an interval. Its probabilities are areas under a density curve.
Normal random variable	A continuous random variable whose distribution is a normal curve $N(\mu, \sigma)$.

Sum of independent normal random variables	If X and Y are independent and both normal, then $X + Y$ and $X - Y$ are also normal, with mean and standard deviation found by the rules of Lesson 7.3.
Clearance	The leftover space when one part fits inside another, such as the gap width minus the board width. A positive clearance means the part fits.

CORE IDEAS

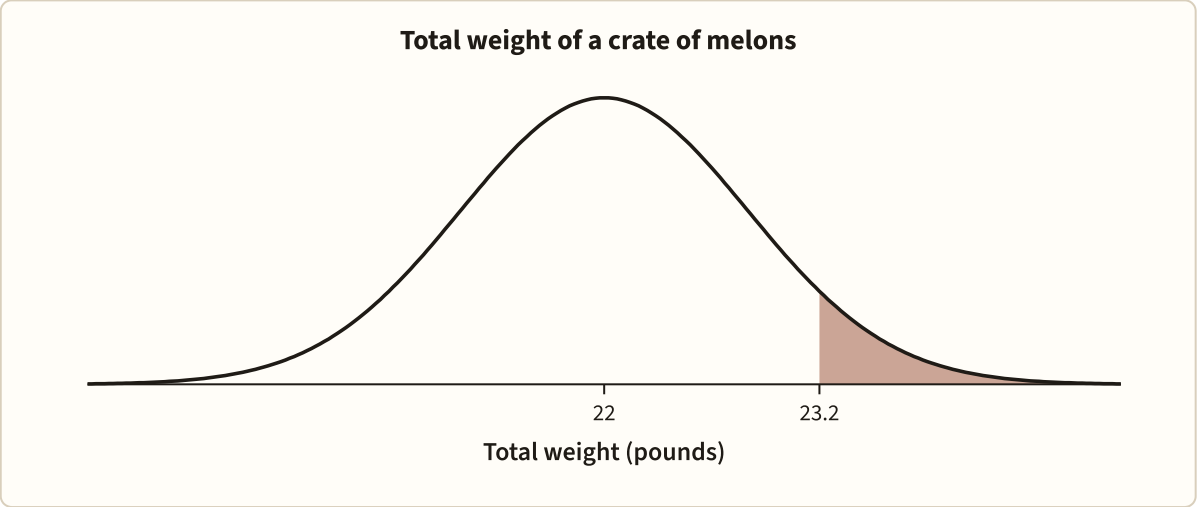
A **continuous random variable** can take any value in an interval: the weight of a melon, the time of a race, the length of a board. We cannot list its values one by one, so we describe it with a density curve, as in Unit 3, and probability is AREA under the curve. For example, if the wait for a farm-tour wagon is equally likely to be anywhere from 0 to 20 minutes, the probability of waiting between 5 and 12 minutes is the area of a rectangle: $(12 - 5) \times \frac{1}{20} = 0.35$.

Since a single value is a line with no width, its area is 0: for a continuous random variable, $P(X = 7) = 0$. This is why $P(X < 7)$ and $P(X \leq 7)$ are the same for a continuous random variable, unlike the discrete case in Lesson 7.1. It does not mean a wait of 7 minutes is impossible; it means that exactly 7.000... minutes, to infinitely many decimal places, has no area of its own.

A **normal random variable** has a normal density curve $N(\mu, \sigma)$. You already know how to find its probabilities: standardize with $z = \frac{x - \mu}{\sigma}$, then use the standard normal table or technology for the area, and always sketch and shade the curve first.

Here is a powerful fact, stated without proof: if X and Y are independent normal random variables, then $X + Y$ and $X - Y$ are also normal. Their means and standard deviations come from Lesson 7.3: means add or subtract, and variances add. So one normal calculation answers questions about totals and differences.

Example: a crate holds 4 melons whose weights are independent and each $N(5.0, 0.4)$ pounds, and the empty crate weighs 2.0 pounds. The total T has mean $4(5.0) + 2.0 = 22.0$ pounds and variance $4(0.4^2) = 0.64$, so $\sigma_T = 0.8$ pound, and T is normal. The probability that a crate weighs more than 23.2 pounds is found with $z = \dfrac{23.2 - 22.0}{0.8} = 1.50$, giving $P(T > 23.2) = 1 - 0.9332 = 0.0668$. About 6.7% of crates weigh more than 23.2 pounds.



Differences answer "which is bigger?" questions. Whether a board fits a gap, or whether one runner beats another, depends on the sign of a difference D. Turn the question into a statement about D (such as $D > 0$), find μ_D and σ_D, and find the area. The rules need independence, so ask whether the two quantities could be linked before you add the variances.

KEY FORMULAS

$$z = \frac{x - \mu}{\sigma}$$

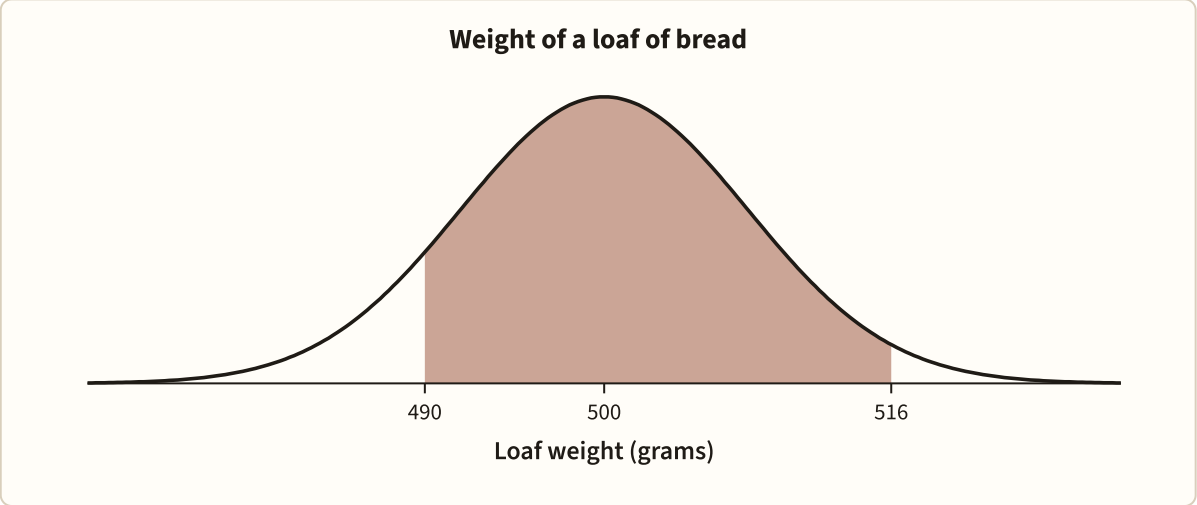
For a continuous random variable, $P(X = a) = 0$, so $P(X < a) = P(X \le a)$

$$X, Y \text{ independent and normal:} \quad X \pm Y \sim N\left(\mu_X \pm \mu_Y,\ \sqrt{\sigma_X^2 + \sigma_Y^2}\right)$$

Practice habit. Before any calculation about a total or a difference, write three things: the random variable in words ("D = gap width minus board width"), its mean, and its standard deviation found by adding variances. Then draw and shade the normal curve for that variable.

WORKED EXAMPLES

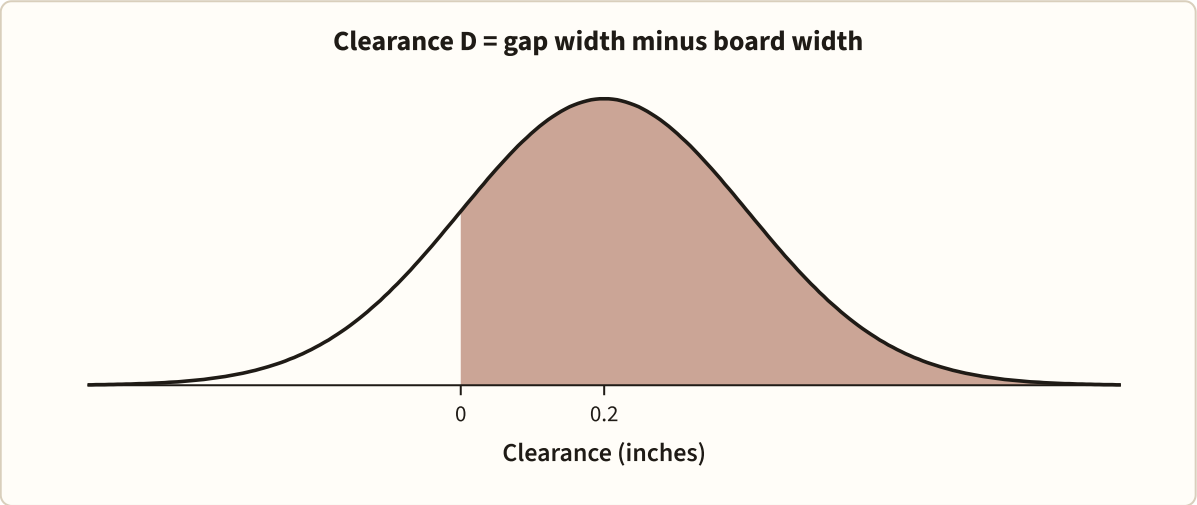
Example 1. The weight X of a loaf of bread from a small bakery is normal with mean 500 grams and standard deviation 8 grams. What is the probability that a randomly chosen loaf weighs between 490 and 516 grams?



1. Sketch $N(500, 8)$ and shade the area between 490 and 516.
2. Standardize the left end: $z = \dfrac{490 - 500}{8} = -1.25$.
3. Standardize the right end: $z = \dfrac{516 - 500}{8} = 2.00$.
4. Area to the left of 2.00 is 0.9772; area to the left of -1.25 is 0.1056.
5. Area between: $0.97725 - 0.10565 = 0.8716$ (table: $0.9772 - 0.1056 = 0.8716$).
6. Conclusion: about 87% of this bakery's loaves weigh between 490 and 516 grams.

Answer: $P(490 < X < 516) = 0.8716$

Example 2. A carpenter is fitting boards into gaps in a fence. The gap widths G are normal with mean 48.00 inches and standard deviation 0.12 inch. The board widths B are normal with mean 47.80 inches and standard deviation 0.16 inch, independent of the gaps. What is the probability that a randomly chosen board fits a randomly chosen gap (that is, the board is narrower than the gap)?



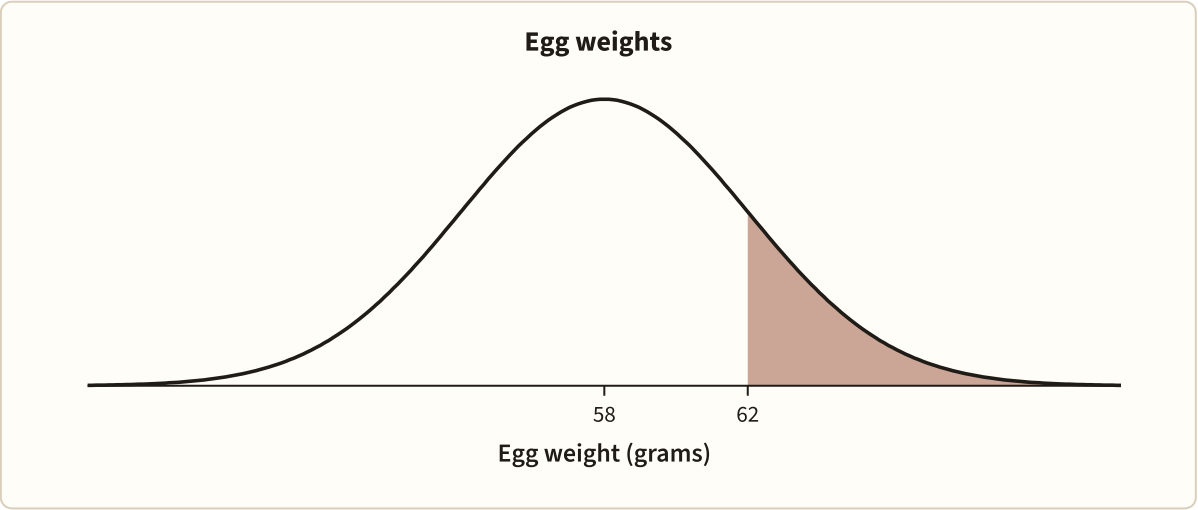
1. Define the clearance $D = G - B$. The board fits when $D > 0$.
2. Mean: $\mu_D = 48.00 - 47.80 = 0.20$ inch.
3. Variances add because G and B are independent: $\sigma_D^2 = 0.12^2 + 0.16^2 = 0.0144 + 0.0256 = 0.04$, so $\sigma_D = 0.20$ inch.
4. G and B are independent normal random variables, so D is normal: $D \sim N(0.20, 0.20)$.
5. Standardize: $z = \dfrac{0 - 0.20}{0.20} = -1.00$.
6. $P(D > 0) = P(Z > -1.00) = 1 - 0.1587 = 0.8413$.
7. Conclusion: about 84% of randomly paired boards and gaps fit, so about 16% of the time the carpenter will need to trim a board or try another one.

Answer: $P(D > 0) = 0.8413$

PRACTICE

1. EASY The rainfall in April on a farm is a continuous random variable. (a) What is the probability that the April rainfall is exactly 2 inches (to infinitely many decimal places)? (b) Explain why.

2. EASY The weights of eggs from a flock are normal with mean 58 grams and standard deviation 4 grams. Find the probability that a randomly chosen egg weighs more than 62 grams.



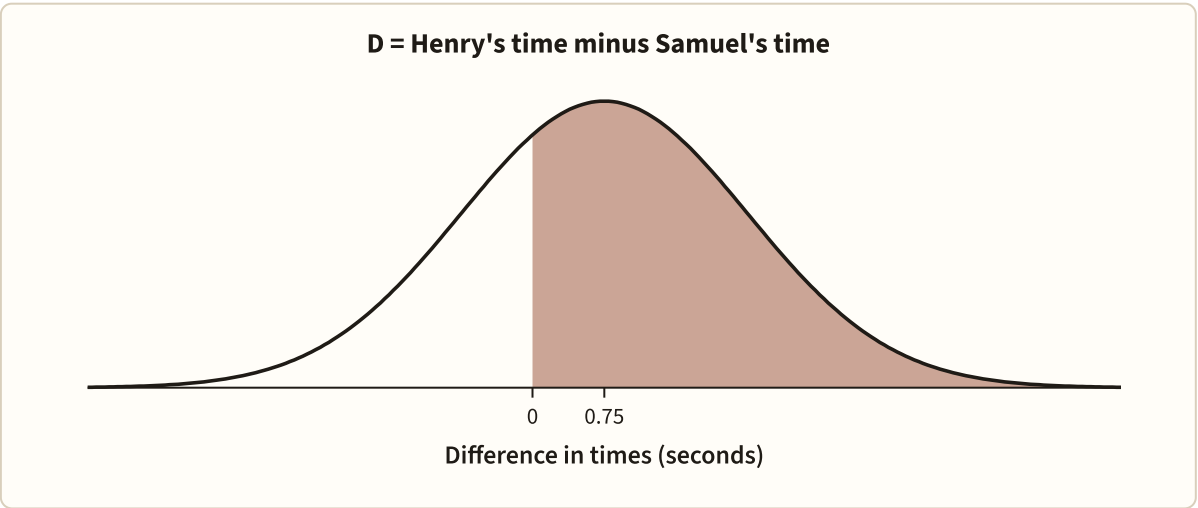
3. EASY For the eggs in Problem 2, find the probability that an egg weighs at most 52 grams.

4. EASY The weights of hay bales are normal with mean 60 pounds and standard deviation 3 pounds, independently. Describe the distribution of the total weight of two bales: give its shape, mean and standard deviation. Round the standard deviation to two decimal places.

5. MEDIUM Henry's golf drives are normal with mean 240 yards and standard deviation 16 yards. Find the probability that a drive goes between 220 and 260 yards.

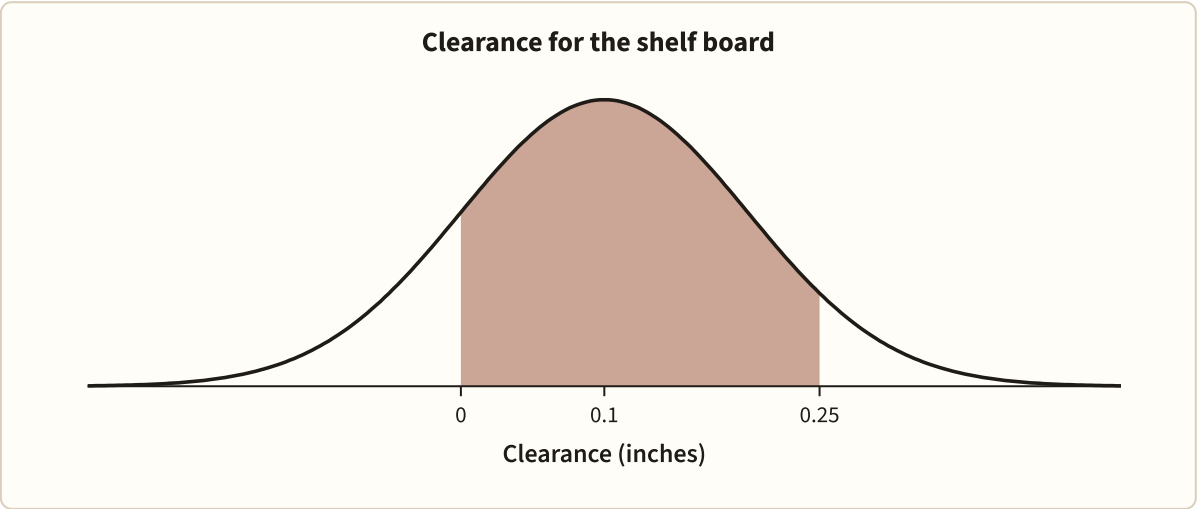
6. MEDIUM A beekeeper packs 9 jars of honey in a case. The weight of honey in each jar is normal with mean 16.2 ounces and standard deviation 0.2 ounce, independently. Find the probability that the total weight of honey in a case is less than 145.2 ounces.

7. MEDIUM In the 400-meter race, Samuel's times are normal with mean 62.0 seconds and standard deviation 1.2 seconds, and Henry's times are normal with mean 62.75 seconds and standard deviation 0.9 second, independently. Let $D =$ Henry's time $-$ Samuel's time. Find the probability that Samuel wins (finishes in less time).



8. MEDIUM In Problem 7, Ruth computes the standard deviation of D as $1.2 + 0.9 = 2.1$ seconds. Explain her mistake and why the correct value is smaller.

9. HARD A shelf board's width is normal with mean 23.90 inches and standard deviation 0.06 inch. The opening in the bookcase is normal with mean 24.00 inches and standard deviation 0.08 inch, independent of the board. Let D = opening − board. Find the probability that the board fits with a clearance of less than 0.25 inch, $P(0 < D < 0.25)$.



10. HARD Each morning a dairy farmer milks the cows (time normal, mean 40 minutes, standard deviation 4 minutes) and then feeds the calves (time normal, mean 25 minutes, standard deviation 3 minutes), independently. Find the probability that the two chores together take less than 72 minutes.

11. HARD Leghorn eggs weigh $N(60, 3)$ grams and bantam eggs weigh $N(45, 4)$ grams. One egg of each kind is chosen at random, independently. Find the probability that the Leghorn egg is more than 22 grams heavier than the bantam egg.

12. HARD In Problem 10, is it reasonable to assume that the milking time and the feeding time are independent? Describe a situation that could make them dependent, and say how that would affect the calculation.

13. CHALLENGE For the chores in Problem 10, how much time should the farmer set aside so that both chores are finished in that time on 95% of mornings? Round to the nearest tenth of a minute.

14. CHALLENGE Bookcase openings are normal with mean 24.00 inches and standard deviation 0.08 inch. A shop can cut shelf boards with standard deviation 0.06 inch at any mean width m it chooses, independently of the openings. What mean width m makes 99% of boards fit their openings? Round to the nearest hundredth of an inch.

COMMON MISTAKES

- Adding standard deviations when combining normal random variables. Add the variances, then take the square root; for 9 jars each with $\sigma = 0.2$, the total has $\sigma = \sqrt{9(0.04)} = 0.6$, not 1.8.

- Forgetting to turn a "which is larger?" question into a statement about a difference. Define D in words, decide whether you need $D > 0$ or $D < 0$, then standardize.
- Standardizing a total with the standard deviation of one item. The z-score for a total must use the mean and standard deviation of the TOTAL.
- Assuming independence without thinking. Quantities affected by the same weather, the same worker or the same machine setting may move together, and then variances do not simply add.

ANSWERS TO PRACTICE

1. (a) 0 (b) For a continuous random variable, probability is area under the density curve, and a single value is a line with no width, so its area is 0. Rainfall near 2 inches, such as between 1.9 and 2.1 inches, has a positive probability.
2. $z = \dfrac{62 - 58}{4} = 1.00$; $P(X > 62) = 1 - 0.8413 = 0.1587$
3. $z = \dfrac{52 - 58}{4} = -1.50$; $P(X \le 52) = 0.0668$
4. Normal, with mean $60 + 60 = 120$ pounds and standard deviation $\sqrt{3^2 + 3^2} = \sqrt{18} \approx 4.24$ pounds
5. $z = \dfrac{220 - 240}{16} = -1.25$ and $z = \dfrac{260 - 240}{16} = 1.25$; area $= 0.89435 - 0.10565 = 0.7887$
6. Total: $\mu = 9(16.2) = 145.8$ ounces; $\sigma = \sqrt{9(0.2^2)} = \sqrt{0.36} = 0.6$ ounce; normal. $z = \dfrac{145.2 - 145.8}{0.6} = -1.00$; $P = 0.1587$
7. $\mu_D = 62.75 - 62.0 = 0.75$ s; $\sigma_D = \sqrt{1.2^2 + 0.9^2} = \sqrt{1.44 + 0.81} = \sqrt{2.25} = 1.5$ s. Samuel wins when $D > 0$: $z = \dfrac{0 - 0.75}{1.5} = -0.50$; $P(D > 0) = 1 - 0.3085 = 0.6915$
8. Standard deviations do not add; variances do. The value 1.2 + 0.9 = 2.1 seconds would be right only if every slow race by Henry came with a fast race by Samuel (and the other way round), so that their deviations always combined in the same direction. Because the two times vary independently, the runners are often both slow or both fast, and then their deviations partly cancel in the difference. So $\sigma_D = \sqrt{1.44 + 0.81} = 1.5$ seconds, which is less than $1.2 + 0.9 = 2.1$ seconds.
9. $\mu_D = 24.00 - 23.90 = 0.10$ in; $\sigma_D = \sqrt{0.0036 + 0.0064} = \sqrt{0.01} = 0.10$ in. $z = \dfrac{0 - 0.10}{0.10} = -1.00$ and $z = \dfrac{0.25 - 0.10}{0.10} = 1.50$; area $= 0.93319 - 0.15866 = 0.7745$
10. Total: $\mu = 40 + 25 = 65$ min; $\sigma = \sqrt{16 + 9} = 5$ min; normal. $z = \dfrac{72 - 65}{5} = 1.40$; $P = 0.9192$
11. $D = L - B$: $\mu_D = 60 - 45 = 15$ g; $\sigma_D = \sqrt{9 + 16} = 5$ g; normal. $z = \dfrac{22 - 15}{5} = 1.40$; $P(D > 22) = 1 - 0.9192 = 0.0808$

12. Often it is reasonable, since the two chores involve different animals. But they could be linked: on a cold, icy morning both chores may run long, or a farmer who is rushed may hurry through both. Then long milking times would go with long feeding times, the total would vary more than $\sqrt{16+9} = 5$ minutes, and the probability we found would be too high.
13. The 95th percentile of $N(65, 5)$: $z = 1.645$, so $x = 65 + 1.645(5) = 73.225 \approx 73.2$ minutes
14. $D = \text{opening} - \text{board}$ has $\mu_D = 24.00 - m$ and $\sigma_D = \sqrt{0.0064 + 0.0036} = 0.10$. We need $P(D > 0) = 0.99$, so 0 must be at the 1st percentile of D: $z = -2.326$. Then $\dfrac{0 - (24.00 - m)}{0.10} = -2.326$, so $24.00 - m = 0.2326$ and $m \approx 23.77$ inches.

Unit 7 test

You have 60 minutes. You may use a scientific calculator and the standard normal table at the back of the course. Show your work: write the formula, the numbers you substitute and the result, and answer in context where asked. Keep full precision until the last step. Unless a problem says otherwise, round probabilities to 4 decimal places. The test is worth 100 points.

1. **5 PTS** Classify each random variable as discrete or continuous. (a) The number of foals born at a horse farm in a year. (b) The length of a carrot pulled from a garden.

2. **6 PTS** Let X = the number of cars that arrive at a roadside farm stand in a 10-minute period. The table gives the distribution with one probability missing. (a) Find $P(X = 3)$. (b) Find the probability that at least 2 cars arrive.

   Cars arriving at a farm stand in 10 minutes

Cars, x	P(x)
0	0.1
1	0.25
2	0.3
3	?
4	0.1

3. **7 PTS** For the farm stand distribution in Problem 2, find the mean and standard deviation of X. Round the standard deviation to two decimal places.

   Cars arriving at a farm stand in 10 minutes

Cars, x	P(x)
0	0.1
1	0.25
2	0.3
3	?
4	0.1

4. **5 PTS** Interpret the mean from Problem 3 in context.

5. **8 PTS** A spinner has three equal sections numbered 1, 2 and 3. It is spun twice, and X = the sum of the two numbers. (a) Find the probability distribution of X; give $P(2), P(3), P(4), P(5), P(6)$ rounded to 4 decimal places. (b) Find $P(X > 4)$.

6. **7 PTS** A 4-H club runs a ring toss at the county fair to raise money. It costs \$2 to play. From past fairs, 5% of players win a \$20 prize and 20% win a \$4 prize; the table shows a player's net gain. (a) Find a player's expected net gain. (b) About how much will the club raise from 400 plays?

Ring toss at the county fair: player's net gain

Result	Net gain (dollars)	Probability
Large prize	18	0.05
Small prize	2	0.2
No prize	-2	0.75

7. **7 PTS** A store offers a \$90 two-year repair plan on a sewing machine. The table shows the repair costs the store would pay and their probabilities. (a) Find the expected repair cost. (b) Explain why the store can expect to make money on the plan, and why a family might still choose to buy it.

Repair costs for a sewing machine over two years

Repair cost (dollars)	Probability
0	0.8
300	0.15
600	0.05

8. **6 PTS** At a small sawmill, the number of boards X cut in an hour has mean 40 and standard deviation 5. Each board sells for \$8, and running the mill costs \$60 per hour, so the hourly profit is $P = 8X - 60$. Find the mean and standard deviation of P.

9. **6 PTS** A family hikes up a mountain and back down. The uphill time U and the downhill time D are independent, with the means and standard deviations shown. Find the mean and standard deviation of the total hiking time $U + D$.

Mountain hike times (minutes)

Part of hike	Mean	Standard deviation
Uphill, U	95	12
Downhill, D	60	5

10. **7 PTS** For the hike in Problem 9, let $U - D$ be how much longer the uphill part takes than the downhill part. (a) Find the mean and standard deviation of $U - D$. (b) Henry says the standard deviation should be $12 - 5 = 7$ minutes. Explain why he is wrong.

Mountain hike times (minutes)

Part of hike	Mean	Standard deviation
Uphill, U	95	12
Downhill, D	60	5

11. **7 PTS** Birth weights of calves on a ranch are normal with mean 80 pounds and standard deviation 6 pounds. Find the probability that a newborn calf weighs between 74 and 92 pounds.

12. **8 PTS** Grace makes bars of soap whose weights are normal with mean 4.0 ounces and standard deviation 0.25 ounce, independently. She packs 16 bars in a box that weighs exactly 6.0 ounces empty. Find the probability that a filled box weighs more than 71.5 ounces.

13. **5 PTS** In Problem 12, the variance of the total weight of the 16 bars is 16 times the variance of one bar, but the standard deviation is only 4 times as large. Explain why the total does not vary 16 times as much as a single bar.

14. **8 PTS** In an orchard, Honeycrisp apples weigh $N(220, 15)$ grams and Gala apples weigh $N(180, 20)$ grams. One apple of each variety is picked at random, independently. Find the probability that the Gala apple is heavier than the Honeycrisp apple.

15. **8 PTS** A carpenter fits wooden pegs into drilled holes. Hole diameters are normal with mean 1.010 inches and standard deviation 0.003 inch; peg diameters are normal with mean 1.000 inch and standard deviation 0.004 inch, independently. Let C = hole $-$ peg. (a) Find the probability that a randomly chosen peg fits its hole ($C > 0$). (b) Find the probability that the fit is too loose, $C > 0.015$ inch.

UNIT 7 TEST: ANSWER KEY

1. (a) Discrete (a count) (b) Continuous (a length can be any value in an interval)
2. (a) $1 - (0.10 + 0.25 + 0.30 + 0.10) = 1 - 0.75 = 0.25$ (b) $P(X \geq 2) = 0.30 + 0.25 + 0.10 = 0.65$
3. $\mu_X = 0(0.10) + 1(0.25) + 2(0.30) + 3(0.25) + 4(0.10) = 0 + 0.25 + 0.60 + 0.75 + 0.40 = 2$ cars. $\sigma_X^2 = 4(0.10) + 1(0.25) + 0(0.30) + 1(0.25) + 4(0.10) = 0.4 + 0.25 + 0 + 0.25 + 0.4 = 1.3$; $\sigma_X = \sqrt{1.3} \approx 1.14$ cars
4. Over many 10-minute periods, an average of about 2 cars arrive at the farm stand per period. In any one period the number may be anywhere from 0 to 4.
5. There are $3 \times 3 = 9$ equally likely outcomes. Sum 2: 1 outcome; 3: 2; 4: 3; 5: 2; 6: 1. (a) $P(2) = \frac{1}{9} \approx 0.1111$, $P(3) = \frac{2}{9} \approx 0.2222$, $P(4) = \frac{3}{9} \approx 0.3333$, $P(5) \approx 0.2222$, $P(6) \approx 0.1111$ (b) $P(X > 4) = \frac{2 + 1}{9} = \frac{3}{9} \approx 0.3333$
6. (a) $18(0.05) + 2(0.20) + (-2)(0.75) = 0.90 + 0.40 - 1.50 = -\$0.20$ per play (b) The club gains about \$0.20 per play: $400 \times \$0.20 = \$80$
7. (a) $0(0.80) + 300(0.15) + 600(0.05) = 0 + 45 + 30 = \$75$ (b) The store collects \$90 but pays out only \$75 per plan on average, so over many plans it gains about \$15 each. A family has only one machine, and a plan protects them from a sudden \$300 or \$600 bill; they pay a little more on average in exchange for certainty.
8. $\mu_P = 8(40) - 60 = \$260$; $\sigma_P = 8(5) = \$40$
9. $\mu = 95 + 60 = 155$ minutes; $\sigma = \sqrt{12^2 + 5^2} = \sqrt{144 + 25} = \sqrt{169} = 13$ minutes
10. (a) $\mu = 95 - 60 = 35$ minutes; $\sigma = \sqrt{144 + 25} = 13$ minutes (b) Variances add for a difference of independent variables, just as for a sum. The difference is largest when the climb is slow and the descent quick, so both sources of variation appear in it; subtracting the times does not cancel their variability.
11. $z = \frac{74 - 80}{6} = -1.00$ and $z = \frac{92 - 80}{6} = 2.00$; area $= 0.97725 - 0.15866 = 0.8186$
12. Total: $\mu = 16(4.0) + 6.0 = 70.0$ ounces; $\sigma = \sqrt{16(0.25^2)} = \sqrt{1} = 1.0$ ounce; normal. $z = \frac{71.5 - 70.0}{1.0} = 1.50$; $P = 1 - 0.9332 = 0.0668$
13. The bars vary independently, so some are a little heavy and others a little light, and their errors partly cancel in the total. Variances add, giving $16(0.0625) = 1$; the standard deviation is the square root, $\sqrt{16} \times 0.25 = 4 \times 0.25 = 1$ ounce, not $16 \times 0.25 = 4$ ounces.
14. Let $D =$ Honeycrisp $-$ Gala: $\mu_D = 220 - 180 = 40$ g; $\sigma_D = \sqrt{225 + 400} = \sqrt{625} = 25$ g; normal. The Gala is heavier when $D < 0$: $z = \frac{0 - 40}{25} = -1.60$; $P(D < 0) = 0.0548$

15. $\mu_C = 1.010 - 1.000 = 0.010$ in; $\sigma_C = \sqrt{0.000009 + 0.000016} = \sqrt{0.000025} = 0.005$ in; normal. (a) $z = \dfrac{0 - 0.010}{0.005} = -2.00$; $P(C > 0) = 1 - 0.0228 = 0.9772$ (b) $z = \dfrac{0.015 - 0.010}{0.005} = 1.00$; $P(C > 0.015) = 1 - 0.8413 = 0.1587$

UNIT 8

Binomial and Geometric Distributions

Many chance questions on a farm, in sports or at home are counting questions: how many of 20 seeds will sprout, how many free throws out of 10 will go in, how many rolls until a 6 appears. In this unit you meet the two most important models for such counts. The binomial distribution counts successes in a fixed number of independent trials, and the geometric distribution counts the trials until the first success. For each you will learn not only the formulas but why every part of them is there.

The unit builds directly on Unit 6 (the multiplication rule, complements and combinations) and Unit 7 (random variables, their means and standard deviations, and the fact that variances add). In Lesson 8.3 you will see the normal curve appear out of the binomial, just as it did for Abraham de Moivre in 1733. That discovery is the bridge to Unit 9, where the same idea explains how sample proportions behave, and to all the inference that follows.

PREREQUISITES

- Use the multiplication rule for independent events and the complement rule (Lessons 6.2 and 6.4).
- Compute combinations $\binom{n}{k}$ (Lesson 6.6).
- Find the mean and standard deviation of a discrete random variable from its distribution (Lessons 7.1 and 7.2).
- Explain that means add and that variances add for independent random variables (Lesson 7.3).
- Find normal probabilities and work backward from an area to a value (Lessons 3.4 and 3.5).

MASTERY CHECKLIST

☐ I can check the four BINS conditions and decide whether a count is binomial.
☐ I can explain why the binomial formula contains $\binom{n}{k}$, p^k and $(1-p)^{n-k}$.
☐ I can compute binomial probabilities $P(X = k)$ by hand and with a calculator.
☐ I can find cumulative probabilities such as $P(X \le 2)$ and $P(X \ge 3)$, using the complement when it is quicker.
☐ I can compute and interpret $\mu = np$ and $\sigma = \sqrt{np(1-p)}$, and explain why np makes sense.
☐ I can describe the shape of a binomial distribution from n and p, and check the 10% condition.
☐ I can check the Large Counts condition and use the normal approximation to the binomial.
☐ I can recognize a geometric setting and compute $P(X = k)$, $P(X > k)$ and $P(X \le k)$.

☐ I can compute and interpret the mean $\mu = 1/p$ of a geometric distribution and explain why it is always skewed right.

LESSON 8.1 · 2 SESSIONS OF 45 MINUTES

Binomial Settings and the Binomial Formula

OBJECTIVES

- Decide whether a chance process is a binomial setting by checking the four BINS conditions.
- Explain why the binomial formula contains $\binom{n}{k}$, p^k and $(1-p)^{n-k}$.
- Compute binomial probabilities $P(X = k)$ by hand and interpret them in context.

VIDEO

WATCH

Binomial Probability (1 of 2: Preliminary Example w/ Probability Tree)

Eddie Woo · 13:26

When to watch: Before Example 1: building the binomial formula from a tree.

https://www.youtube.com/watch?v=zc9QKwbuUQk

BACKUP

Binomial Probability (2 of 2: Further Example & General Case)

Eddie Woo · 7:20

When to watch: After Example 1: the general binomial formula.

https://www.youtube.com/watch?v=KFrmwThf2Vk

KEY TERMS

Trial	One repetition of a chance process, such as one seed planted, one free throw or one apple inspected.
Success and failure	The two possible outcomes of a trial. A success is simply the outcome we are counting; it does not have to be a good thing (a blemished apple can be a success).
Binomial setting	A chance process with Binary outcomes, Independent trials, a fixed Number n of trials, and the Same probability p of success on every trial (BINS).
Binomial random variable	X = the number of successes in a binomial setting. Its possible values are 0, 1, 2, ..., n. We write X is B(n, p).
Binomial coefficient	The number $\binom{n}{k}$ of different ways to choose which k of the n trials are the successes.

CORE IDEAS

Many questions in statistics are counting questions. Of 20 seeds planted, how many will sprout? Of 4 free throws, how many will Ruth make? Of 6 apples, how many will have a blemish? When the chance process has a certain structure, one formula answers every question like these. In this lesson you will build that formula yourself, so that you know why each piece is there.

The structure has four conditions, easy to remember as **BINS**. **B**inary: each trial has exactly two outcomes, success or failure. **I**ndependent: the outcome of one trial does not change the probabilities for another. **N**umber: the number of trials n is fixed in advance. **S**ame probability: the probability of success p is the same on every trial. When all four hold, the count X of successes is a **binomial random variable** with parameters n and p, written B(n, p). Watch for settings that fail: rolling a die until a 6 appears has no fixed n (that is Lesson 8.4), and drawing 3 names from a choir of 10 without putting them back changes p after every draw, so the draws are not independent.

Let us build the formula with a small case. A gardener plants 3 seeds, and each sprouts with probability 0.8, independently. Write S for a seed that sprouts and F for one that does not. There are 8 possible outcomes. Because the seeds are independent, the probability of each outcome is a product (the multiplication rule from Lesson 6.4).

Three seeds, each sprouts with probability 0.8

Outcome	Product	Probability	X (number that sprout)
SSS	0.8 × 0.8 × 0.8	0.512	3
SSF	0.8 × 0.8 × 0.2	0.128	2
SFS	0.8 × 0.2 × 0.8	0.128	2
FSS	0.2 × 0.8 × 0.8	0.128	2
SFF	0.8 × 0.2 × 0.2	0.032	1
FSF	0.2 × 0.8 × 0.2	0.032	1
FFS	0.2 × 0.2 × 0.8	0.032	1
FFF	0.2 × 0.2 × 0.2	0.008	0
Total		1	

To find $P(X = 2)$, add the probabilities of the outcomes with exactly 2 sprouts: SSF, SFS and FSS. Each contains two S's and one F, so each has probability $p \times p \times (1 - p) = 0.8 \times 0.8 \times 0.2 = 0.128$; changing the order does not change the product. So $P(X = 2) = 3 \times 0.128 = 0.384$. Every binomial probability follows this pattern: (the number of arrangements) $\times$ (the probability of any one arrangement).

How many arrangements are there in general? An arrangement with k successes is decided by choosing which k of the n trial positions hold an S. The order in which we choose the positions does

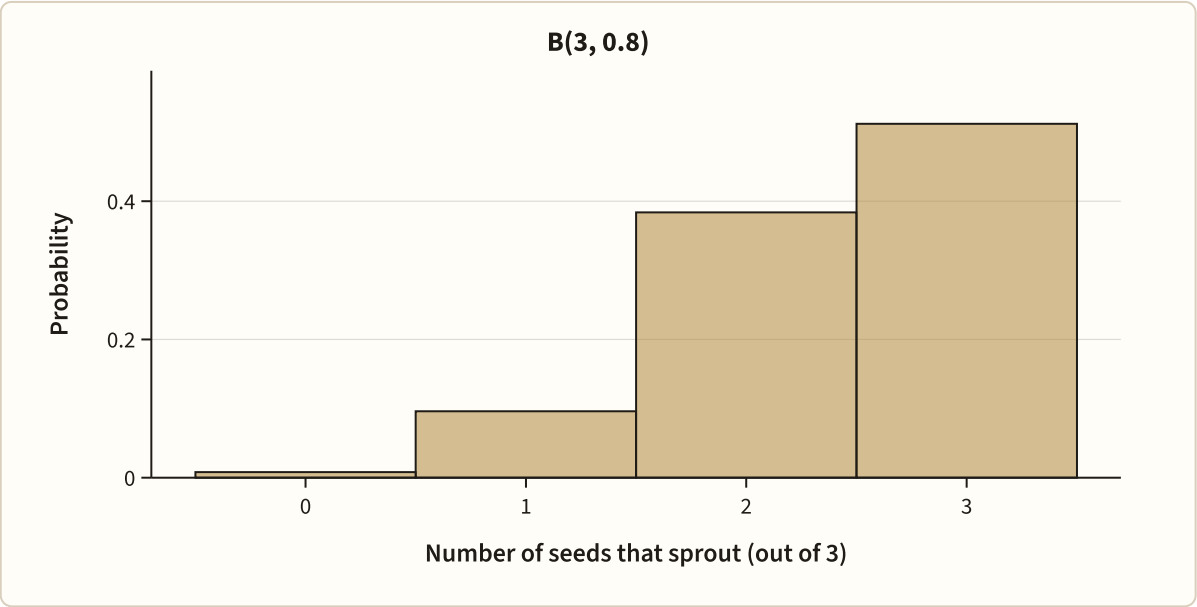
not matter, so the count is the number of combinations from Lesson 6.6: $\binom{n}{k} = \dfrac{n!}{k!(n-k)!}$. For 3 seeds and 2 sprouts, $\binom{3}{2} = 3$, matching SSF, SFS, FSS. These numbers are the entries of Pascal's triangle.

Values of C(n, k): Pascal's triangle

n	k = 0	k = 1	k = 2	k = 3	k = 4	k = 5	k = 6
1	1	1					
2	1	2	1				
3	1	3	3	1			
4	1	4	6	4	1		
5	1	5	10	10	5	1	
6	1	6	15	20	15	6	1

Now every piece of the binomial formula $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$ has a reason. The factor p^k is the k successes, multiplied because the trials are independent. The factor $(1-p)^{n-k}$ is the n − k failures, multiplied for the same reason. The factor $\binom{n}{k}$ counts the different orders, and we add (multiply by the count) because no two orders can happen at once. The exponents always add to n, since every trial is either a success or a failure.

Using the formula for k = 0, 1, 2, 3 gives the whole probability distribution of X for the 3 seeds: 0.008, 0.096, 0.384 and 0.512, which add to 1, as they must. The probability histogram shows that 3 sprouts is the most likely result, which makes sense when each seed sprouts 80% of the time. A graphing calculator does the same arithmetic with binompdf(n, p, k), but you should always be able to say what it is computing.



KEY FORMULAS

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}, \quad k = 0, 1, 2, \ldots, n$$

$$\binom{n}{k} = \frac{n!}{k!(n-k)!}$$

BINS: Binary outcomes, Independent trials, fixed Number n of trials, Same probability p of success.

From history. Trials with two outcomes are often called Bernoulli trials, after the Swiss mathematician Jacob Bernoulli, whose book Ars Conjectandi (published in 1713, after his death) studied them and proved the first law of large numbers. The numbers $\binom{n}{k}$ are the entries of the arithmetic triangle that Blaise Pascal wrote about, which is why it is called Pascal's triangle.

Practice habit. Before you use the binomial formula, check BINS out loud and write down three things: what counts as a success, the value of n, and the value of p.

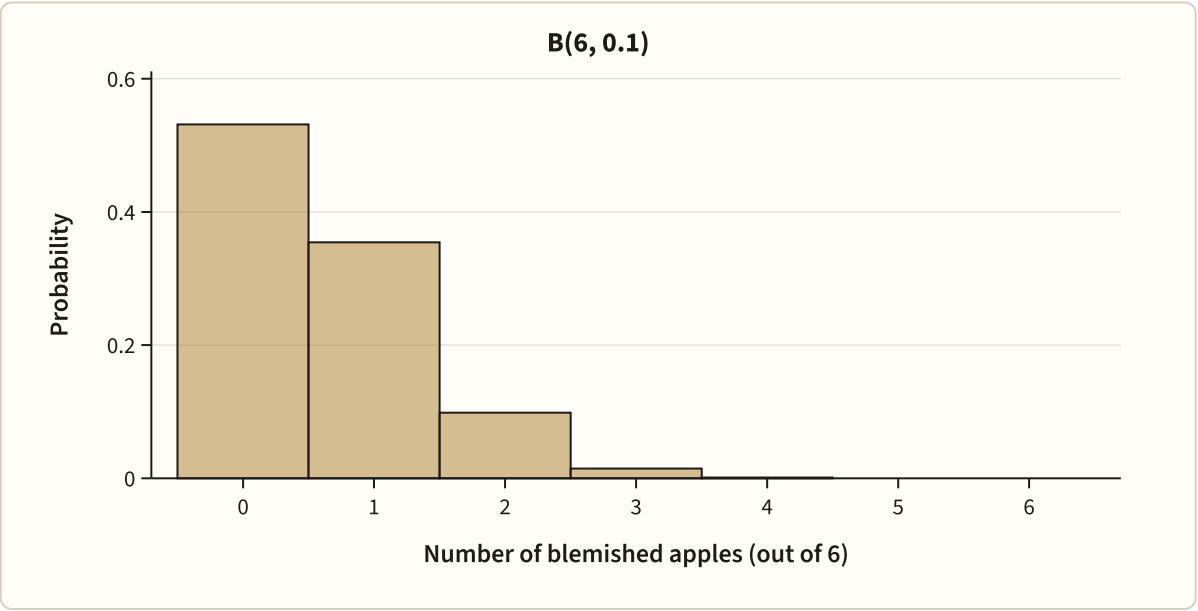
WORKED EXAMPLES

Example 1. Ruth makes 70% of her free throws. In a game she shoots 4 free throws. Assuming her shots are independent, find the probability that she makes exactly 3 of them. Round to 4 decimal places.

1. Check BINS. Binary: each shot is made (success) or missed. Independent: we assume one shot does not affect the next. Number: n = 4 shots, fixed. Same probability: p = 0.7 on every shot. So X = the number made is B(4, 0.7).
2. Count the arrangements: $\binom{4}{3} = 4$ (the one miss can be the 1st, 2nd, 3rd or 4th shot).
3. Probability of one arrangement, such as make, make, make, miss: $(0.7)^3(0.3)^1 = 0.343 \times 0.3 = 0.1029$.
4. Multiply by the number of arrangements: $P(X = 3) = 4 \times 0.1029 = 0.4116$.
5. Conclusion: in about 41% of games in which Ruth shoots 4 free throws, she makes exactly 3 of them.

Answer: $P(X = 3) = \binom{4}{3}(0.7)^3(0.3)^1 = 4 \times 0.1029 = 0.4116$

Example 2. In a large bin of apples from Henry's orchard, 10% of the apples have a blemish. An inspector picks 6 apples at random. Let X = the number of blemished apples. (a) Find $P(X = 1)$. (b) Find $P(X = 2)$. Round to 4 decimal places.



1. Success is the thing we count, a blemished apple, so p = 0.1 and n = 6. The bin is large, so taking out a few apples hardly changes the 10%; we treat the picks as independent. X is B(6, 0.1).
2. (a) $P(X = 1) = \binom{6}{1}(0.1)^1(0.9)^5 = 6 \times 0.1 \times 0.59049$.
3. $= 0.354294 \approx 0.3543$.
4. (b) $P(X = 2) = \binom{6}{2}(0.1)^2(0.9)^4 = 15 \times 0.01 \times 0.6561$.
5. $= 0.098415 \approx 0.0984$.
6. Conclusion: about 35% of random samples of 6 apples contain exactly one blemished apple, and only about 10% contain exactly two. The probability histogram shows that 0 or 1 blemished apples is by far the most common result.

Answer: (a) $P(X = 1) = 6(0.1)(0.9)^5 \approx 0.3543$ (b) $P(X = 2) = 15(0.1)^2(0.9)^4 \approx 0.0984$

PRACTICE

1. EASY A farmer plants 20 seeds of a bean variety whose seeds sprout 85% of the time, independently of one another. Let X = the number of seeds that sprout. (a) Is X a binomial random variable? (b) What is n? (c) What is p?
2. EASY In a board game, Grace must roll a 6 to move her first piece onto the board. She rolls a fair die again and again until she gets a 6. Let X = the number of rolls she needs. Is X a binomial random variable? If not, which BINS condition fails?
3. EASY Compute $\binom{5}{2}$ and explain what it counts when a batter has 5 at-bats and gets exactly 2 hits.

4. EASY Henry flips a fair coin 4 times. Find the probability that he gets exactly 2 heads.

5. MEDIUM Grace hits the bull's-eye on 20% of her arrows, independently from arrow to arrow. She shoots 5 arrows. Find the probability that she hits the bull's-eye exactly once. Round to 4 decimal places.

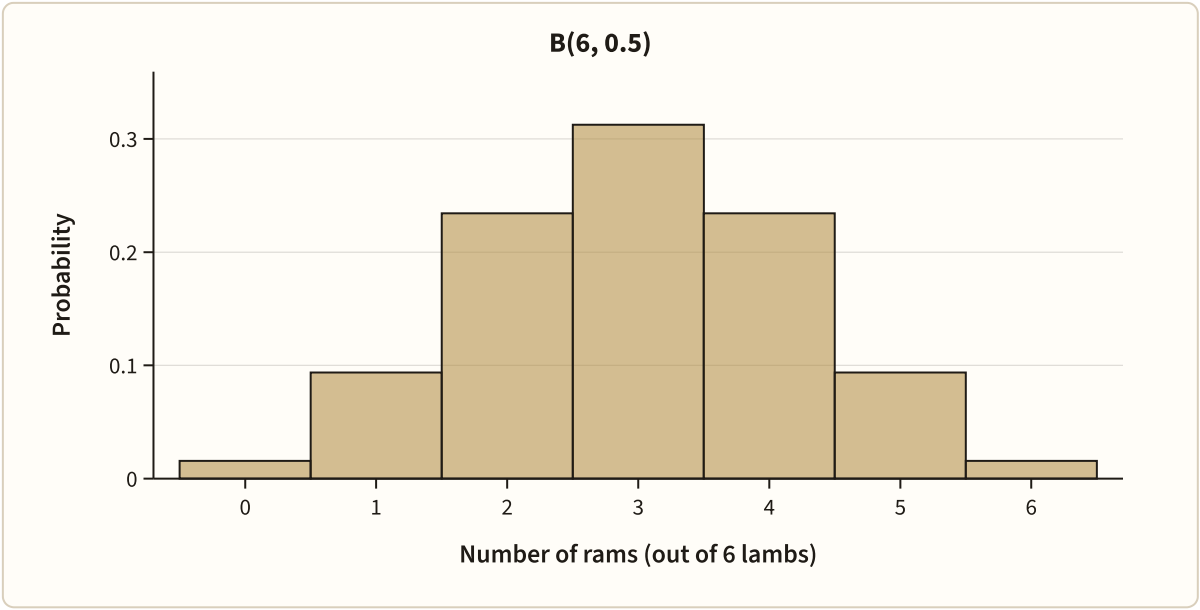
6. MEDIUM Samuel gets a hit in 30% of his at-bats. Assume his at-bats are independent. In a game with 4 at-bats, find the probability that he gets exactly 2 hits. Round to 4 decimal places.

7. MEDIUM In a large orchard, 10% of the apples have a blemish. A picker chooses 5 apples at random. Find the probability that none of them has a blemish. Round to 4 decimal places.

8. MEDIUM A small church choir has 10 members: 4 tenors and 6 others. The director draws 3 names from a hat, without putting names back, to choose a trio. Let X = the number of tenors chosen. Is X a binomial random variable? Explain.

9. MEDIUM Three seeds each sprout with probability 0.8, independently. Ruth writes $P(X = 2) = (0.8)^2(0.2) = 0.128$. (a) Explain her mistake. (b) Give the correct probability.

10. HARD Six lambs are born on the Miller farm this spring. Each lamb is equally likely to be a ram (male) or a ewe lamb (female), independently of the others. The probability histogram of X = the number of rams is shown. (a) Find $P(X = 3)$. (b) Explain why the histogram is symmetric.



11. HARD Henry makes 60% of his free throws, independently from shot to shot. He shoots 3 free throws. Let X = the number he makes. Find $P(X = 0)$, $P(X = 1)$, $P(X = 2)$ and $P(X = 3)$, and check that they add to 1.

12. HARD A baseball pitcher throws a first-pitch strike to 70% of the batters he faces. Assume batters are independent. Of the next 5 batters, find the probability that exactly 3 get a first-pitch strike. Round to 4 decimal places.

13. **HARD** A seed catalog says that if you plant 3 seeds of a new squash variety, the probability that all 3 sprout is 0.512. Seeds sprout independently with the same probability p. (a) Find p. (b) Find the probability that none of the 3 seeds sprouts.

14. **CHALLENGE** Samuel rolls a fair die 6 times. Is he more likely to roll no sixes at all or exactly one six? Find both probabilities, rounded to 4 decimal places, and explain why the answer is not obvious.

COMMON MISTAKES

- Leaving out $\binom{n}{k}$. The product $p^k(1-p)^{n-k}$ is the probability of ONE particular order; multiply by the number of orders.
- Using the binomial formula when it does not apply. Check BINS first: rolling until a success has no fixed n, and drawing without replacement from a small group makes the trials dependent.
- Putting the exponents on the wrong numbers. The exponent on p is the number of successes k and the exponent on 1 − p is the number of failures n − k; they always add to n.
- Defining success carelessly. If you are counting blemished apples, p is the probability of a blemish (0.1), not the probability of a good apple (0.9).

ANSWERS TO PRACTICE

1. (a) Yes: each seed sprouts or not (binary), the seeds sprout independently, the number of seeds is fixed, and every seed has the same chance. (b) n = 20 (c) p = 0.85
2. No. The number of trials is not fixed in advance (the N condition fails): X counts the rolls until the first success, not the successes in a fixed number of rolls. This is a geometric setting (Lesson 8.4).
3. $\binom{5}{2} = \dfrac{5!}{2!\,3!} = \dfrac{120}{2 \times 6} = 10$. It counts the 10 different ways to choose which 2 of the 5 at-bats were the hits (for example hit, hit, out, out, out or out, hit, out, out, hit).
4. $\binom{4}{2}(0.5)^2(0.5)^2 = 6 \times 0.0625 = 0.375$
5. $\binom{5}{1}(0.2)^1(0.8)^4 = 5 \times 0.2 \times 0.4096 = 0.4096$
6. $\binom{4}{2}(0.3)^2(0.7)^2 = 6 \times 0.09 \times 0.49 = 0.2646$
7. Success = blemished apple, p = 0.1, n = 5. $P(X = 0) = \binom{5}{0}(0.1)^0(0.9)^5 = 1 \times 1 \times 0.59049 \approx 0.5905$
8. No. The draws are not independent and p does not stay the same. The first name is a tenor with probability 4/10 = 0.4, but if the first name is a tenor, the second is a tenor with probability 3/9, about 0.33. Each draw changes the chances for the next.
9. (a) $(0.8)^2(0.2) = 0.128$ is the probability of one particular order, such as sprout, sprout, fail. Two sprouts can happen in $\binom{3}{2} = 3$ different orders (SSF, SFS, FSS), and she must add all of them. (b) $P(X = 2) = 3 \times 0.128 = 0.384$

10. (a) $\binom{6}{3}(0.5)^3(0.5)^3 = 20 \times \dfrac{1}{64} = 0.3125$ (b) Because p = 0.5, rams and ewe lambs are equally likely, so k rams has the same probability as k ewe lambs, that is, 6 − k rams: $P(X = k) = P(X = 6 - k)$. The bars on the left mirror the bars on the right.
11. $P(X = 0) = (0.4)^3 = 0.064$; $P(X = 1) = 3(0.6)(0.4)^2 = 0.288$; $P(X = 2) = 3(0.6)^2(0.4) = 0.432$; $P(X = 3) = (0.6)^3 = 0.216$. Check: $\sum P(x) = 0.064 + 0.288 + 0.432 + 0.216 = 1$.
12. $\binom{5}{3}(0.7)^3(0.3)^2 = 10 \times 0.343 \times 0.09 = 0.3087$
13. (a) $P(X = 3) = p^3 = 0.512$, so $p = \sqrt[3]{0.512} = 0.8$. (b) $P(X = 0) = (1 - 0.8)^3 = (0.2)^3 = 0.008$
14. $P(X = 0) = \left(\dfrac{5}{6}\right)^6 \approx 0.3349$; $P(X = 1) = 6\left(\dfrac{1}{6}\right)\left(\dfrac{5}{6}\right)^5 = \left(\dfrac{5}{6}\right)^5 \approx 0.4019$. Exactly one six is more likely. It is not obvious because any one particular order with one six, such as six first, is less likely than no sixes; but there are 6 different places the six can appear, and those 6 orders together outweigh the single all-fail order.

LESSON 8.2 · 2 SESSIONS OF 45 MINUTES

Binomial Mean, Standard Deviation and Cumulative Probabilities

OBJECTIVES

- Compute cumulative binomial probabilities such as $P(X \leq 2)$ and $P(X \geq 3)$ by adding terms or by using the complement.
- Compute and interpret the mean $\mu = np$ and standard deviation $\sigma = \sqrt{np(1-p)}$ of a binomial random variable, and explain why np makes sense.
- Describe the shape of a binomial distribution from the values of n and p.
- Check the 10% condition before treating sampling without replacement as binomial.

VIDEO

WATCH

Binomial Setting & Binomial Distribution in Statistics Pt 2

ProfRobBob · 12:14

When to watch: Before Example 1: binomial mean and standard deviation, and cumulative probabilities.

https://www.youtube.com/watch?v=P90cT7BN0EY

BACKUP

The Binomial Distribution: Mathematically Deriving the Mean and Variance

jbstatistics · 13:54

When to watch: Enrichment: deriving the binomial mean and variance (uses more algebra than the lesson).

https://www.youtube.com/watch?v=8fqkQRjcR1M

KEY TERMS

Cumulative probability	The probability that X is at most some value, $P(X \leq k)$, found by adding the probabilities of 0, 1, ..., k.
Complement	Everything that is not the event. Since all the probabilities add to 1, $P(\text{event}) = 1 - P(\text{not the event})$.
Binomial mean	$\mu = np$, the long-run average number of successes in many repetitions of n trials.
Binomial standard deviation	$\sigma = \sqrt{np(1-p)}$, the typical distance between the number of successes and the mean np.

10% condition When we sample without replacement, we may treat the draws as independent (binomial) if the sample size n is at most 10% of the population size N.

CORE IDEAS

Real questions are often about a range of values: at least 4 of 5 seeds sprout, at most 2 apples are blemished. Translate the words into an inequality first. "At most 2" means $X \le 2$ (0, 1 or 2); "at least 3" means $X \ge 3$; "fewer than 3" means $X \le 2$; "more than 3" means $X \ge 4$. Then add the probabilities of the values you listed, because different values of X cannot happen at the same time. The table shows B(5, 0.8), the number of 5 seeds that sprout, with the running total $P(X \le x)$ in the last column. For example, $P(X \ge 4) = 0.4096 + 0.32768 = 0.73728$.

B(5, 0.8): number of 5 seeds that sprout

x	P(X = x)	P(X ≤ x)
0	0.0003	0.0003
1	0.0064	0.0067
2	0.0512	0.0579
3	0.2048	0.2627
4	0.4096	0.6723
5	0.3277	1

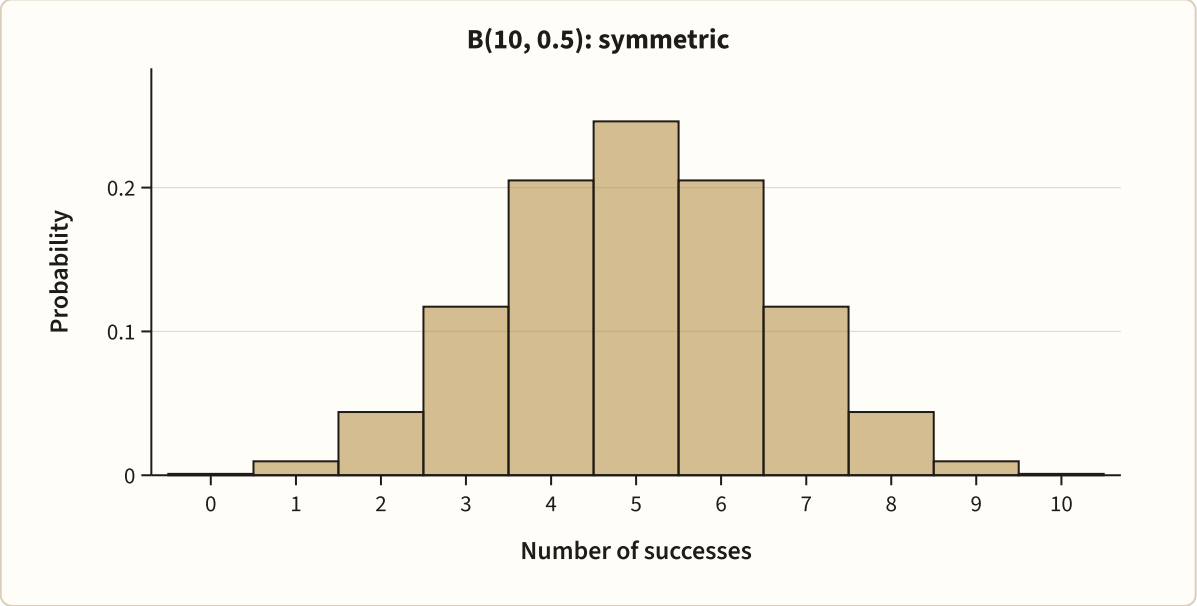
Often the quickest route is the **complement**. All the probabilities add to 1, so $P(X \ge k) = 1 - P(X \le k - 1)$. Choose whichever side has fewer terms. The probability that at least one of 5 seeds sprouts needs five terms the long way, but only one with the complement: $P(X \ge 1) = 1 - P(X = 0) = 1 - (0.2)^5 = 0.99968$. A graphing calculator gives cumulative probabilities $P(X \le k)$ with binomcdf(n, p, k).

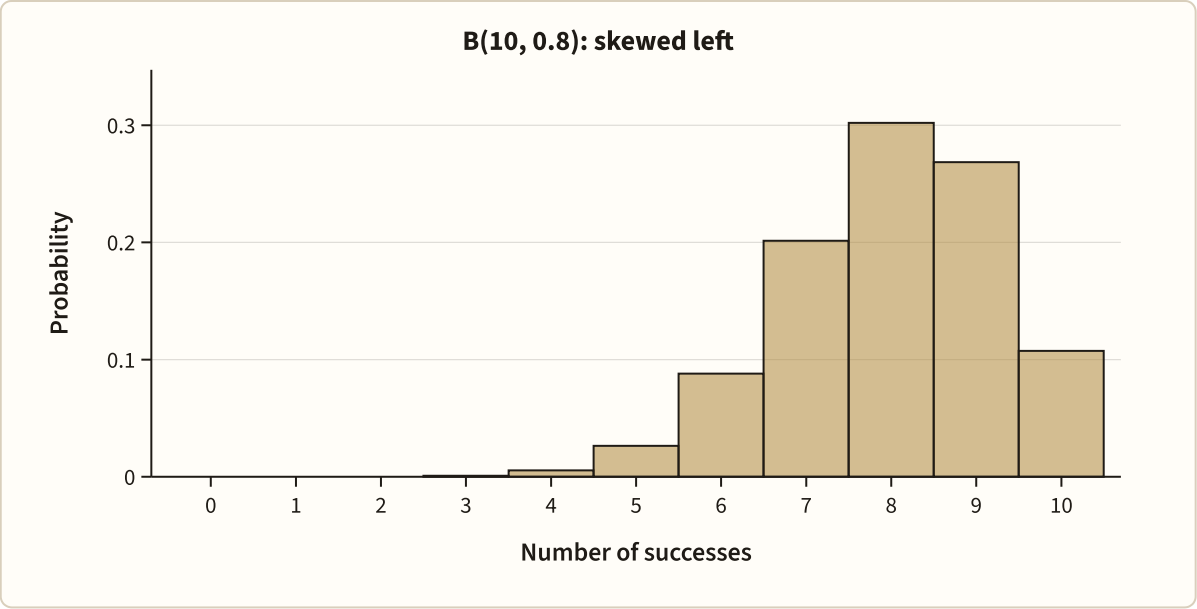
If you plant 50 seeds that each sprout 80% of the time, you expect about 40 to sprout: $n \times p = 50 \times 0.8$. Here is why $\mu = np$ always works. Think of X as a sum of n small random variables, one per trial, each equal to 1 for a success and 0 for a failure. One trial has mean $\mu = 1 \cdot p + 0 \cdot (1 - p) = p$. Means add (Lesson 7.3), so the sum of n trials has mean $p + p + \cdots + p = np$. The mean is a long-run average: over many plantings of 50 seeds, the average number that sprout will be close to 40, though any one planting may give 37 or 43.

The same idea gives the standard deviation. One 0-or-1 trial has variance $(1 - p)^2 p + (0 - p)^2(1 - p) = p(1 - p)[(1 - p) + p] = p(1 - p)$. The trials are independent, so their variances add (never the standard deviations): the variance of X is $np(1 - p)$ and $\sigma = \sqrt{np(1 - p)}$. Check it on B(5, 0.8) from the table:

$\sum x \cdot P(x) = 0.0064 + 0.1024 + 0.6144 + 1.6384 + 1.6384 = 4 = 5 \times 0.8$, and $\sigma = \sqrt{5 \times 0.8 \times 0.2} = \sqrt{0.8} \approx 0.894$. In sets of 5 seeds, the number that sprout is typically about 0.9 seed away from 4.

The shape of a binomial distribution depends on p. When p = 0.5, successes and failures are equally likely and the histogram is symmetric. When p is small, most of the probability piles up near 0 and a tail stretches right: skewed right. When p is large, the probability piles up near n and the tail stretches left: skewed left. The three histograms below all have n = 10, with p = 0.2, 0.5 and 0.8. For larger n the skewness fades, which is the key to Lesson 8.3.





Finally, a practical question. When we pick a sample, we usually pick without replacement, which strictly breaks independence. But if the population is large it hardly matters. Choosing tenors from a choir of 10 with 4 tenors, the chance changes from 4/10 = 0.4 to 3/9, about 0.33, after one tenor is picked. From a choral festival of 1,000 singers with 400 tenors, it changes only from 0.4 to 399/999, about 0.399. The **10% condition** makes this precise: if the sample is at most 10% of the population ($n \le 0.10N$), treat X as binomial.

KEY FORMULAS

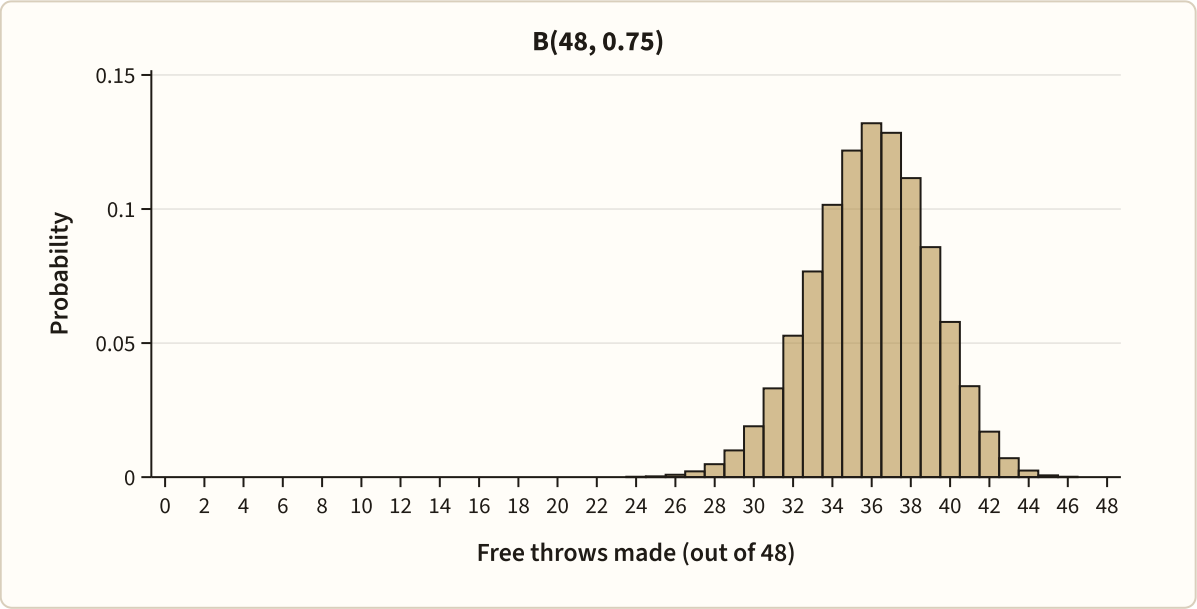
$$\mu = np \qquad \sigma = \sqrt{np(1-p)}$$

$$P(X \ge k) = 1 - P(X \le k-1) \qquad P(X \ge 1) = 1 - P(X = 0)$$

10% condition: when sampling without replacement, $n \le 0.10N$.

Practice habit. Translate the words into an inequality and list the values of X before adding: "at least 3 of 5" means X = 3, 4 or 5. Then ask whether the complement has fewer terms.

Example 1. Ruth makes 75% of her free throws. At practice she shoots 48 free throws, and we assume the shots are independent. Find the mean and standard deviation of X = the number she makes, and interpret both in context.



1. X counts successes in a fixed n = 48 independent shots with the same p = 0.75, so X is B(48, 0.75).
2. Mean: $\mu = np = 48 \times 0.75 = 36$.
3. Standard deviation: $\sigma = \sqrt{np(1-p)} = \sqrt{48 \times 0.75 \times 0.25} = \sqrt{9} = 3$.
4. Interpret the mean: over many practices of 48 shots, Ruth makes 36 free throws on average.
5. Interpret the standard deviation: in a typical practice, the number she makes is about 3 shots away from 36, so results from about 33 to 39 are ordinary.
6. Use it: a practice with only 28 makes would be $\dfrac{28-36}{3} \approx -2.67$ standard deviations from the mean, an unusually poor day (or a sign that something about her shooting has changed).

Answer: $\mu = 36$ free throws, $\sigma = 3$ free throws. On average she makes 36 of 48, and a typical practice is about 3 makes away from 36.

Example 2. A shipment of 2,000 apples contains 10% with a blemish. An inspector picks 8 apples at random, without replacement. Let X = the number of blemished apples. (a) Find the probability that at least one is blemished. (b) Find the probability that at most one is blemished. Round to 4 decimal places.

1. 10% condition: $n = 8 \le 0.10 \times 2000 = 200$, so we may treat X as B(8, 0.1).
2. (a) "At least one" is $X \ge 1$; its complement is $X = 0$, one term: $P(X = 0) = (0.9)^8 = 0.43046721$.
3. $P(X \ge 1) = 1 - 0.43046721 = 0.56953279 \approx 0.5695$.
4. (b) "At most one" is $X \le 1$, that is X = 0 or 1. $P(X = 1) = \binom{8}{1}(0.1)(0.9)^7 = 8 \times 0.1 \times 0.4782969 = 0.38263752$.
5. $P(X \le 1) = 0.43046721 + 0.38263752 = 0.81310473 \approx 0.8131$.
6. Conclusion: about 57% of random samples of 8 apples contain at least one blemished apple, and about 81% contain no more than one.

Answer: (a) $P(X \ge 1) = 1 - (0.9)^8 \approx 0.5695$ (b) $(0.9)^8 + 8(0.1)(0.9)^7 \approx 0.8131$

PRACTICE

1. EASY A seed company sells cucumber seeds that sprout 90% of the time, independently of one another. A gardener plants 100 seeds. Find the mean and standard deviation of the number that sprout.
2. EASY Samuel makes 30% of his three-point shots. The table gives the distribution of X = the number he makes out of 6 attempts. Find $P(X \le 1)$. Round to 4 decimal places.

B(6, 0.3): three-pointers made out of 6

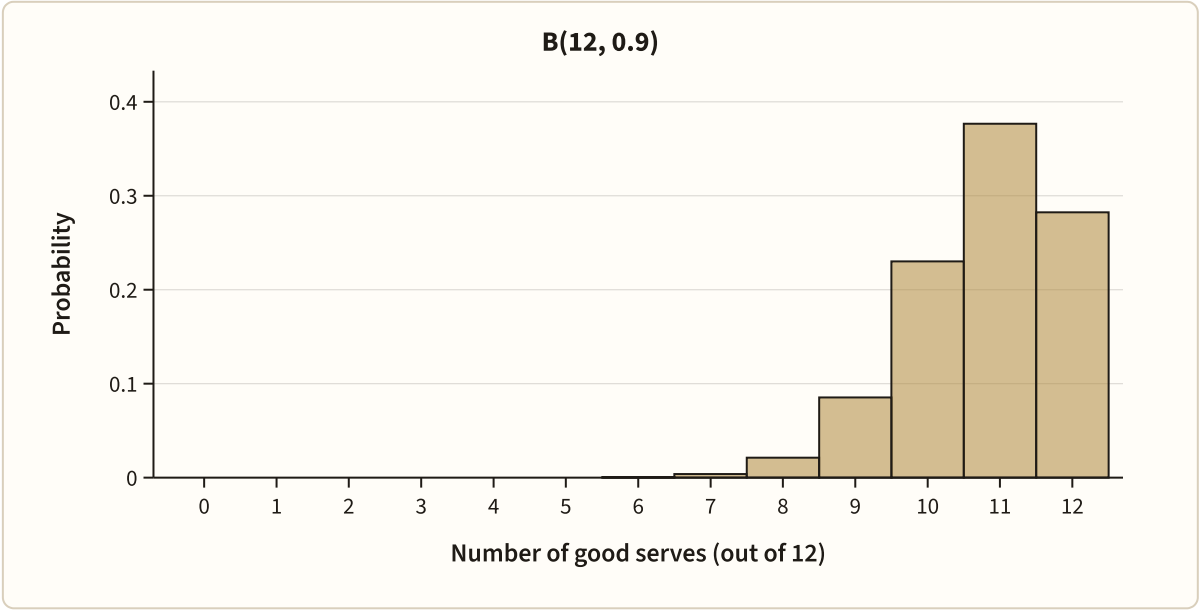
x	P(X = x)
0	0.1176
1	0.3025
2	0.3241
3	0.1852
4	0.0595
5	0.0102
6	0.0007

3. **EASY** Use the table for Samuel's 6 three-point attempts. Find the probability that he makes more than 4. Round to 4 decimal places.

B(6, 0.3): three-pointers made out of 6

x	P(X = x)
0	0.1176
1	0.3025
2	0.3241
3	0.1852
4	0.0595
5	0.0102
6	0.0007

4. **EASY** Grace's volleyball serves land in bounds 90% of the time. The probability histogram shows the number of good serves out of 12. Describe the shape of the distribution.



5. **MEDIUM** Henry plants 4 pepper seeds, each of which sprouts with probability 0.6, independently. Find the probability that at least one sprouts. Round to 4 decimal places.

6. **MEDIUM** Ruth makes 80% of her free throws, independently from shot to shot. She shoots 25 in a practice. (a) Find the mean of X = the number she makes. (b) Find the standard deviation. (c) Interpret the mean in context.

7. MEDIUM Use the table for Samuel's 6 three-point attempts. Find the probability that he makes fewer than 3. Round to 4 decimal places.

B(6, 0.3): three-pointers made out of 6

x	P(X = x)
0	0.1176
1	0.3025
2	0.3241
3	0.1852
4	0.0595
5	0.0102
6	0.0007

8. MEDIUM Decide whether the 10% condition is met. (a) A church choir has 40 members, 10 of them sopranos. The director picks 5 members at random to sing at a wedding; X = the number of sopranos. (b) A county has 2,000 4-H members, 25% of whom raise sheep. A newsletter editor picks 50 members at random; X = the number who raise sheep.

9. MEDIUM A gardener plants 50 seeds that each sprout with probability 0.8. Explain in your own words why the expected number of sprouts is 40, and what "expected" means here. Will exactly 40 sprout?

10. HARD For Samuel's 6 three-point attempts, a classmate writes $P(X \geq 2) = 1 - P(X \leq 2)$. (a) Explain the mistake. (b) Find the correct probability, rounded to 4 decimal places, using the table.

B(6, 0.3): three-pointers made out of 6

x	P(X = x)
0	0.1176
1	0.3025
2	0.3241
3	0.1852
4	0.0595
5	0.0102
6	0.0007

11. **HARD** A field-goal kicker makes 90% of his short kicks, independently. In a game he attempts 4 short kicks. (a) Find the probability that he makes at most 2. (b) Find the mean number made. (c) Find the standard deviation.

12. **HARD** A seed company claims that 90% of its carrot seeds sprout, independently of one another. A farmer plants 400 of the seeds and 342 sprout. (a) If the claim is true, find the mean and (b) the standard deviation of the number that sprout out of 400. (c) Is 342 an unusually low result if the claim is true? Explain.

13. **CHALLENGE** Grace hits the bull's-eye on 20% of her arrows, independently. What is the smallest number of arrows she must shoot so that the probability of at least one bull's-eye is at least 0.9?

14. **CHALLENGE** Henry makes 60% of his free throws and shoots 3. The distribution of X = the number made is $P(0) = 0.064$, $P(1) = 0.288$, $P(2) = 0.432$, $P(3) = 0.216$. (a) Compute the mean from $\mu = \sum x \cdot P(x)$ and check that it equals np. (b) Compute the standard deviation from $\sigma^2 = \sum (x - \mu)^2 P(x)$, round to 3 decimal places, and check it against $\sqrt{np(1-p)}$.

COMMON MISTAKES

- Using the wrong complement. The complement of $X \geq k$ is $X \leq k - 1$, not $X \leq k$. List the values on each side before you subtract.
- Mixing up "fewer than" and "at most". "Fewer than 3" leaves out 3 ($X \leq 2$); "at most 3" includes it ($X \leq 3$).
- Forgetting the square root: $np(1-p)$ is the variance; the standard deviation is $\sqrt{np(1-p)}$.
- Treating a sample from a small group as binomial. Check that the sample is at most 10% of the population first.

ANSWERS TO PRACTICE

1. $\mu = 100 \times 0.9 = 90$ seeds; $\sigma = \sqrt{100 \times 0.9 \times 0.1} = \sqrt{9} = 3$ seeds
2. $P(X \leq 1) = 0.117649 + 0.302526 = 0.420175 \approx 0.4202$
3. "More than 4" means X = 5 or 6: $P(X > 4) = 0.010206 + 0.000729 = 0.010935 \approx 0.0109$
4. Skewed left: most of the probability is piled near 11 and 12, with a tail stretching toward smaller values, as expected when p is large.
5. $P(X \geq 1) = 1 - P(X = 0) = 1 - (0.4)^4 = 1 - 0.0256 = 0.9744$
6. (a) $\mu = 25 \times 0.8 = 20$ (b) $\sigma = \sqrt{25 \times 0.8 \times 0.2} = \sqrt{4} = 2$ (c) Over many practices of 25 free throws, Ruth makes 20 on average.
7. "Fewer than 3" means X = 0, 1 or 2: $P(X < 3) = 0.117649 + 0.302526 + 0.324135 = 0.74431 \approx 0.7443$
8. (a) No: 10% of 40 is 4, and the sample of 5 is larger than that, so X should not be treated as binomial. (b) Yes: $n = 50 \leq 0.10 \times 2000 = 200$, so X is approximately B(50, 0.25).

9. Each seed contributes on average 0.8 of a sprout (it is 1 sprout 80% of the time and 0 otherwise), and 50 seeds contribute $np = 50 \times 0.8 = 40$ on average. "Expected" means the long-run average over many plantings of 50 seeds. Any one planting will vary; it may give 37 or 43, and exactly 40 is not guaranteed.
10. (a) The complement of "at least 2" is "at most 1" (X = 0 or 1), not "at most 2"; subtracting $P(X \le 2)$ wrongly removes X = 2, which belongs to the event. (b) $P(X \ge 2) = 1 - P(X \le 1) = 1 - 0.420175 = 0.579825 \approx 0.5798$
11. (a) $P(X \le 2) = 1 - P(X = 3) - P(X = 4) = 1 - 4(0.9)^3(0.1) - (0.9)^4 = 1 - 0.2916 - 0.6561 = 0.0523$ (b) $\mu = 4 \times 0.9 = 3.6$ kicks (c) $\sigma = \sqrt{4 \times 0.9 \times 0.1} = \sqrt{0.36} = 0.6$ kicks
12. (a) $\mu = 400 \times 0.9 = 360$ (b) $\sigma = \sqrt{400 \times 0.9 \times 0.1} = \sqrt{36} = 6$ (c) Yes. 342 is 18 seeds below 360, which is $\dfrac{342 - 360}{6} = -3$ standard deviations. A result that far below the mean would be very rare if the claim were true, so the farmer has reason to doubt the 90% claim (or to ask whether the seeds were stored badly).
13. $P(\text{at least one}) = 1 - (0.8)^n \ge 0.9$ means $(0.8)^n \le 0.1$. $(0.8)^{10} \approx 0.1074$ is too big and $(0.8)^{11} \approx 0.0859$ works, so she needs 11 arrows.
14. (a) $\mu = 0 + 0.288 + 0.864 + 0.648 = 1.8 = 3 \times 0.6$ (b) $\sigma^2 = (3.24)(0.064) + (0.64)(0.288) + (0.04)(0.432) + (1.44)(0.216) = 0.20736 + 0.18432 + 0.01728 + 0.31104 = 0.72$, so $\sigma = \sqrt{0.72} \approx 0.849$, the same as $\sqrt{3 \times 0.6 \times 0.4} = \sqrt{0.72}$.

LESSON 8.3 · 2 SESSIONS OF 45 MINUTES

The Normal Approximation to the Binomial

OBJECTIVES

- Check the conditions $np \geq 10$ and $n(1-p) \geq 10$ and explain why they are needed.
- Use a normal distribution with mean np and standard deviation $\sqrt{np(1-p)}$ to approximate binomial probabilities when n is large.
- Compare a normal approximation with the exact binomial probability and explain the small difference.
- Interpret approximate binomial probabilities in context.

VIDEO

WATCH

The Normal Approximation to the Binomial Distribution

jbstatistics · 14:09

When to watch: Before Example 1: approximating a binomial count with a normal curve.

https://www.youtube.com/watch?v=CCqWkJ_pqNU

BACKUP

Normal Approximation to Binomial Distribution (example question)

Eddie Woo · 13:13

When to watch: After Example 2: the same idea for a sample proportion, a preview of Lesson 9.2.

https://www.youtube.com/watch?v=rTPoXiQHI6M

> **Note:** The main video also uses the continuity correction, which this course treats as optional.

KEY TERMS

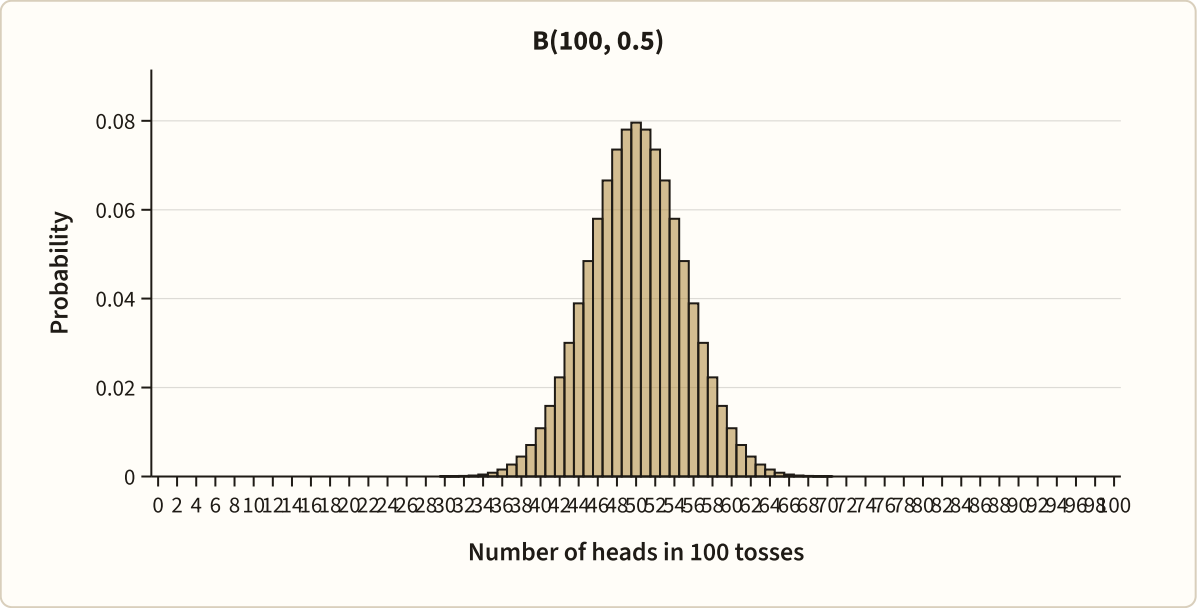
Normal approximation to the binomial	Using the normal distribution $N(np, \sqrt{np(1-p)})$ in place of the exact binomial distribution B(n, p) to find probabilities when n is large.
Large Counts condition	The expected numbers of successes and failures are both at least 10: $np \geq 10$ and $n(1-p) \geq 10$. When it holds, the binomial histogram is close enough to a bell shape.

Continuity correction

An optional refinement: since each bar of the binomial histogram covers half a unit on each side of a whole number, use 40.5 instead of 40 (for example) when finding $P(X \leq 40)$ with the normal curve.

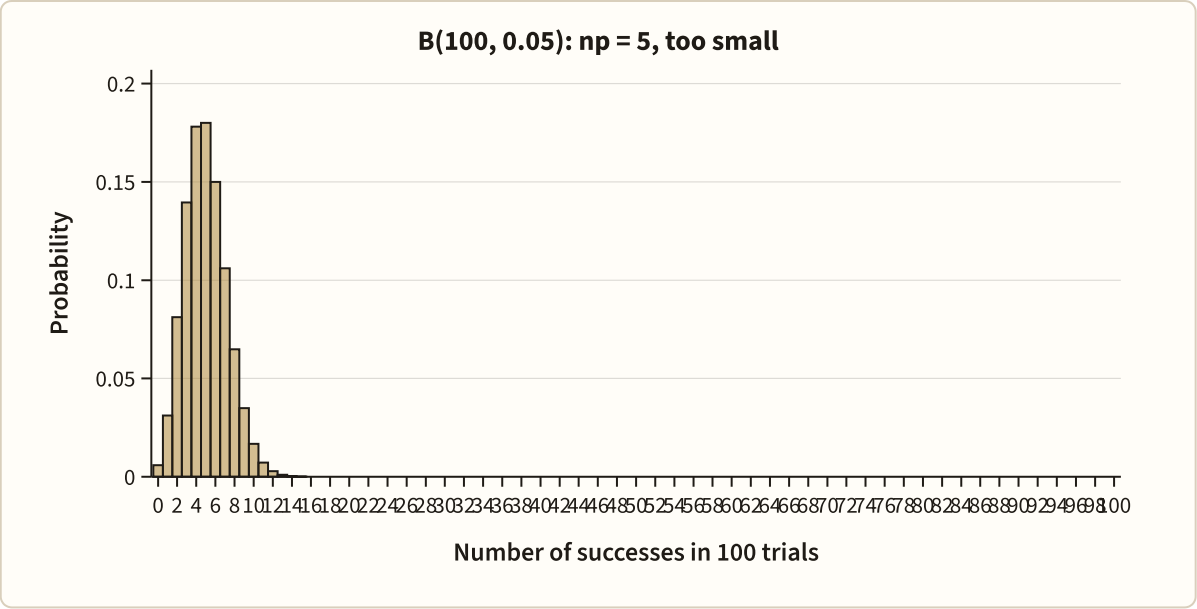
CORE IDEAS

Suppose a farmer plants 400 seeds that each sprout with probability 0.8 and wants $P(X \leq 308)$. The binomial formula works, but it would mean adding 309 terms. There is a better way, and it comes from looking at the shape of a binomial distribution when n is large. Here is B(100, 0.5), the number of heads in 100 tosses of a fair coin:

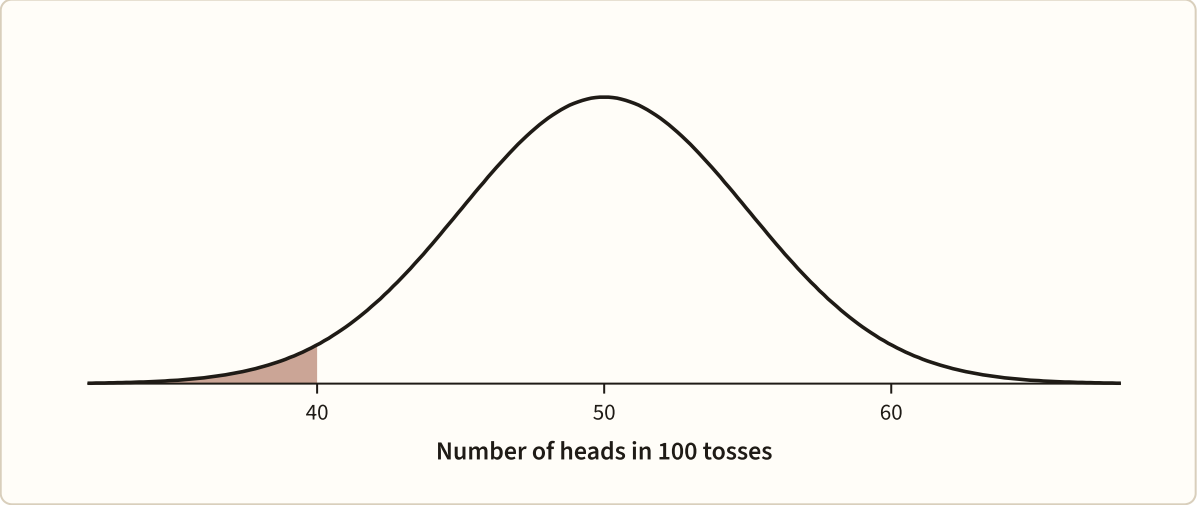


The histogram is a bell. That is no accident. A binomial count is a sum of many independent 0-or-1 trials, and when many small independent pieces are added, their ups and downs tend to balance out: totals near the mean are common and totals far from it are rare, in a symmetric pattern. So for large n, X is approximately normal with the mean and standard deviation from Lesson 8.2: $\mu = np$ and $\sigma = \sqrt{np(1-p)}$.

How large must n be? The answer depends on p. When np is small, the distribution is jammed against 0 and skewed right, like B(100, 0.05) below, where the expected count is only 5. A normal curve with mean 5 and standard deviation $\sqrt{100 \times 0.05 \times 0.95} \approx 2.2$ would stretch below 0, where a count cannot go. The same trouble happens near n when $n(1-p)$ is small. The rule of thumb is the **Large Counts condition**: $np \geq 10$ and $n(1-p) \geq 10$, so that both the expected successes and the expected failures leave room for a bell on each side.



The method uses the normal skills of Lesson 3.4. Check the conditions; find $\mu = np$ and $\sigma = \sqrt{np(1-p)}$; draw and shade the normal curve; standardize with $z = \dfrac{x - np}{\sqrt{np(1-p)}}$; find the area with the table or technology; and conclude in context. For 100 tosses of a fair coin, $np = n(1-p) = 50 \geq 10$, $\mu = 50$ and $\sigma = \sqrt{25} = 5$. The probability of 40 or fewer heads is about $P(Z \leq \dfrac{40 - 50}{5}) = P(Z \leq -2.00) = 0.0228$.



How good is it? The exact binomial answer, from binomcdf(100, 0.5, 40), is 0.0284. The approximation is close, but a little low, and here is why: in the histogram the bar for 40 covers 39.5 to 40.5, and the curve cut off at 40 misses the right half of that bar. Using 40.5 instead gives $z = \dfrac{40.5 - 50}{5} = -1.90$ and an area of 0.0287, very close to the exact value. This adjustment is the **continuity correction**; this course does not require it, and for large n the difference matters less and less.

The same picture explains a strange-sounding fact. A normal curve gives every single value an area of 0 (Lesson 3.4), so it cannot answer "exactly 50 heads" without the half-unit trick. Use the normal approximation for ranges of values, such as "at most 308" or "between 110 and 140". With technology

the exact binomial is easy, but the approximation still matters: with only np and $\sqrt{np(1-p)}$, you can judge in your head whether a count is ordinary or surprising.

KEY FORMULAS

Large Counts condition: $np \geq 10$ and $n(1-p) \geq 10$

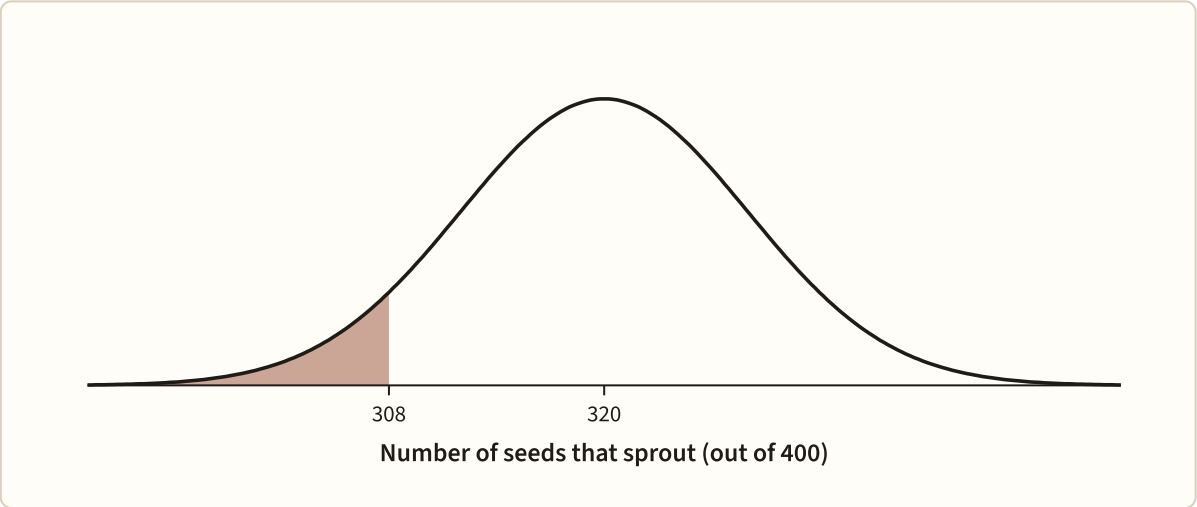
$$X \text{ is approximately } N\left(np, \sqrt{np(1-p)}\right)$$

$$z = \frac{x - np}{\sqrt{np(1-p)}}$$

From history. In 1733 Abraham de Moivre, a French-born mathematician living in London, was studying the number of heads in a large number of coin tosses. Adding the binomial terms by hand was hopeless, so he found a smooth curve that approximated them very closely. That curve was the normal curve, making its first appearance in mathematics as an approximation to the binomial.

Practice habit. Before you use the normal curve for a count, check $np \geq 10$ and $n(1-p) \geq 10$; then sketch the curve with $\mu = np$ marked and shade the region you want.

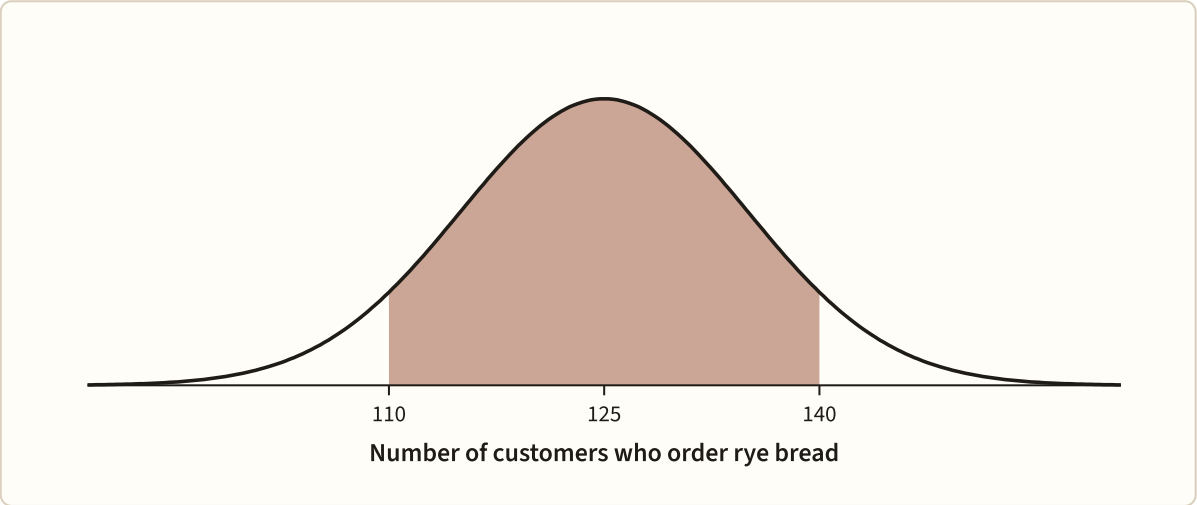
Example 1. A farmer plants 400 seeds of a corn variety in which each seed sprouts with probability 0.8, independently. Use the normal approximation to find the probability that at most 308 seeds sprout. Round to 4 decimal places.



1. X = the number that sprout is B(400, 0.8).
2. Check Large Counts: $np = 400 \times 0.8 = 320 \geq 10$ and $n(1-p) = 400 \times 0.2 = 80 \geq 10$.
3. Mean: $\mu = np = 320$ seeds.
4. Standard deviation: $\sigma = \sqrt{400 \times 0.8 \times 0.2} = \sqrt{64} = 8$ seeds.
5. Standardize: $z = \dfrac{308 - 320}{8} = -1.50$.
6. Area to the left of -1.50: $P(Z \leq -1.50) = 0.0668$.
7. Conclusion: if the 80% sprouting rate is right, only about 7% of plantings of 400 seeds would give 308 or fewer sprouts.

Answer: $z = \dfrac{308 - 320}{8} = -1.50$; $P(X \leq 308) \approx 0.0668$

Example 2. On a busy Saturday the Miller family bakery expects 625 customers. Each customer, independently, buys a loaf of rye bread with probability 0.2. (a) Use the normal approximation to find the probability that between 110 and 140 customers buy rye bread. Round to 4 decimal places. (b) How many rye loaves should the bakery bake so that the chance of running out is only about 5%?



1. X = the number of rye buyers is B(625, 0.2). Check: $np = 125 \geq 10$ and $n(1-p) = 500 \geq 10$.
2. $\mu = 625 \times 0.2 = 125$ and $\sigma = \sqrt{625 \times 0.2 \times 0.8} = \sqrt{100} = 10$.
3. (a) Standardize both ends: $z = \dfrac{110 - 125}{10} = -1.50$ and $z = \dfrac{140 - 125}{10} = 1.50$.
4. Area between: $P(-1.50 \leq Z \leq 1.50) = 0.9332 - 0.0668 = 0.8664$.
5. (b) Running out means demand above the number baked, L. We want an area of 0.05 above L, so 0.95 below: $z = 1.645$.
6. $L = \mu + z\sigma = 125 + 1.645 \times 10 = 141.45$.
7. Round UP to 142 loaves: 141 would leave slightly more than a 5% chance of running out.
8. Conclusion: on about 87% of such Saturdays, between 110 and 140 customers buy rye; baking 142 loaves means running out on only about 1 Saturday in 20.

Answer: (a) $P(110 \leq X \leq 140) \approx P(-1.50 \leq Z \leq 1.50) = 0.8664$ (b) $L = 125 + 1.645(10) = 141.45$, so bake 142 loaves

PRACTICE

1. EASY On a farm, 10% of the harvested potatoes are too small to sell. A worker picks 50 potatoes at random. Is the Large Counts condition met, so that the number of small potatoes can be approximated by a normal distribution?

2. **EASY** In a large maple grove, 60% of the trees are sugar maples. A forester picks 150 trees at random. (a) Is the Large Counts condition met? (b) Find the mean and (c) the standard deviation of the number of sugar maples.

3. **EASY** Henry tosses a fair coin 100 times. Use the normal approximation to find the probability of 60 or more heads. Round to 4 decimal places.

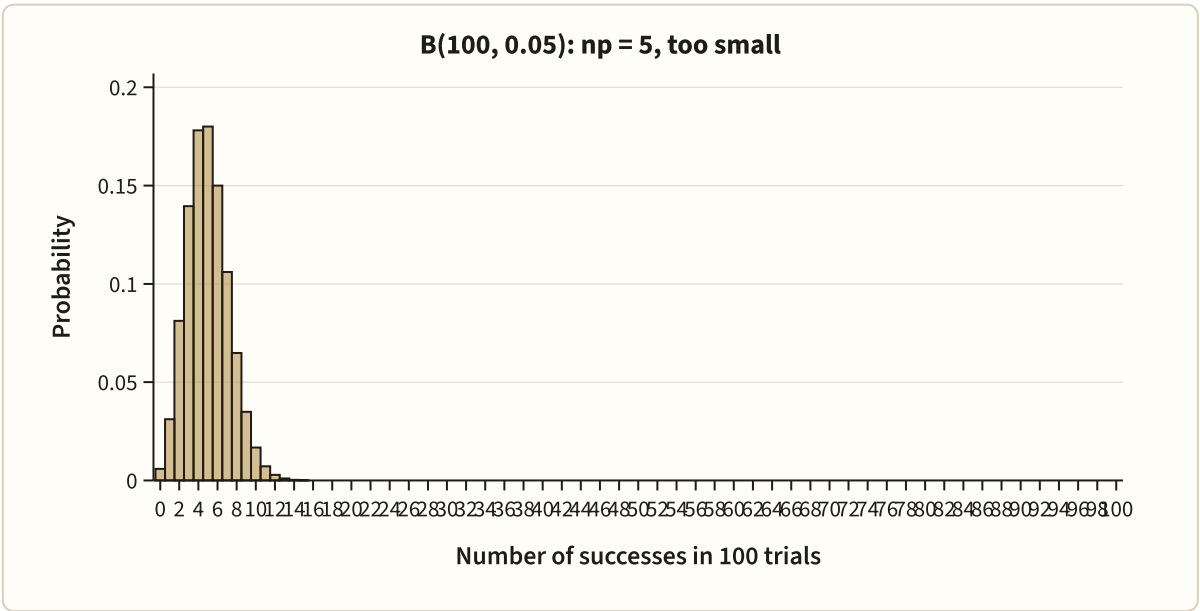
4. **EASY** Henry found that the probability of 60 or more heads in 100 tosses of a fair coin is about 0.0228. Interpret this probability in context.

5. **MEDIUM** In a large hatchery, 90% of fertile eggs hatch, independently of one another. Use the normal approximation to find the probability that at most 801 of a batch of 900 eggs hatch. Round to 4 decimal places.

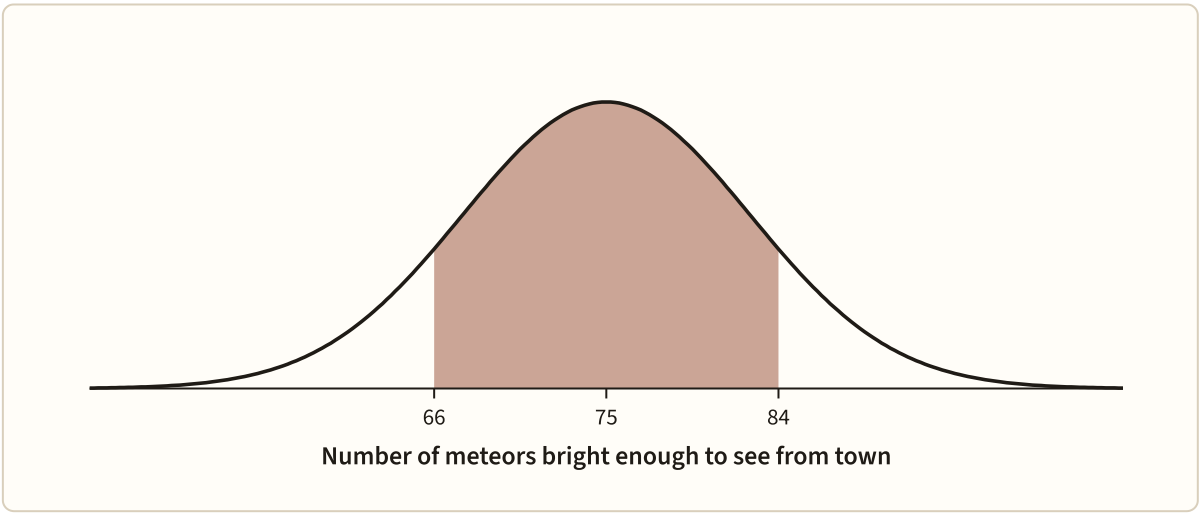
6. **MEDIUM** A machine makes bolts, and 2% of them are defective, independently from bolt to bolt. Use the normal approximation to find the probability that a box of 2,500 bolts contains between 43 and 57 defective bolts. Round to 4 decimal places.

7. **MEDIUM** Ruth sinks 60% of her putts from 8 feet, independently from putt to putt. In a season she attempts 600 such putts. Use the normal approximation to find the probability that she sinks at least 390. Round to 4 decimal places.

8. **MEDIUM** The probability histogram shows B(100, 0.05). Explain why the normal approximation works well for B(100, 0.5) but not for B(100, 0.05), even though n = 100 in both.



9. HARD During a meteor shower, about 25% of the meteors seen from a dark hilltop are bright enough to be seen from town, independently from meteor to meteor. An astronomy club counts 300 meteors from the hilltop. Use the normal approximation to find the probability that between 66 and 84 of them would be bright enough to see from town. Round to 4 decimal places.



10. HARD A seed catalog claims that 80% of its tomato seeds sprout. Grace plants 625 of them and only 470 sprout. (a) If the claim is true, use the normal approximation to find the probability that 470 or fewer would sprout. Round to 4 decimal places. (b) What does this suggest about the claim?

11. HARD A pumpkin patch expects 600 visitors on a fall Saturday, and each visitor, independently, buys a pumpkin with probability 0.4. How many pumpkins should the patch have on hand so that the probability of running out is only about 10%? Use the normal approximation.

12. CHALLENGE At a sawmill, 4% of the boards have a knot too large for furniture grade. What is the smallest number of boards n for which the Large Counts condition holds for the number of boards with large knots?

13. CHALLENGE Henry tosses a fair coin 100 times. The exact binomial probability of exactly 50 heads is 0.0796. (a) What does a normal curve with $\mu = 50$ and $\sigma = 5$ give for $P(X = 50)$ if you use the single value 50? (b) Using the continuity correction, approximate $P(X = 50)$ by the area under the normal curve between 49.5 and 50.5. Round to 4 decimal places.

COMMON MISTAKES

- Skipping the Large Counts check. If np or n(1 − p) is less than 10, the binomial is too skewed for a normal curve; use the exact binomial instead.
- Using $np(1 - p)$ as the standard deviation. That is the variance; take the square root before you standardize.
- Using the normal curve for a single value such as "exactly 50". The curve gives 0; use the binomial formula (or the continuity correction) for one value.
- Rounding a planning number the wrong way. When you need enough loaves or pumpkins, round UP, or the chance of running out will be larger than you wanted.

ANSWERS TO PRACTICE

1. No: $np = 50 \times 0.1 = 5$, which is less than 10. The distribution is too skewed right for a normal curve.
2. (a) Yes: $np = 90 \geq 10$ and $n(1-p) = 60 \geq 10$. (b) $\mu = 150 \times 0.6 = 90$ (c) $\sigma = \sqrt{150 \times 0.6 \times 0.4} = \sqrt{36} = 6$
3. $\mu = 50$, $\sigma = 5$; $z = \dfrac{60 - 50}{5} = 2.00$; $P(Z \geq 2.00) = 1 - 0.9772 = 0.0228$
4. If Henry tossed a fair coin 100 times over and over, he would get 60 or more heads in only about 2 out of every 100 sets of tosses (about 2.3%). Getting that many heads is possible but rare.
5. Check: $np = 810$, $n(1-p) = 90$. $\mu = 810$, $\sigma = \sqrt{900 \times 0.9 \times 0.1} = \sqrt{81} = 9$; $z = \dfrac{801 - 810}{9} = -1.00$; $P(Z \leq -1.00) = 0.1587$
6. Check: $np = 50$, $n(1-p) = 2450$. $\mu = 50$, $\sigma = \sqrt{2500 \times 0.02 \times 0.98} = \sqrt{49} = 7$; $z = \pm 1.00$; area $= 0.8413 - 0.1587$, which is 0.6827 by technology (table: 0.6826)
7. $\mu = 600 \times 0.6 = 360$, $\sigma = \sqrt{600 \times 0.6 \times 0.4} = \sqrt{144} = 12$; $z = \dfrac{390 - 360}{12} = 2.50$; $P(Z \geq 2.50) = 1 - 0.9938 = 0.0062$
8. For B(100, 0.05) the expected count is only $np = 5$, less than 10. The distribution is piled up against 0 and skewed right, and a normal curve centered at 5 with standard deviation about 2.2 would put some area below 0, where a count is impossible. For B(100, 0.5), $np = n(1-p) = 50$, so the histogram has room on both sides and is bell-shaped.
9. Check: $np = 75$, $n(1-p) = 225$. $\mu = 75$, $\sigma = \sqrt{300 \times 0.25 \times 0.75} = \sqrt{56.25} = 7.5$; $z = \dfrac{66 - 75}{7.5} = -1.20$ and $z = \dfrac{84 - 75}{7.5} = 1.20$; area $= 0.8849 - 0.1151$, which is 0.7699 by technology (table: 0.7698)
10. (a) $\mu = 625 \times 0.8 = 500$, $\sigma = \sqrt{625 \times 0.8 \times 0.2} = \sqrt{100} = 10$; $z = \dfrac{470 - 500}{10} = -3.00$; $P(Z \leq -3.00) = 0.0013$ (b) If 80% really sprouted, a result as low as 470 would happen in fewer than 2 plantings out of 1,000. That is so unlikely that Grace has good reason to doubt the 80% claim for her seeds.
11. $\mu = 240$, $\sigma = \sqrt{600 \times 0.4 \times 0.6} = \sqrt{144} = 12$. For an area of 0.90 below, $z = 1.28$ (technology 1.2816). $\mu + z\sigma = 240 + 1.2816 \times 12 \approx 255.4$; round up to 256 pumpkins.
12. Need $n(0.04) \geq 10$, so $n \geq \dfrac{10}{0.04} = 250$; then $n(1-p) = 240 \geq 10$ too. The smallest is n = 250 boards.
13. (a) 0: a single value has no area under a continuous curve. (b) $z = \dfrac{49.5 - 50}{5} = -0.10$ and $z = \dfrac{50.5 - 50}{5} = 0.10$; area $= 0.5398 - 0.4602$, which is 0.0797 by technology (table: 0.0796), very close to the exact 0.0796.

LESSON 8.4 · 2 SESSIONS OF 45 MINUTES

Geometric Distributions

OBJECTIVES

- Recognize a geometric setting and tell it apart from a binomial setting.
- Compute $P(X = k)$, $P(X > k)$ and $P(X \leq k)$ for a geometric random variable and explain why each formula works.
- Compute and interpret the mean $\mu = 1/p$ of a geometric random variable.
- Explain why every geometric distribution is skewed right.

VIDEO

WATCH

An Introduction to the Geometric Distribution

jbstatistics · 10:47

When to watch: Before Example 1: the geometric setting and formula.

https://www.youtube.com/watch?v=zq9Oz82iHf0

BACKUP

Pokémon and Geometric Distributions - Numberphile

Numberphile · 23:21

When to watch: Enrichment: a geometric distribution worked out with a game of catching Pokémon.

https://www.youtube.com/watch?v=K79aOe-F0Mk

KEY TERMS

Geometric setting	Independent trials, each a success or failure with the same probability p of success, repeated until the first success occurs. The number of trials is not fixed in advance.
Geometric random variable	X = the number of the trial on which the first success occurs. Its possible values are 1, 2, 3, ..., with no upper limit.
Mean of a geometric random variable	$\mu = 1/p$, the average number of trials needed to get the first success, over many repetitions.

CORE IDEAS

Sometimes the question is not "how many successes in n trials?" but "how long until the first success?" In a board game you must roll a 6 to start; Grace shoots arrows until she hits the bull's-eye; a fisherman casts until a fish bites. The trials are still binary and independent with the same probability p, but now we do not fix the number of trials; we keep going until the first success. The random variable X = the trial on which the first success occurs is a **geometric random variable**. The table compares the two settings.

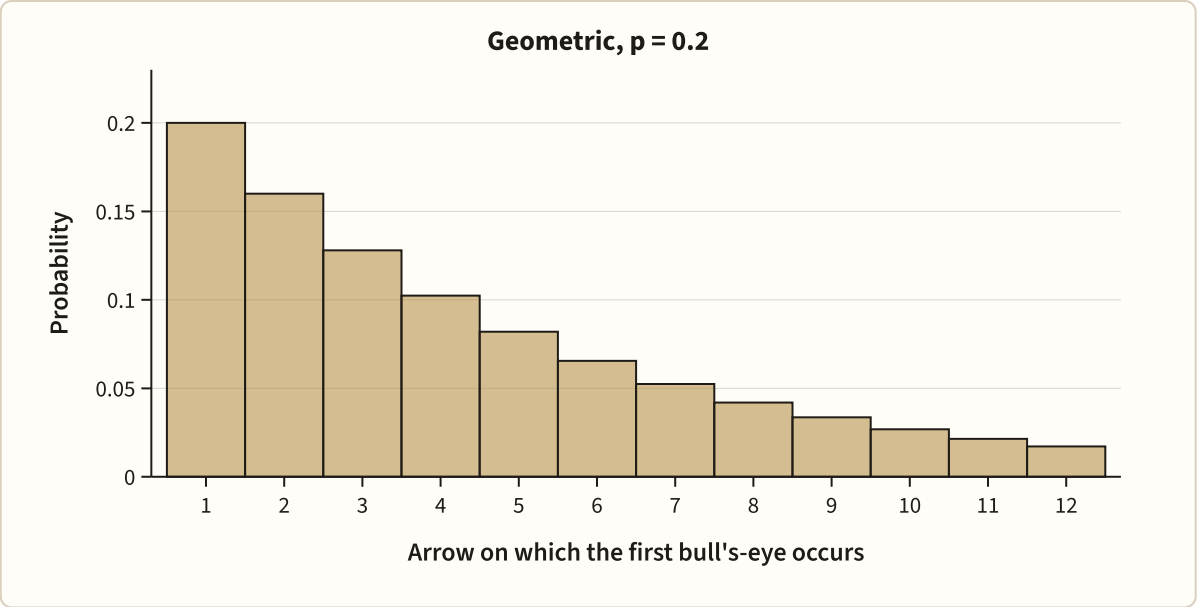
Binomial or geometric?

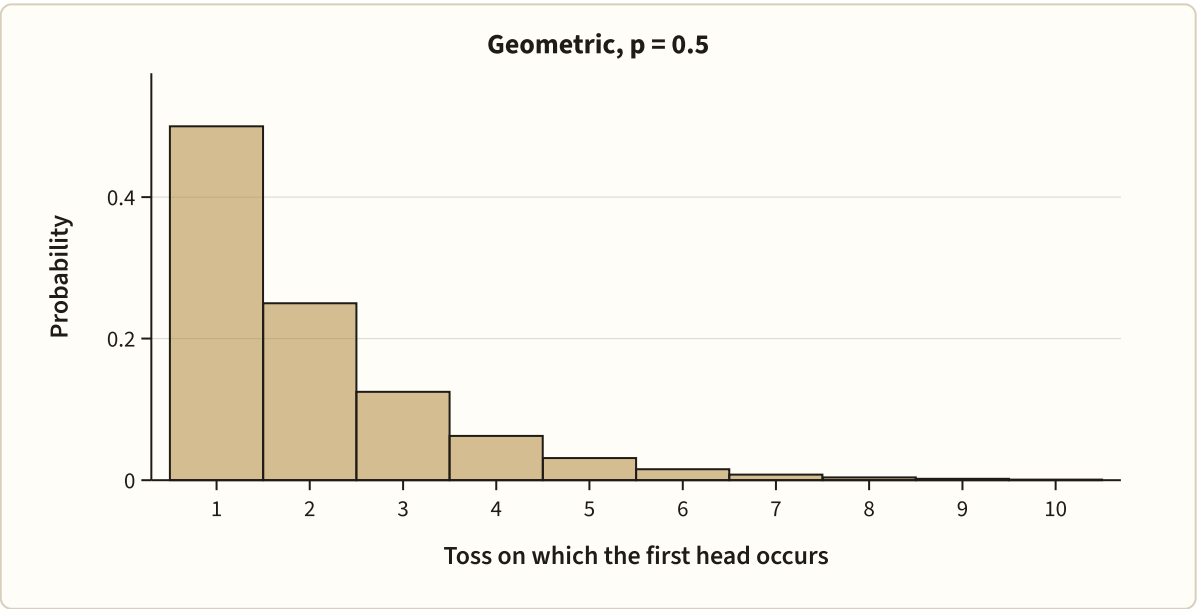
	Binomial	Geometric
Number of trials	Fixed in advance (n)	Not fixed: continue until the first success
X counts	Successes in n trials	The trial on which the first success occurs
Possible values	0, 1, 2, ..., n	1, 2, 3, ... (no upper limit)
P(X = k)	C(n, k) p^k (1 - p)^(n - k)	(1 - p)^(k - 1) p
Mean	np	1/p

The formula is simpler than the binomial formula. The event X = k means the first k − 1 trials are failures and trial k is a success. There is only one such arrangement (F, F, ..., F, S), so there is no $\binom{n}{k}$. Multiplying because the trials are independent gives $P(X = k) = (1 - p)^{k-1}p$. For Grace, with p = 0.2: $P(X = 1) = 0.2$, $P(X = 2) = 0.8 \times 0.2 = 0.16$, $P(X = 3) = 0.8^2 \times 0.2 = 0.128$, and so on.

Questions such as "more than 5 arrows" would seem to need infinitely many terms: X = 6, 7, 8, ... But there is a shortcut with a simple reason. Needing more than k trials means exactly that the first k trials were all failures, so $P(X > k) = (1 - p)^k$. Its complement is getting the first success within k trials: $P(X \leq k) = 1 - (1 - p)^k$. For Grace, $P(X > 5) = 0.8^5 = 0.32768$: about one time in three she still has not hit the bull's-eye after 5 arrows.

Every geometric distribution has the same shape. Each bar is $(1 - p)$ times the bar before it, so the bars always shrink as k grows: the tallest bar is at X = 1 and a long tail stretches to the right. Every geometric distribution is skewed right, whatever p is. It may surprise you that even for a rare success the single most likely value is 1. That is because getting the first success on trial 1 needs only one thing to go right, while any later value also needs every earlier trial to fail.





The mean is $\mu = \frac{1}{p}$, and the reason is easy to see. In the long run a fraction p of all trials are successes, so there is about one success for every $\frac{1}{p}$ trials, and the average wait from one success to the next is $\frac{1}{p}$ trials. A fair die gives a 6 with p = 1/6, so the average wait is 6 rolls; Grace waits 5 arrows on average. Because of the long right tail, the mean is larger than the most likely value, just as the mean is pulled toward the tail in any skewed-right distribution.

Finally, the trials have no memory (recall Lesson 6.1). If Grace has missed 3 arrows in a row, the next arrow still hits with probability 0.2; she is not "due". Starting from any point, the wait for the next success has the same geometric distribution as the wait from the beginning.

KEY FORMULAS

$$P(X = k) = (1-p)^{k-1}p, \quad k = 1, 2, 3, \ldots$$

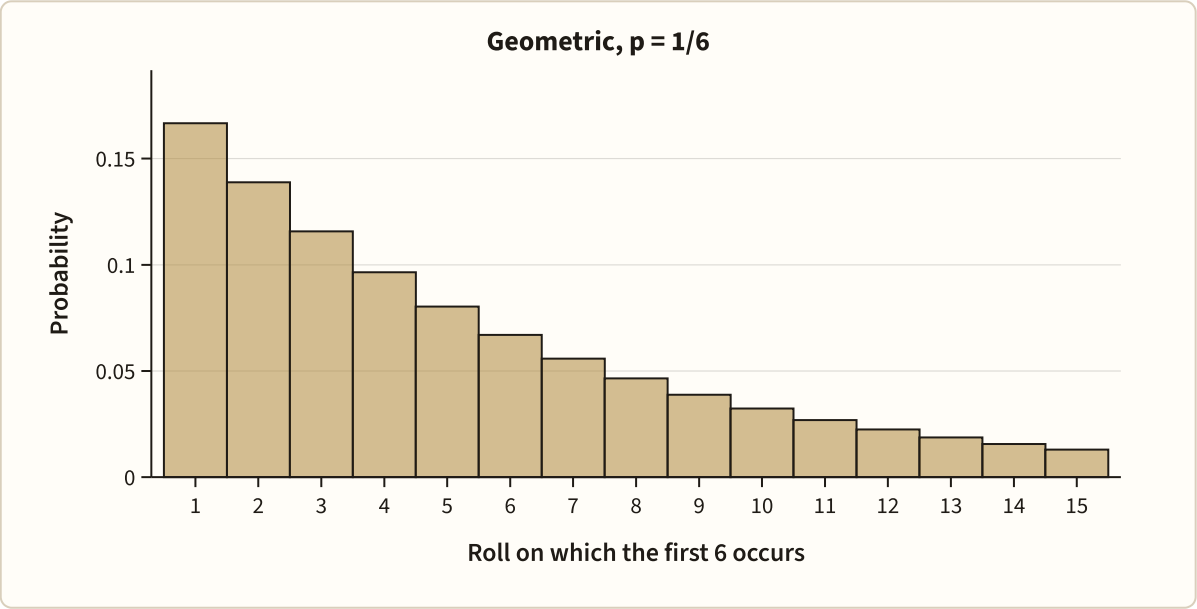
$$P(X > k) = (1-p)^k \qquad P(X \le k) = 1 - (1-p)^k$$

$$\mu = \frac{1}{p}$$

Practice habit. Ask first: is the number of trials fixed in advance (binomial, count the successes) or am I waiting for the first success (geometric, count the trials)?

WORKED EXAMPLES

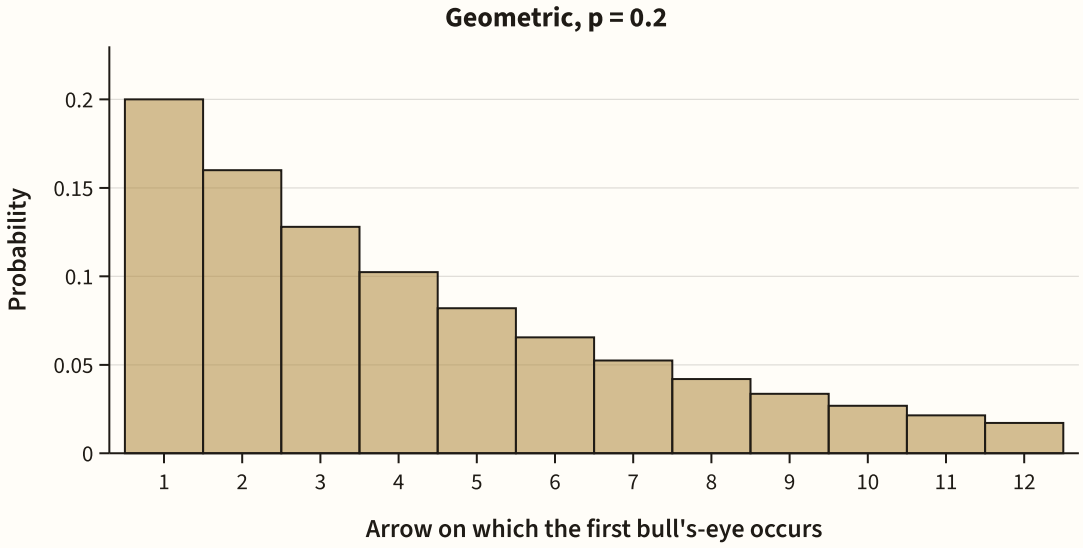
Example 1. In a board game, Samuel must roll a 6 on a fair die before he can move his first piece. Let X = the roll on which he gets his first 6. (a) Find the probability that his first 6 comes on the third roll. Round to 4 decimal places. (b) Find the mean of X and interpret it.



1. Check: each roll is a 6 or not (binary), rolls are independent, p = 1/6 every time, and he rolls until the first 6. X is geometric with p = 1/6.
2. (a) First 6 on roll 3 means not-6, not-6, 6: $P(X = 3) = \left(\frac{5}{6}\right)^{2}\left(\frac{1}{6}\right)$.
3. $= \frac{25}{36} \times \frac{1}{6} = \frac{25}{216} \approx 0.1157$.
4. (b) $\mu = \frac{1}{p} = \frac{1}{1/6} = 6$ rolls.
5. Conclusion: about 12% of the time Samuel's first 6 comes exactly on his third roll, and over many games he needs 6 rolls on average to start.

Answer: (a) $\left(\frac{5}{6}\right)^{2}\left(\frac{1}{6}\right) = \frac{25}{216} \approx 0.1157$ (b) $\mu = 6$ rolls: on average Samuel needs 6 rolls to get his first 6.

Example 2. Grace hits the bull's-eye on 20% of her arrows, independently. Let X = the arrow on which she first hits the bull's-eye. (a) Find $P(X = 4)$. (b) Find the probability that she needs more than 5 arrows. Round to 4 decimal places. (c) Find the mean of X.



1. X is geometric with p = 0.2: binary, independent, same p, and she shoots until the first bull's-eye.
2. (a) Three misses, then a hit: $P(X = 4) = (0.8)^3(0.2) = 0.512 \times 0.2 = 0.1024$.
3. (b) Needing more than 5 arrows means the first 5 all miss: $P(X > 5) = (0.8)^5 = 0.32768 \approx 0.3277$.
4. (c) $\mu = \dfrac{1}{0.2} = 5$ arrows.
5. Conclusion: about 10% of the time her first bull's-eye comes on exactly the 4th arrow; about a third of the time she has still not hit it after 5 arrows; and on average she needs 5 arrows. The histogram shows the long right tail that makes waits of 8 or 10 arrows possible.

Answer: (a) $(0.8)^3(0.2) = 0.1024$ (b) $(0.8)^5 \approx 0.3277$ (c) $\mu = 5$ arrows

PRACTICE

1. EASY Say whether X is binomial or geometric, and why. (a) Ruth shoots 10 free throws; X = the number she makes. (b) Ruth shoots free throws until she makes one; X = the number of shots she takes.
2. EASY Henry tosses a fair coin until he gets heads. Find the probability that the first head comes on the third toss.
3. EASY In a mixed flock, 25% of the eggs are brown. Grace takes eggs one at a time at random from a large basket until she gets a brown one. On average, how many eggs will she take?

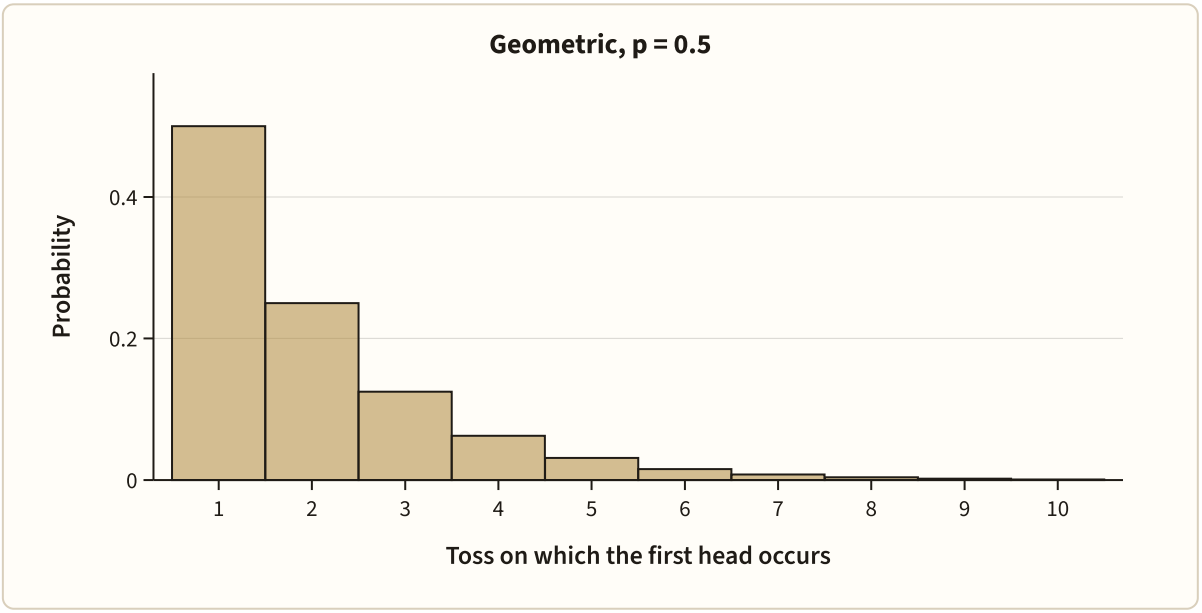
4. EASY Ruth makes 80% of her free throws, independently from shot to shot. She shoots until she makes one. Find the probability that her first make is on her second shot.

5. MEDIUM At a large cattle sale, 30% of the cattle are Herefords. A buyer looks at cattle one at a time, chosen at random, until she finds a Hereford. Find the probability that the first Hereford is the 4th animal she looks at. Round to 4 decimal places.

6. MEDIUM On a night of a meteor shower, Samuel has a 0.1 chance of seeing a fireball (a very bright meteor) in any one hour of watching, independently from hour to hour. (a) Find the probability that he has to watch more than 5 hours to see his first fireball. Round to 4 decimal places. (b) Find the mean number of hours until his first fireball.

7. MEDIUM Samuel rolls a fair die until he gets a 6. Find the probability that he gets his first 6 within his first 3 rolls. Round to 4 decimal places.

8. MEDIUM The probability histogram shows the toss on which the first head occurs when a fair coin is tossed. Explain why the tallest bar is at X = 1 and why the distribution is skewed right.



9. MEDIUM Henry catches a fish on 30% of his casts, independently. Explain why the probability that he needs more than 4 casts to catch his first fish is $(0.7)^4$, without adding up an endless list of terms. Then compute it.

10. HARD A church volunteer phones families to sign up for the fall potluck. Each call is answered with probability 0.4, independently. (a) Find the probability that she gets her first answer within her first 3 calls. (b) Find the mean number of calls until her first answer.

11. HARD Over many fishing trips, Ruth needs on average 8 casts to catch her first fish. Her casts are independent with the same probability p of a catch. (a) Find p. (b) Find the probability that she needs more than 2 casts. Round to 4 decimal places.

12. HARD A batter gets a hit in 25% of her at-bats, independently. What is the smallest number k of at-bats such that the probability of getting her first hit within the first k at-bats is at least 0.95?

13. **CHALLENGE** Grace hits the bull's-eye on 20% of her arrows. She has missed with her first 3 arrows. Find the probability that her first bull's-eye comes on the 5th arrow, given that the first 3 missed. Compare your answer with $P(X = 2)$ and explain.

COMMON MISTAKES

- Using the wrong exponent. The first success on trial k needs k − 1 failures, so the formula is $(1-p)^{k-1}p$, not $(1-p)^k p$.
- Including $\binom{n}{k}$. There is only one arrangement (all failures, then the success), so a geometric probability has no binomial coefficient.
- Mixing up "more than k" and "at least k". $P(X > k) = (1-p)^k$, but $P(X \geq k) = (1-p)^{k-1}$, because $X \geq k$ only needs the first k − 1 trials to fail.
- Thinking a success is "due" after many failures, or that the mean is the most likely value. The trials have no memory, and the most likely value is always 1.

ANSWERS TO PRACTICE

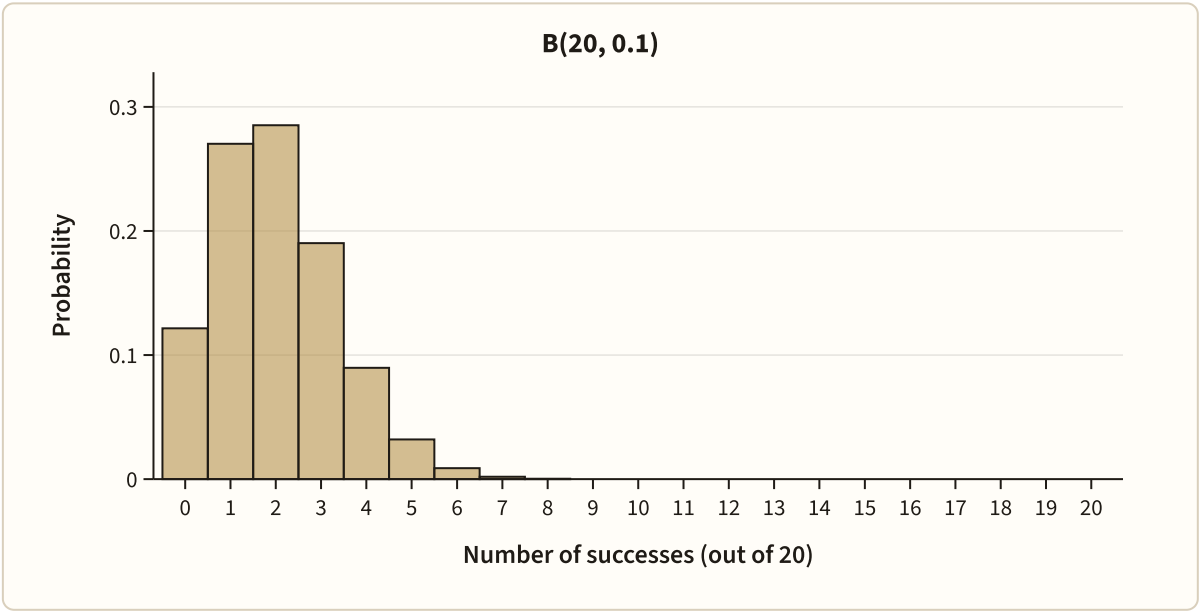
1. (a) Binomial: the number of trials is fixed (n = 10) and X counts successes. (b) Geometric: the number of trials is not fixed, and X is the trial on which the first success occurs.
2. $P(X = 3) = (0.5)^2(0.5) = 0.125$
3. $\mu = \dfrac{1}{0.25} = 4$ eggs
4. Miss, then make: $P(X = 2) = (0.2)(0.8) = 0.16$
5. $P(X = 4) = (0.7)^3(0.3) = 0.343 \times 0.3 = 0.1029$
6. (a) $P(X > 5) = (0.9)^5 = 0.59049 \approx 0.5905$ (b) $\mu = \dfrac{1}{0.1} = 10$ hours
7. $P(X \leq 3) = 1 - \left(\dfrac{5}{6}\right)^3 = 1 - \dfrac{125}{216} = \dfrac{91}{216} \approx 0.4213$
8. Each bar is $(1-p) = 0.5$ times the bar before it, so the bars keep shrinking: $P(X = 1) = 0.5$, $P(X = 2) = 0.25$, $P(X = 3) = 0.125$, and so on. Getting heads on the first toss needs only one thing to go right; a later first head also needs every earlier toss to be tails. The shrinking bars make a long tail to the right.
9. Needing more than 4 casts happens exactly when the first 4 casts all fail. Each fails with probability 0.7, and the casts are independent, so the probability is $(0.7)^4 = 0.2401$. What happens after the 4th cast does not matter.
10. (a) $P(X \leq 3) = 1 - (0.6)^3 = 1 - 0.216 = 0.784$ (b) $\mu = \dfrac{1}{0.4} = 2.5$ calls
11. (a) $\mu = \dfrac{1}{p} = 8$, so $p = \dfrac{1}{8} = 0.125$ (b) $P(X > 2) = (0.875)^2 = 0.765625 \approx 0.7656$
12. Need $P(X \leq k) = 1 - (0.75)^k \geq 0.95$, that is $(0.75)^k \leq 0.05$. $(0.75)^{10} \approx 0.0563$ is too big and $(0.75)^{11} \approx 0.0422$ works, so k = 11.

13. $P(X = 5 \mid X > 3) = \dfrac{(0.8)^4(0.2)}{(0.8)^3} = 0.8 \times 0.2 = 0.16$. This equals $P(X = 2) = 0.16$: after 3 misses, the 4th arrow must miss and the 5th hit, exactly like starting over and waiting 2 arrows. The arrows have no memory; she is not "due" for a hit.

Unit 8 test

You have 60 minutes. You may use a scientific or graphing calculator and the normal table at the back of the course. Show your work: write the distribution you are using (with n and p), the formula, the substitution and the result, and answer in context when asked. Round probabilities to 4 decimal places unless a problem says otherwise. The test is worth 100 points.

1. **6 PTS** In a county with 400 beehives, 20% of the hives have mites. An inspector checks 15 hives chosen at random. Let X = the number of hives with mites. (a) May X be treated as binomial? Check the 10% condition. (b) Give n. (c) Give p.
2. **5 PTS** A quarterback throws passes in practice until one is caught. Let X = the number of passes he throws. Is X a binomial random variable? Explain.
3. **6 PTS** A sow gives birth to a litter of 6 piglets. Each piglet is equally likely to be male or female, independently. Find the probability that exactly 2 piglets are male.
4. **7 PTS** A softball pitcher strikes out 40% of the batters she faces, independently. Find the probability that she strikes out exactly 2 of the next 5 batters.
5. **7 PTS** A gardener plants 225 tulip bulbs, each of which blooms with probability 0.8, independently. (a) Find the mean number of bulbs that bloom. (b) Find the standard deviation. (c) Interpret the mean in context.
6. **5 PTS** The probability histogram shows B(20, 0.1). Describe its shape.



7. **8 PTS** In a mixed packet of sunflower seeds, 30% grow into red sunflowers. Grace plants 4 seeds; the table gives the distribution of X = the number of red sunflowers. (a) Find the probability of at least 2 red sunflowers. (b) Use the complement to find the probability of at most 3.

B(4, 0.3): red sunflowers out of 4 seeds

x	P(X = x)
0	0.2401
1	0.4116
2	0.2646
3	0.0756
4	0.0081

8. **7 PTS** Each tomato plant in Henry's garden has a 0.1 chance of being attacked by hornworms this summer, independently. He has 6 plants. Find the probability that at least one plant is attacked.

9. **7 PTS** At a sawmill, 25% of the boards are grade A. An inspector checks boards one at a time, in random order, until she finds a grade A board. (a) Find the probability that the first grade A board is the third one she checks. (b) Find the mean number of boards she checks.

10. **7 PTS** At a church fall festival, a child lands a ring on a bottle with probability 0.3 on each toss, independently, and keeps tossing until a ring lands. (a) Find the probability that it takes more than 3 tosses. (b) Find the mean number of tosses, rounded to 2 decimal places.

11. **8 PTS** In a large commercial flock, 10% of the eggs are jumbo size. A packer selects 1,600 eggs at random. (a) Find the mean and (b) the standard deviation of the number of jumbo eggs. (c) Check the Large Counts condition and use the normal approximation to find the probability of at least 175 jumbo eggs.

12. **8 PTS** On a large sheep ranch, 400 lambs are born this spring. Each lamb is equally likely to be a ram or a ewe lamb, independently. Use the normal approximation to find the probability that between 190 and 215 of the lambs are rams.

13. **8 PTS** A rancher's cows each have a heifer (female) calf with probability 0.5, independently. Explain why the probability of exactly 2 heifers among 3 calves is $\binom{3}{2}(0.5)^2(0.5)^1$. Say what each of the three factors means, and why the factors are multiplied.

14. **6 PTS** Only 5% of the apples from an orchard are too large for the standard box. A packer picks 60 apples at random and wants to use a normal curve to find probabilities for the number of oversized apples. (a) Is the Large Counts condition met? (b) Explain what would go wrong with the normal approximation.

15. 5 PTS Grace has rare heirloom bean seeds, each of which sprouts with probability 0.6, independently. What is the smallest number of seeds she must plant so that the probability that at least one sprouts is at least 0.99?

UNIT 8 TEST: ANSWER KEY

1. (a) Yes: each hive has mites or not, the sample of 15 is at most 10% of 400 (that is, 40), so the checks are nearly independent with the same p. (b) n = 15 (c) p = 0.2
2. No. The number of trials is not fixed in advance; X counts the trials until the first success, so it is a geometric random variable.
3. $\binom{6}{2}(0.5)^2(0.5)^4 = 15 \times \dfrac{1}{64} = 0.234375 \approx 0.2344$
4. $\binom{5}{2}(0.4)^2(0.6)^3 = 10 \times 0.16 \times 0.216 = 0.3456$
5. (a) $\mu = 225 \times 0.8 = 180$ bulbs (b) $\sigma = \sqrt{225 \times 0.8 \times 0.2} = \sqrt{36} = 6$ bulbs (c) If the gardener planted 225 such bulbs year after year, the average number that bloom would be about 180.
6. Skewed right: most of the probability is at 0, 1, 2 and 3, with a long tail toward larger counts, as expected when p is small.
7. (a) $P(X \geq 2) = 0.2646 + 0.0756 + 0.0081 = 0.3483$ (b) $P(X \leq 3) = 1 - P(X = 4) = 1 - 0.0081 = 0.9919$
8. $P(X \geq 1) = 1 - P(X = 0) = 1 - (0.9)^6 = 1 - 0.531441 = 0.468559 \approx 0.4686$
9. (a) $P(X = 3) = (0.75)^2(0.25) = 0.5625 \times 0.25 = 0.140625 \approx 0.1406$ (b) $\mu = \dfrac{1}{0.25} = 4$ boards
10. (a) $P(X > 3) = (0.7)^3 = 0.343$ (b) $\mu = \dfrac{1}{0.3} \approx 3.33$ tosses
11. (a) $\mu = 1600 \times 0.1 = 160$ (b) $\sigma = \sqrt{1600 \times 0.1 \times 0.9} = \sqrt{144} = 12$ (c) $np = 160 \geq 10$ and $n(1-p) = 1440 \geq 10$. $z = \dfrac{175 - 160}{12} = 1.25$; $P(Z \geq 1.25) = 1 - 0.8944 = 0.1056$
12. $np = n(1-p) = 200 \geq 10$. $\mu = 200$, $\sigma = \sqrt{400 \times 0.5 \times 0.5} = \sqrt{100} = 10$. $z = \dfrac{190 - 200}{10} = -1.00$ and $z = \dfrac{215 - 200}{10} = 1.50$; area $= 0.9332 - 0.1587 = 0.7745$
13. $(0.5)^2$ is the probability that two particular calves are heifers and $(0.5)^1$ that the third is a bull calf; these are multiplied because the births are independent. That product is the probability of one particular order, such as heifer, heifer, bull. $\binom{3}{2} = 3$ counts the orders in which 2 of the 3 calves can be heifers (HHB, HBH, BHH); these orders cannot happen together, so their equal probabilities add, which is the same as multiplying by 3. The result is $P(X = 2) = 3 \times 0.125 = 0.375$.
14. (a) No: $np = 60 \times 0.05 = 3$, less than 10. (b) The binomial distribution is piled up near 0 and strongly skewed right. A normal curve with mean 3 and standard deviation $\sqrt{60 \times 0.05 \times 0.95} \approx 1.7$ is symmetric and would put area below 0, where a count is impossible, so its probabilities would be poor. Use the exact binomial instead.
15. Need $P(X \geq 1) = 1 - (0.4)^n \geq 0.99$, that is $(0.4)^n \leq 0.01$. $(0.4)^5 = 0.01024$ is too big and $(0.4)^6 = 0.004096$ works, so she must plant 6 seeds.

UNIT 9

Sampling Distributions

Every sample tells a slightly different story. If you weigh 25 eggs from a flock today and another 25 tomorrow, the two sample means will not agree, even though the flock has not changed. This unit studies that sampling variability directly. A statistic such as $\hat{p}$ or $\bar{x}$ is a random variable: it has a distribution, called its sampling distribution, with a center, a spread and a shape. You will build sampling distributions by listing every sample from tiny populations and by simulation, and then find their center, spread and shape with formulas.

The unit pulls together much of the course so far. The binomial distribution of Unit 8 gives the sampling distribution of a sample proportion; the rule that variances add (Unit 7) explains why the spread of $\bar{x}$ is $\frac{\sigma}{\sqrt{n}}$; and the normal calculations of Unit 3 let us find probabilities. The central limit theorem then shows why the normal curve appears for sample means from almost any population.

This unit is the bridge to inference. In Units 10 through 13 you will turn the reasoning around: from one sample you will estimate an unknown parameter with a confidence interval or test a claim about it. Those methods make sense only because we know how statistics vary from sample to sample, which is exactly what this unit teaches.

PREREQUISITES

- Find normal probabilities and percentiles with z-scores, the table and technology (Lessons 3.4 and 3.5).
- Use the rules for means and variances of combined random variables, including "variances add" (Lesson 7.3).
- Use the binomial mean np, standard deviation $\sqrt{np(1-p)}$ and the Large Counts idea (Lessons 8.2 and 8.3).
- Count samples with $\binom{n}{k}$ (Lesson 6.6).
- Explain random sampling and bias (Lessons 5.1 and 5.2).

MASTERY CHECKLIST

☐ I can tell a parameter from a statistic and use the symbols p, μ, σ, $\hat{p}$, $\bar{x}$ and s correctly.

☐ I can build a sampling distribution by listing all samples from a small population and read one made by simulation.

☐ I can decide whether a statistic is unbiased and explain bias and variability with the target picture.

☐ I can explain why larger samples give less variability and why population size matters little when it is at least 10 times the sample.

☐ I can find the mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$ of $\hat{p}$ and check the 10% and Large Counts conditions.

☐ I can explain why four times the sample size cuts the standard deviation of a statistic in half.

☐ I can find the mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ of $\bar{x}$ and explain why averages vary less than single values.

☐ I can state the central limit theorem and decide when $\bar{x}$ is approximately normal.

☐ I can compute and interpret normal probabilities for $\hat{p}$, for $\bar{x}$ and for totals, in context.

☐ I can tell whether a question is about one individual or about a sample statistic, and use the right standard deviation.

LESSON 9.1 · 2 SESSIONS OF 45 MINUTES

Parameters, Statistics and Sampling Variability

OBJECTIVES

- Distinguish a parameter from a statistic and use the symbols p, μ, σ and $\hat{p}$, $\bar{x}$, s correctly.
- Build the sampling distribution of a statistic by listing all samples from a tiny population, and read one made by simulation.
- Decide whether a statistic is an unbiased estimator by comparing the center of its sampling distribution with the parameter.
- Explain why larger samples give statistics with less variability, and why the size of the population matters little when it is at least 10 times the sample.

VIDEO

WATCH

Sampling Distributions: Introduction to the Concept

jbstatistics · 7:51

When to watch: Before Example 1: parameters, statistics and sampling variability.

https://www.youtube.com/watch?v=Zbw-YvELsaM

BACKUP

Introduction to Sampling & Populations (1 of 4: Graphing the sample means)

Eddie Woo · 4:11

When to watch: After Example 2: graphing the means of many samples.

https://www.youtube.com/watch?v=70JM-26dJfQ

KEY TERMS

Parameter	A number that describes a whole population, such as the population proportion p, mean μ or standard deviation σ. It is fixed, but usually unknown.
Statistic	A number computed from a sample, such as $\hat{p}$, $\bar{x}$ or s. We know it once we have the sample, and we use it to estimate a parameter.
Sampling variability	The fact that different random samples from the same population give different values of a statistic.
Sampling distribution	The distribution of the values a statistic takes in all possible samples of the same size from the same population.

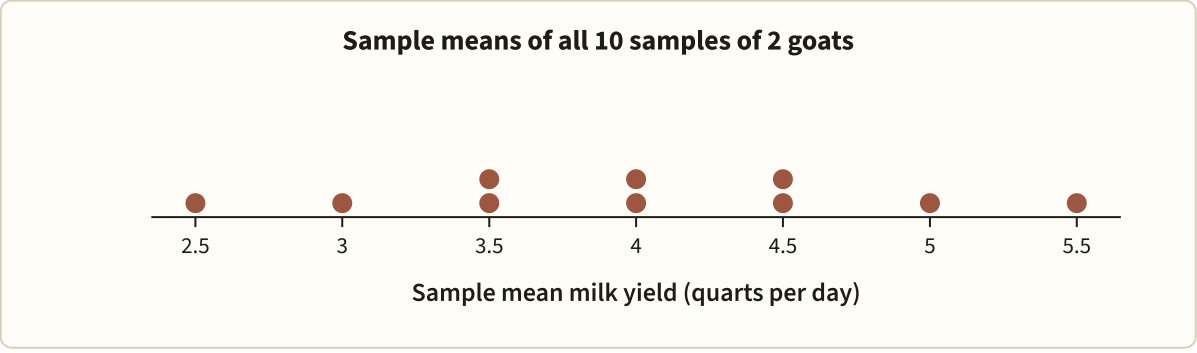
Unbiased estimator	A statistic whose sampling distribution has its mean (center) exactly equal to the parameter it estimates. It does not systematically overestimate or underestimate.
Variability of a statistic	The spread of its sampling distribution: how far the statistic typically lands from sample to sample. Larger random samples give smaller variability.

CORE IDEAS

Back in Unit 1 you met two words that now become the center of the course. A **parameter** is a number that describes a whole population; a **statistic** is a number computed from a sample. The usual parameters are the population proportion p, the population mean μ and the population standard deviation σ. The matching statistics are the sample proportion $\hat{p}$, the sample mean $\bar{x}$ and the sample standard deviation s. A parameter is a fixed number, but we almost never know it, because measuring a whole population is too costly or impossible. A statistic is known as soon as we have our sample, so we use it to estimate the parameter. A memory aid: **p**arameter goes with **p**opulation, **s**tatistic goes with **s**ample.

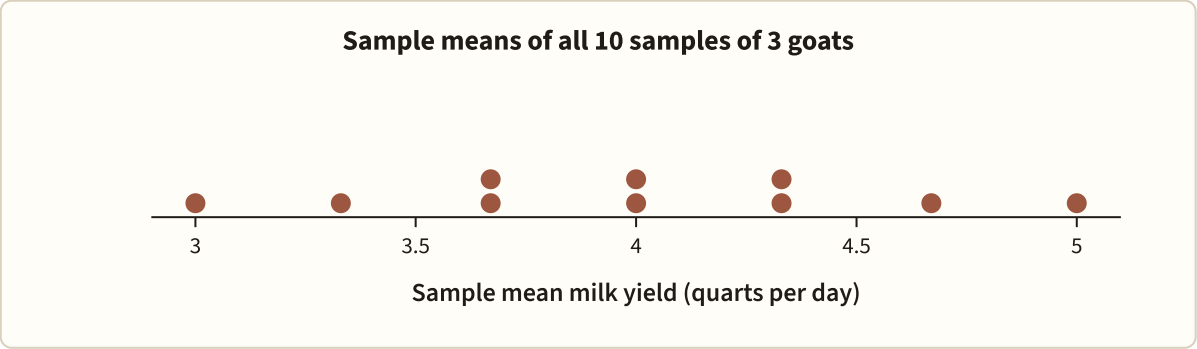
Take two random samples of 50 apples from the same orchard and you will almost certainly get two different values of $\hat{p}$, the proportion with blemishes. This is **sampling variability**. It is not a mistake by anyone; it is simply what chance does. The question statistics asks is: how much do statistics vary, and so how far from the parameter might our one estimate be? To answer it we study the **sampling distribution**: the distribution of the values a statistic takes in all possible samples of the same size from the same population.

Here is a population small enough to list every sample. The Miller family has 5 goats that give 2, 3, 4, 5 and 6 quarts of milk a day, so $\mu = \dfrac{2+3+4+5+6}{5} = 4$ quarts. There are $\binom{5}{2} = 10$ possible samples of 2 goats. Their sample means are 2.5, 3, 3.5, 4, 3.5, 4, 4.5, 4.5, 5 and 5.5 quarts. The dotplot shows this sampling distribution of $\bar{x}$ for samples of size 2.

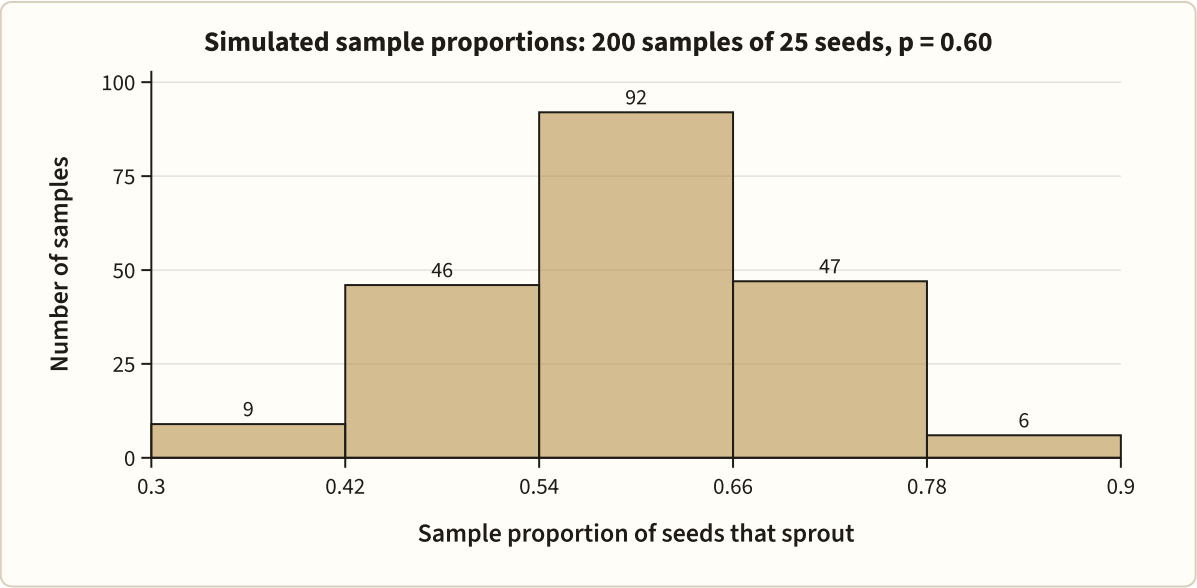


Notice two things. First, the 10 sample means add to 40, so their mean is $40 \div 10 = 4$, exactly μ. The sampling distribution is centered on the parameter, so $\bar{x}$ is an **unbiased estimator** of μ: it does not lean high or low. Second, the means spread from 2.5 to 5.5, so one unlucky sample could miss μ by 1.5 quarts. Now take samples of 3 goats instead (again 10 samples). Their means run only from 3 to 5.

Why? In a bigger sample a high value is more likely to be balanced by a low one, so the average lands closer to the middle. **Larger random samples give less variability.**



Real populations are far too big to list every sample, so we **simulate**. Suppose 60% of the seeds in a large seed lot sprout. A computer drew 200 random samples of 25 seeds from such a lot and recorded $\hat{p}$ for each. The histogram shows the result: it is centered near 0.60, roughly symmetric, and about 92% of the samples (185 of 200) gave $\hat{p}$ between 0.42 and 0.78. A single sample of 25 seeds could easily miss p by 0.10 or more. In the next three lessons we find the center, spread and shape of such sampling distributions without any simulation at all.



Picture the parameter as the bull's-eye of a target and each sample's statistic as one arrow. **Bias** means the arrows land consistently off to one side: the method aims at the wrong spot. High **variability** means the arrows are scattered widely. A good method has both low bias and low variability: arrows tightly grouped around the bull's-eye. Random sampling takes care of bias (with a fair method, $\hat{p}$ and $\bar{x}$ are unbiased; dividing by $n-1$ also makes the sample variance s^2 an unbiased estimator of σ^2, which is the deeper reason for $n-1$ in Unit 2). A larger sample takes care of variability. But a larger sample never fixes bias: a huge convenience sample still aims at the wrong spot.

One fact surprises almost everyone: the variability of a statistic depends on the size of the **sample**, not the size of the population, as long as the population is at least 10 times as large as the sample. A cook

tastes one spoonful from a well-stirred pot, and that spoonful tells as much about a large pot as a small one. So a random sample of 1,000 farms tells about as much about a large state as about a small one.

KEY FORMULAS

$$\text{parameters: } p,\ \mu,\ \sigma \qquad \text{statistics: } \hat{p},\ \bar{x},\ s$$

$$\text{unbiased estimator: mean of the sampling distribution} = \text{parameter}$$

$$10\%\text{ condition: } n \le 0.10N \quad (N = \text{population size})$$

Practice habit. Before you compute anything, ask: "Does this number describe the whole population or just the sample?" Parameters go with Populations, Statistics go with Samples.

WORKED EXAMPLES

Example 1. A county extension agent wants to learn about the 2,400 lambs born in the county this spring. The agent weighs a random sample of 36 of these lambs at 60 days old and finds a mean weight of 48.2 pounds; 9 of the 36 lambs are twins. (a) Identify the population, the two parameters of interest and the two statistics, with their symbols. (b) Find the value of the sample proportion of twins.

1. The population is the whole group we want to learn about: all 2,400 lambs born in the county this spring.
2. The parameters describe all 2,400 lambs: μ = the mean 60-day weight of all of them, and p = the proportion of all of them that are twins. Both are unknown.
3. The statistics come from the 36 lambs in the sample: $\bar{x} = 48.2$ pounds and $\hat{p}$.
4. $\hat{p} = \dfrac{\text{twins in sample}}{n} = \dfrac{9}{36} = 0.25$.
5. In context: we use $\bar{x} = 48.2$ pounds to estimate the unknown mean weight μ of all 2,400 lambs, and $\hat{p} = 0.25$ to estimate the unknown proportion p of twins. Another sample of 36 would give somewhat different statistics.

Answer: Population: the 2,400 lambs. Parameters: μ (mean weight of all 2,400) and p (proportion of all 2,400 that are twins). Statistics: $\bar{x} = 48.2$ pounds and $\hat{p} = \dfrac{9}{36} = 0.25$.

Example 2. A small orchard has just 4 young apple trees, A, B, C and D, which gave 4, 6, 8 and 10 bushels of apples. Treat these 4 trees as the population. (a) List all samples of 2 trees, and find the sample mean and the sample range (largest minus smallest) for each. (b) Find the mean of the 6 sample means and the mean of the 6 sample ranges (round to two decimal places). (c) Which statistic is an unbiased estimator of its parameter?

Sample	Bushels	Sample mean	Sample range
A, B	4, 6	5	2
A, C	4, 8	6	4
A, D	4, 10	7	6
B, C	6, 8	7	2
B, D	6, 10	8	4
C, D	8, 10	9	2

1. Population parameters: $\mu = \dfrac{4+6+8+10}{4} = \dfrac{28}{4} = 7$ bushels; population range $= 10 - 4 = 6$ bushels.
2. There are $\binom{4}{2} = 6$ samples of 2 trees. The table lists each with its mean and range.
3. Mean of the sample means: $\dfrac{5+6+7+7+8+9}{6} = \dfrac{42}{6} = 7$ bushels, exactly μ.
4. Mean of the sample ranges: $\dfrac{2+4+6+2+4+2}{6} = \dfrac{20}{6} \approx 3.33$ bushels, far below the population range of 6.
5. The sample range can never be larger than the population range, and it is usually smaller, so it systematically underestimates: it is biased.
6. Conclusion: the sample mean is an unbiased estimator of the mean yield of the 4 trees; the sample range is a biased estimator of the range, because on average it falls short by about 2.67 bushels.

Answer: Mean of the sample means $= 7 = \mu$; mean of the sample ranges ≈ 3.33 bushels, less than the population range of 6. The sample mean is unbiased; the sample range is biased (too small on average).

PRACTICE

1. EASY A beekeeper weighs the honey from every one of the 80 hives in the apiary and finds a mean of 52 pounds per hive. Is 52 pounds a parameter or a statistic? Give its symbol and explain.

2. **EASY** A sawmill inspector chooses 20 boards at random from the 3,000 boards cut that day. The 20 boards have a mean length of 96.1 inches, and 3 of them are warped. (a) Find the sample proportion $\hat{p}$ of warped boards. (b) Is 96.1 inches a parameter or a statistic? Give its symbol.

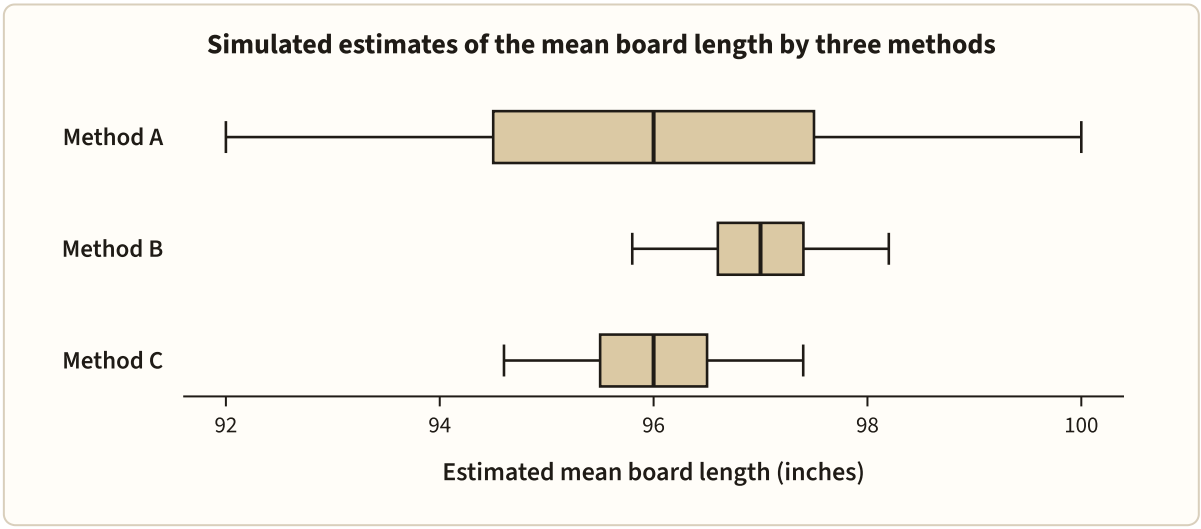
3. **EASY** Grace has 5 hens and wants to study every possible sample of hens. (a) How many different samples of 2 hens are there? (b) How many different samples of 3 hens?

4. **EASY** Samuel and Ruth each take a random sample of 30 families from the same parish and ask how many miles each family drives to church. Samuel gets $\bar{x} = 6.4$ miles and Ruth gets $\bar{x} = 5.7$ miles. Did one of them make a mistake? Explain.

5. **MEDIUM** One morning the 4 rain gauges on the Hale farm read 0.2, 0.4, 0.6 and 1.0 inch. Treat these 4 readings as the population. (a) Find μ. (b) List the sample means of all 6 samples of 2 gauges and find the mean of these 6 sample means. (c) How many of the 6 sample means are within 0.10 inch of μ?

6. **MEDIUM** A farmer has 5 calves: 2 heifers and 3 bulls, so the proportion of heifers is $p = 0.4$. Consider all 10 samples of 2 calves, and let $\hat{p}$ be the proportion of heifers in the sample. (a) What fraction of the 10 samples give $\hat{p} = 0.5$? (b) Find the mean of the 10 values of $\hat{p}$. (c) Is $\hat{p}$ an unbiased estimator of p here?

7. **MEDIUM** The true mean length of the boards a sawmill cut last month is $\mu = 96$ inches. Three methods of choosing a sample of boards were each simulated many times; the boxplots show the estimates of the mean length each method gave. Which methods appear unbiased, which has the least variability, and which method is best? Explain using the target picture.



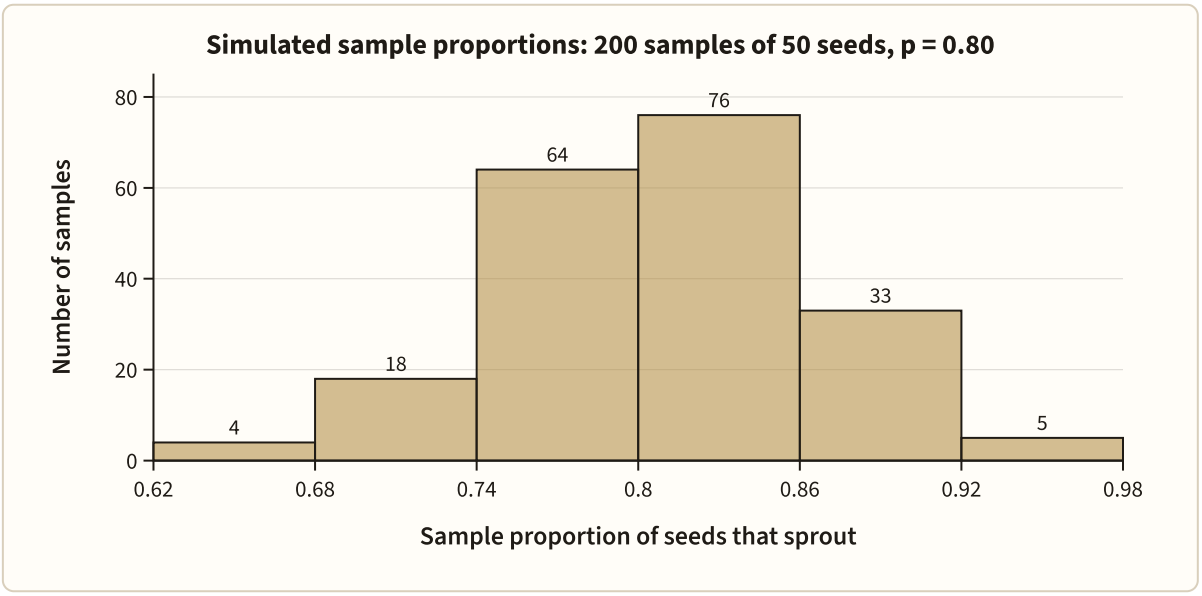
8. **MEDIUM** A seed company simulated 500 random samples of seeds from a lot in which 80% sprout, once with samples of 25 seeds and once with samples of 100 seeds. One histogram of $\hat{p}$ runs from about 0.60 to 0.96; the other runs from about 0.70 to 0.90. Which histogram is for samples of 100 seeds? Explain why.

9. MEDIUM Ruth takes a random sample of 100 households in a town of 2,000 households and asks whether they visited the public library last month. Henry takes a random sample of 100 households in a city of 200,000 households and asks the same question. Whose sample proportion will have less sampling variability? Explain.

10. HARD Five pumpkins in a garden weigh 8, 10, 12, 14 and 16 pounds. Treat them as the population. Consider all 10 samples of 3 pumpkins and use the heaviest pumpkin in the sample (the sample maximum) to estimate the population maximum. (a) Find the mean of the 10 sample maximums. (b) Is the sample maximum a biased estimator of the population maximum?

11. HARD Five bags of potatoes weigh 12, 14, 16, 18 and 20 pounds, so $\mu = 16$ pounds. Consider all 10 samples of 3 bags. (a) How many of the 10 samples have a sample mean within 1 pound of μ (from 15 to 17 pounds)? (b) How many have a sample median within 1 pound of μ? (c) Which statistic has less variability as an estimator of μ here?

12. HARD A seed company claims 80% of its sunflower seeds sprout. The histogram shows $\hat{p}$ for 200 simulated random samples of 50 seeds when $p = 0.80$. (With 50 seeds, $\hat{p}$ is always a multiple of 0.02, so the first class holds only the values 0.62, 0.64 and 0.66.) Grace planted a random sample of 50 seeds and 32 sprouted. (a) How many of the 200 simulated samples gave $\hat{p}$ of 0.66 or less? (b) What percent of the samples is that? (c) What does Grace's result suggest about the company's claim?



13. CHALLENGE The Hale family's 6 lambs weigh 40, 42, 44, 46, 48 and 50 pounds, so $\mu = 45$ pounds. (a) How many samples of 2 lambs are there, and how many samples of 4 lambs? (b) Find the largest possible sample mean for a sample of 2 lambs and for a sample of 4 lambs. (c) Every sample of 4 leaves out exactly 2 lambs. The 6 weights add to 270. Show that the mean of a sample of 4 is always exactly half as far from 45 as the mean of the 2 lambs left out (on the other side), and explain what this says about variability.

COMMON MISTAKES

- Mixing up parameters and statistics. Ask whether the number comes from the whole population (parameter: p, μ, σ) or from the sample in hand (statistic: $\hat{p}$, $\bar{x}$, s).
- Thinking "unbiased" means every sample gives the right answer. It means the statistic is right on average over all samples; any one sample can still miss because of sampling variability.
- Believing a bigger sample cures bias. A larger sample reduces variability, but a badly chosen sample (convenience or voluntary response) stays biased no matter how large it is.
- Thinking a large population needs a much larger sample. As long as the population is at least 10 times the sample, the variability depends on the sample size, not the population size.

ANSWERS TO PRACTICE

1. A parameter, $\mu = 52$ pounds, because it describes all 80 hives, the whole population of interest, not a sample.
2. (a) $\hat{p} = \dfrac{3}{20} = 0.15$. (b) A statistic, $\bar{x} = 96.1$ inches, since it comes from the sample of 20 boards.
3. (a) $\binom{5}{2} = 10$ samples. (b) $\binom{5}{3} = 10$ samples (each sample of 3 leaves out a pair of 2).
4. Not necessarily. Different random samples contain different families, so they give different values of $\bar{x}$. This is sampling variability, a natural result of chance, not an error.
5. (a) $\mu = \dfrac{2.2}{4} = 0.55$ inch. (b) Sample means 0.3, 0.4, 0.6, 0.5, 0.7, 0.8; their mean is $\dfrac{3.3}{6} = 0.55$ inch, equal to μ. (c) 2 of them (0.5 and 0.6); the others miss by 0.15 inch or more.
6. Samples with 0 heifers: $\binom{3}{2} = 3$ ($\hat{p} = 0$); with 1 heifer: $2 \times 3 = 6$ ($\hat{p} = 0.5$); with 2 heifers: 1 ($\hat{p} = 1$). (a) $\dfrac{6}{10} = 0.6$. (b) $\dfrac{3(0) + 6(0.5) + 1(1)}{10} = \dfrac{4}{10} = 0.4$. (c) Yes, the mean of the sampling distribution equals $p = 0.4$.
7. Methods A and C are centered at 96, so they appear unbiased; Method B is centered near 97, so it overestimates (biased). Method B has the least variability (IQR 0.8 inch), with Method C close behind (IQR 1.0 inch); Method A is far more spread out. Method C is best: like arrows grouped tightly around the bull's-eye, it is both unbiased and has low variability, while B's tight group is centered on the wrong spot.
8. The narrower one, from 0.70 to 0.90. Larger random samples give statistics with less variability, because in a big sample the chance runs of sprouting and non-sprouting seeds balance out more, so $\hat{p}$ stays closer to $p = 0.80$. Both are centered near 0.80.
9. About the same. Both populations are at least 10 times the sample size ($100 \le 0.10 \times 2{,}000 = 200$), and then the variability depends on the sample size, 100 in both cases, not on the population size.

10. Maximum 12: 1 sample; maximum 14: $\binom{3}{2} = 3$ samples; maximum 16: $\binom{4}{2} = 6$ samples. (a) $\dfrac{12(1) + 14(3) + 16(6)}{10} = \dfrac{150}{10} = 15$ pounds. (b) Yes; the center 15 is below the true maximum of 16 pounds, and a sample maximum can never exceed it, so it underestimates.
11. Sample means: 14, 14.67, 15.33, 15.33, 16, 16.67, 16, 16.67, 17.33, 18. Sample medians: 14 (3 samples), 16 (4 samples), 18 (3 samples). (a) 6. (b) 4. (c) The sample mean: its values cluster more tightly around 16, because averaging all three values lets highs and lows cancel, while the median uses only the middle value.
12. (a) 4 samples. (b) $\dfrac{4}{200} = 2\%$. (c) Grace's $\hat{p} = \dfrac{32}{50} = 0.64$ is as low as only about 2% of samples when $p = 0.80$. A result this low would be unusual if the claim were true, so it gives some reason to doubt that 80% of the seeds sprout.
13. (a) $\binom{6}{2} = 15$ and $\binom{6}{4} = 15$. (b) $\dfrac{48 + 50}{2} = 49$ pounds; $\dfrac{44 + 46 + 48 + 50}{4} = 47$ pounds. (c) If the 2 left-out lambs have mean m, they weigh $2m$ in total, so the 4 in the sample have mean $\dfrac{270 - 2m}{4}$, and $\dfrac{270 - 2m}{4} - 45 = \dfrac{90 - 2m}{4} = \dfrac{45 - m}{2}$. So every sample mean of size 4 lies half as far from μ as some sample mean of size 2: the sampling distribution for $n = 4$ has exactly half the spread. Larger samples, less variability.

LESSON 9.2 · 2 SESSIONS OF 45 MINUTES

The Sampling Distribution of a Sample Proportion

OBJECTIVES

- Find the mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$ of the sampling distribution of $\hat{p}$, and explain where they come from.
- Check the 10% condition and the Large Counts condition, and explain what each one is for.
- Explain why the spread of $\hat{p}$ shrinks like $\frac{1}{\sqrt{n}}$, so that four times the sample size cuts the spread in half.
- Compute and interpret normal probabilities for $\hat{p}$ in context.

VIDEO

WATCH

The Sampling Distribution of the Sample Proportion

jbstatistics · 9:49

When to watch: Before Example 1: the mean, standard deviation and shape of the sampling distribution of $\hat{p}$.

https://www.youtube.com/watch?v=fuGwbG9_W1c

BACKUP

Statistical Proportions (1 of 2: What are they?)

Eddie Woo · 11:17

When to watch: Review: sample proportions.

https://www.youtube.com/watch?v=ueXB7XUVNUs

KEY TERMS

Sample proportion $\hat{p}$	The fraction of the sample with the characteristic of interest: $\hat{p} = \frac{X}{n}$, where X is the count of successes in a sample of size n.
Standard deviation of $\hat{p}$	$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}$: the typical distance between the sample proportion and the true proportion p over many random samples.
10% condition	When we sample without replacement, the sample size should be at most 10% of the population ($n \le 0.10N$), so the standard deviation formula is accurate.

Large Counts condition $np \geq 10$ and $n(1-p) \geq 10$. When both hold, the sampling distribution of $\hat{p}$ is approximately normal.

CORE IDEAS

A sample proportion is a count divided by the sample size: $\hat{p} = \dfrac{X}{n}$. If we take a random sample from a large population in which a proportion p are "successes", the count X is very nearly a binomial random variable (Unit 8), with mean np and standard deviation $\sqrt{np(1-p)}$. Everything about $\hat{p}$ follows from that.

Center. Dividing a random variable by n divides its mean by n (Lesson 7.3). So $\mu_{\hat{p}} = \dfrac{np}{n} = p$. The sampling distribution of $\hat{p}$ is centered exactly on the true proportion: $\hat{p}$ is an unbiased estimator of p. Some samples overestimate and some underestimate, but there is no systematic lean either way.

Spread. Dividing by n also divides the standard deviation by n: $\sigma_{\hat{p}} = \dfrac{\sqrt{np(1-p)}}{n} = \sqrt{\dfrac{np(1-p)}{n^2}} = \sqrt{\dfrac{p(1-p)}{n}}$. The binomial model assumes independent trials. When we sample without replacement the trials are not quite independent, but if the population is at least 10 times the sample (the **10% condition**), the formula is accurate enough to use.

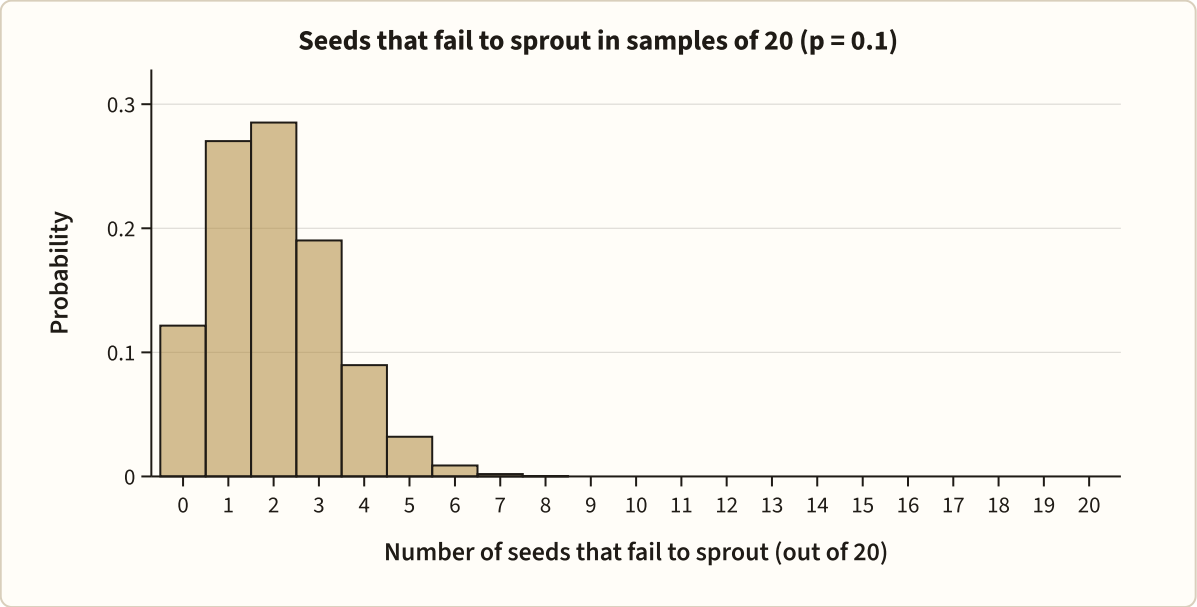
Because n sits under a square root, the spread shrinks like $\dfrac{1}{\sqrt{n}}$, not like $\dfrac{1}{n}$. To cut $\sigma_{\hat{p}}$ in half, you must make n **four** times as large, since $\sqrt{4} = 2$. The table shows this for $p = 0.5$. Going from 25 to 100 seeds is cheap and buys a lot of precision; going from 400 to 1,600 costs 1,200 more seeds to shave off just 0.0125. This is why careful surveys of a few thousand people are common, but surveys of a million are not.

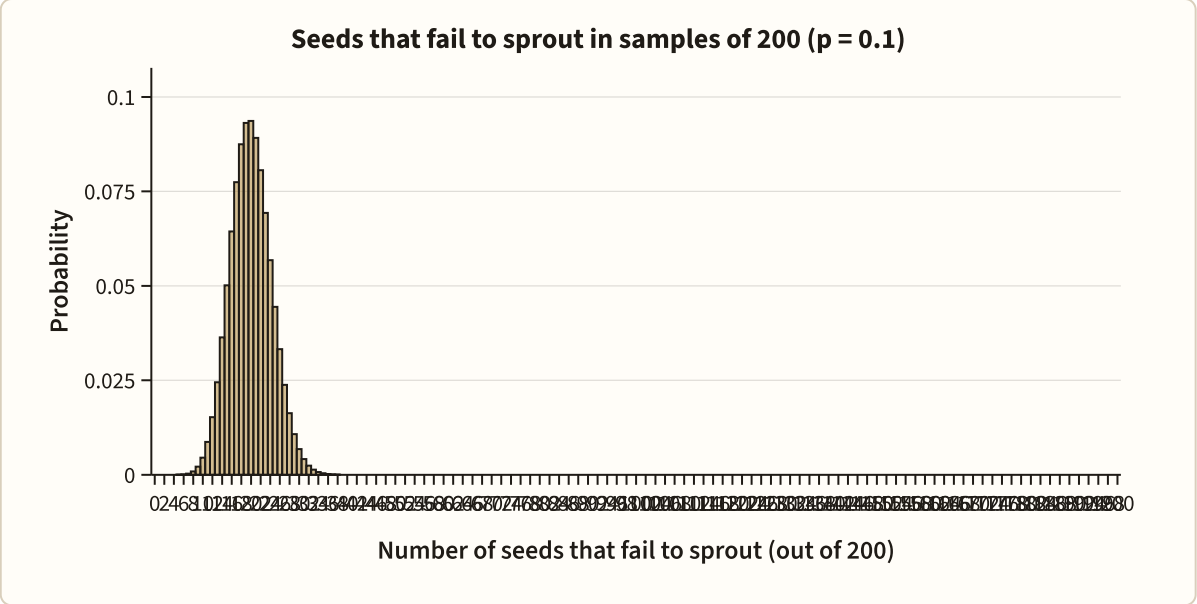
Standard deviation of the sample proportion when p = 0.5

Sample size n	Standard deviation of p-hat
25	0.1
100	0.05
400	0.025
1600	0.0125

Shape. The two probability histograms below show the count X of seeds that fail to sprout when $p = 0.1$, first for samples of 20 seeds and then for samples of 200. (Dividing by n to get $\hat{p}$ only relabels the axis; the shape stays the same.) With $n = 20$ we expect only $np = 2$ failures, the distribution is piled against 0 and strongly skewed right. With $n = 200$ we expect $np = 20$ failures and $n(1-p) = 180$ sprouts, and the histogram is close to a normal curve. The **Large Counts condition**, $np \geq 10$ and $n(1-$

$p) \geq 10$, is the rule of thumb that tells us when the normal shape is close enough to use, just as in Lesson 8.3.





Once the conditions are checked, we find probabilities about $\hat{p}$ just as in Unit 3: standardize with $z = \dfrac{\hat{p} - p}{\sigma_{\hat{p}}}$, draw and shade the curve, and use the table or technology. Notice that in this lesson we **know** p and ask how sample results behave. In Units 10 and 11 we turn this around: we know $\hat{p}$ from one sample and reason back to the unknown p. That reasoning only works because we first understand how $\hat{p}$ varies when p is known.

KEY FORMULAS

$$\mu_{\hat{p}} = p$$

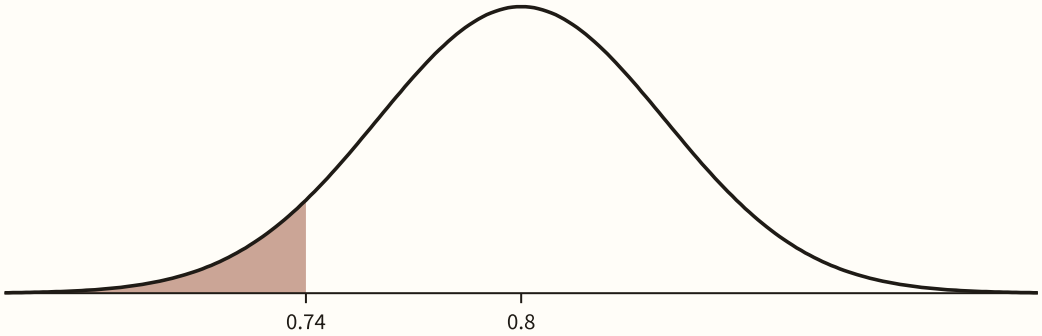
$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}} \quad (10\% \text{ condition: } n \leq 0.10N)$$

$$\text{Large Counts: } np \geq 10 \text{ and } n(1-p) \geq 10$$

$$z = \frac{\hat{p} - p}{\sigma_{\hat{p}}}$$

Practice habit. Check the conditions before calculating: the 10% condition before you use the standard deviation formula, and Large Counts before you use the normal curve. Then draw and shade the curve.

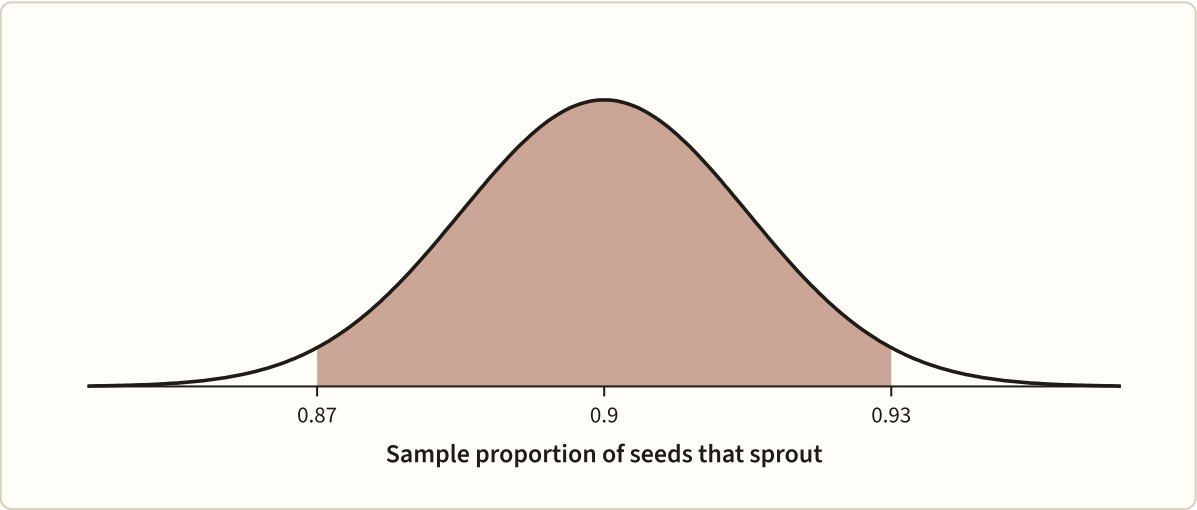
Example 1. A hatchery's records show that 80% of its fertile eggs hatch. A farmer buys a random sample of 100 eggs from the hatchery's stock of many thousands and incubates them. Let $\hat{p}$ be the proportion that hatch. (a) Find the mean and standard deviation of the sampling distribution of $\hat{p}$. (b) Check that the sampling distribution is approximately normal. (c) Find the probability that fewer than 74% of the eggs hatch. Round to 4 decimal places.



1. Center: $\mu_{\hat{p}} = p = 0.80$.
2. 10% condition: 100 eggs is far less than 10% of many thousands, so we may use the formula.
3. Spread: $\sigma_{\hat{p}} = \sqrt{\dfrac{0.80(0.20)}{100}} = \sqrt{0.0016} = 0.04$.
4. Large Counts: $np = 100(0.80) = 80 \geq 10$ and $n(1-p) = 100(0.20) = 20 \geq 10$, so $\hat{p}$ is approximately normal: $N(0.80, 0.04)$.
5. Standardize: $z = \dfrac{0.74 - 0.80}{0.04} = \dfrac{-0.06}{0.04} = -1.50$.
6. Area to the left of $z = -1.50$: $P(\hat{p} < 0.74) = 0.0668$.
7. Conclusion: in about 7% of random samples of 100 eggs, fewer than 74% would hatch. That would be a disappointing batch, but not a shocking one.

Answer: (a) Mean 0.80, standard deviation 0.04. (b) $np = 80$ and $n(1-p) = 20$ are both at least 10. (c) $z = -1.50$, $P(\hat{p} < 0.74) = 0.0668$.

Example 2. A seed company states that 90% of its tomato seeds sprout. Grace plants a random sample of 400 seeds from a lot of many thousands. (a) Find the probability that the proportion of Grace's seeds that sprout is between 0.87 and 0.93. Round to 4 decimal places. (b) How many seeds would Grace need to plant to make the standard deviation of $\hat{p}$ half as large?



1. Center: $\mu_{\hat{p}} = 0.90$. 10% condition: 400 seeds is less than 10% of many thousands.
2. Spread: $\sigma_{\hat{p}} = \sqrt{\dfrac{0.90(0.10)}{400}} = \sqrt{0.000225} = 0.015$.
3. Large Counts: $400(0.90) = 360 \geq 10$ and $400(0.10) = 40 \geq 10$, so $\hat{p}$ is approximately normal.
4. Standardize both ends: $z = \dfrac{0.87 - 0.90}{0.015} = -2.00$ and $z = \dfrac{0.93 - 0.90}{0.015} = 2.00$.
5. $P(-2.00 < Z < 2.00) = 0.9545$ (table: $0.9772 - 0.0228 = 0.9544$).
6. (b) $\sigma_{\hat{p}}$ shrinks like $\dfrac{1}{\sqrt{n}}$, so halving it takes $4 \times 400 = 1{,}600$ seeds. Check: $\sqrt{\dfrac{0.09}{1600}} = 0.0075$, half of 0.015.
7. Conclusion: about 95% of random samples of 400 seeds would have between 87% and 93% sprouting; planting 1,600 seeds would make the typical miss half as large.

Answer: (a) $\sigma_{\hat{p}} = 0.015$, $z = \pm 2.00$, probability $= 0.9545$ (table: 0.9544). (b) $4 \times 400 = 1{,}600$ seeds.

PRACTICE

1. EASY In a parish library of 5,000 books, 40% are hardcover. The librarian chooses a random sample of 96 books. Find the mean and standard deviation of the sampling distribution of $\hat{p}$, the sample proportion of hardcover books.
2. EASY At a large egg farm, 5% of eggs are cracked. An inspector takes a random sample of 100 eggs. Is the sampling distribution of $\hat{p}$, the proportion cracked, approximately normal? Explain.
3. EASY A rancher has 600 cattle and wants to estimate the proportion that are polled (born without horns) by checking a random sample of 80 of them. Is the 10% condition met?

4. EASY For random samples of 50 apple trees in a large county, the standard deviation of the sample proportion of trees with fire blight is 0.06. What is the standard deviation for random samples of 200 trees?

5. MEDIUM In a large orchard, 20% of the apples have blemishes. A worker picks a random sample of 400 apples. Find the probability that more than 25% of the sample have blemishes. Round to 4 decimal places.

6. MEDIUM At a large poultry farm, 75% of the eggs are graded AA. A buyer selects a random sample of 300 eggs. Find the probability that less than 70% of the sample are grade AA. Round to 4 decimal places.

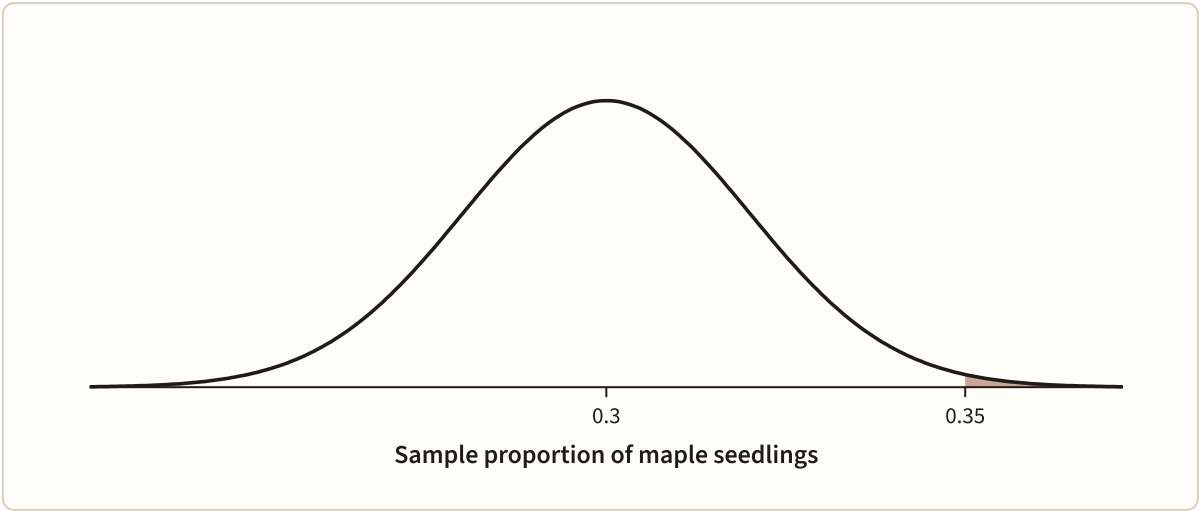
7. MEDIUM At a large nursery, 60% of the strawberry plants are an everbearing variety. A customer buys a random sample of 150 plants. Find the probability that between 55% and 65% of them are everbearing. Round to 4 decimal places.

8. MEDIUM In Problem 7, the standard deviation of $\hat{p}$ is 0.04. Interpret this number in context.

9. MEDIUM Henry says, "If I double my sample size from 150 to 300, the standard deviation of $\hat{p}$ will drop from 0.04 to 0.02." Explain his mistake and find the correct standard deviation. Round to 4 decimal places.

10. HARD A beekeeping club wants a random sample of its members' hives large enough that the standard deviation of $\hat{p}$ is 0.02 when $p = 0.5$. How many hives must it sample? (The club's members keep many thousands of hives.)

11. HARD A tree nursery has 20,000 seedlings, and 30% of them are maples. A crew selects a random sample of 525 seedlings. (a) Find the probability that 35% or more of the sample are maples. Round to 4 decimal places. (b) Would such a sample be surprising?



12. HARD In a large forest, half the trees are oaks. A forester takes a random sample of trees and records $\hat{p}$, the proportion of oaks. Find the probability that $\hat{p}$ is within 0.05 of 0.5 (a) for a sample of 100 trees and (b) for a sample of 400 trees. Round to 4 decimal places.

13. **CHALLENGE** A garden center knows that 80% of its tulip bulbs bloom. It sells bulbs in random lots of 100 drawn from a supply of many thousands. Below what proportion of blooming bulbs will only the worst 5% of lots fall? Round to 4 decimal places.

COMMON MISTAKES

- Thinking that doubling the sample halves the spread. The standard deviation shrinks like $\frac{1}{\sqrt{n}}$: it takes four times the sample to cut it in half.
- Using the normal curve without checking Large Counts. If np or $n(1-p)$ is below 10, the sampling distribution is skewed, and normal calculations can be badly off.
- Using $\sqrt{\frac{p(1-p)}{n}}$ when the sample is more than 10% of the population. Check $n \le 0.10N$ first.
- Plugging $\hat{p}$ into the standard deviation when p is known. In this lesson the parameter p is given, so use p in both the mean and the standard deviation.

ANSWERS TO PRACTICE

1. Mean $= 0.40$. Since $96 \le 0.10(5{,}000) = 500$, $\sigma_{\hat{p}} = \sqrt{\frac{0.40(0.60)}{96}} = \sqrt{0.0025} = 0.05$.
2. No. $np = 100(0.05) = 5$, which is less than 10, so the Large Counts condition fails; the distribution is skewed right.
3. No. 10% of 600 is 60, and the sample of 80 is larger than that, so $\sqrt{\frac{p(1-p)}{n}}$ would overstate the spread.
4. Four times the sample size cuts the standard deviation in half: $\frac{0.06}{2} = 0.03$.
5. $\sigma_{\hat{p}} = \sqrt{\frac{0.20(0.80)}{400}} = 0.02$; Large Counts: 80 and 320. $z = \frac{0.25 - 0.20}{0.02} = 2.50$; $P(\hat{p} > 0.25) = 1 - 0.9938 = 0.0062$.
6. $\sigma_{\hat{p}} = \sqrt{\frac{0.75(0.25)}{300}} = \sqrt{0.000625} = 0.025$; Large Counts: 225 and 75. $z = \frac{0.70 - 0.75}{0.025} = -2.00$; $P(\hat{p} < 0.70) = 0.0228$.
7. $\sigma_{\hat{p}} = \sqrt{\frac{0.60(0.40)}{150}} = \sqrt{0.0016} = 0.04$; Large Counts: 90 and 60. $z = \pm\frac{0.05}{0.04} = \pm 1.25$; probability $= 0.7887$ (table: $0.8944 - 0.1056 = 0.7888$).
8. If the customer took many random samples of 150 plants, the proportion of everbearing plants in a sample would typically differ from the true 0.60 by about 0.04, that is, by about 4 percentage points.
9. The standard deviation shrinks like $\frac{1}{\sqrt{n}}$, not $\frac{1}{n}$. Doubling n divides it by $\sqrt{2}$: $\frac{0.04}{\sqrt{2}} \approx 0.0283$. To reach 0.02 he would need 4 times the sample, 600.
10. $\sqrt{\frac{0.25}{n}} = 0.02$, so $\frac{0.25}{n} = 0.0004$ and $n = \frac{0.25}{0.0004} = 625$ hives.

11. 10%: $525 \le 2{,}000$. $\sigma_{\hat{p}} = \sqrt{\dfrac{0.30(0.70)}{525}} = \sqrt{0.0004} = 0.02$; Large Counts: 157.5 and 367.5. $z = \dfrac{0.35 - 0.30}{0.02} = 2.50$; (a) $P(\hat{p} \ge 0.35) = 0.0062$. (b) Yes: only about 6 in 1,000 random samples would contain that many maples.
12. (a) $\sigma_{\hat{p}} = \sqrt{\dfrac{0.25}{100}} = 0.05$, $z = \pm 1.00$, probability $= 0.6827$ (table: 0.6826). (b) $\sigma_{\hat{p}} = \sqrt{\dfrac{0.25}{400}} = 0.025$, $z = \pm 2.00$, probability $= 0.9545$ (table: 0.9544). Four times the trees turns a 68% chance of landing close into a 95% chance.
13. $\sigma_{\hat{p}} = \sqrt{\dfrac{0.80(0.20)}{100}} = 0.04$; Large Counts: 80 and 20. The bottom 5% has $z = -1.645$, so the cutoff is $0.80 - 1.645(0.04) \approx 0.7342$. Only about 5% of lots have fewer than about 73% of bulbs bloom.

LESSON 9.3 · 2 SESSIONS OF 45 MINUTES

The Sampling Distribution of a Sample Mean

OBJECTIVES

- Find the mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$ of the sampling distribution of $\bar{x}$, and explain why the spread is smaller than σ.
- Explain why averages vary less than individual values, using the rule that variances add.
- State that $\bar{x}$ is exactly normal when the population is normal.
- Compute and interpret normal probabilities for $\bar{x}$ in context, and contrast them with probabilities for a single value.

VIDEO

WATCH

Introduction to Sampling & Populations (2 of 4: Distribution of Sample Means)

Eddie Woo · 12:29

When to watch: Before Example 1: the distribution of sample means.

https://www.youtube.com/watch?v=TEOW-9vEnXQ

BACKUP

The Sampling Distribution of the Sample Mean

jbstatistics · 11:39

When to watch: Review: the mean and standard deviation of x̄.

https://www.youtube.com/watch?v=q50GpTdFYyI

KEY TERMS

Sample mean $\bar{x}$	The average of the n values in a sample. As a statistic, it changes from sample to sample.
Mean of $\bar{x}$	$\mu_{\bar{x}} = \mu$: over all possible samples, the sample means average out to the population mean.
Standard deviation of $\bar{x}$	$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}$: the typical distance between a sample mean and the population mean μ over many random samples of size n.
Population distribution	The distribution of the individual values in the population, with mean μ and standard deviation σ. Do not confuse it with the sampling distribution of $\bar{x}$.

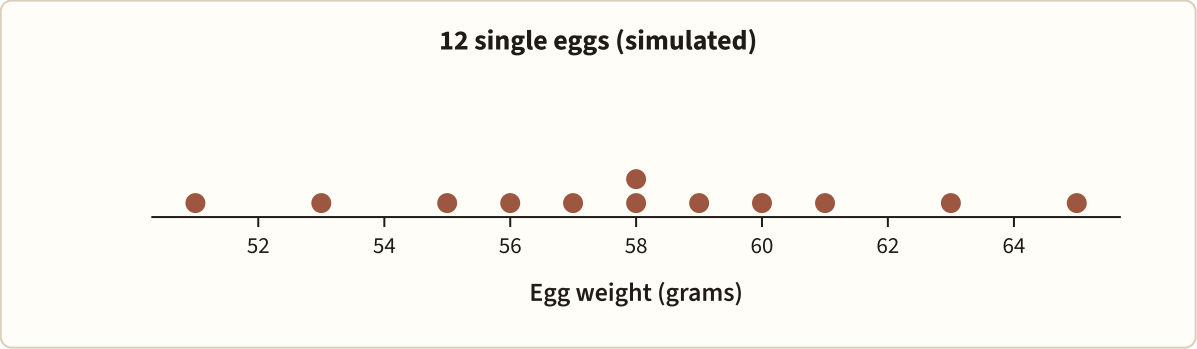
CORE IDEAS

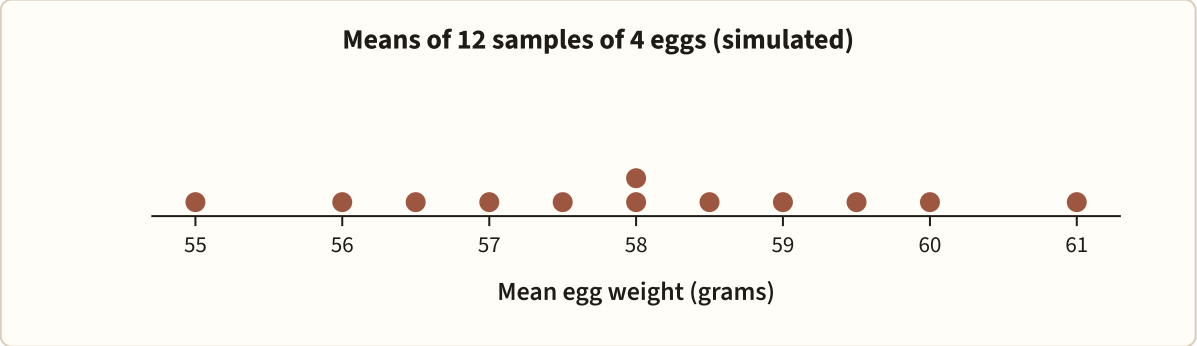
In Lesson 9.2 we studied a sample proportion. Now we study the other workhorse statistic, the sample mean $\bar{x}$. Think of a random sample as n independent draws $X_1, X_2, \ldots, X_n$ from a population with mean μ and standard deviation σ, and $\bar{x} = \dfrac{X_1 + X_2 + \cdots + X_n}{n}$. The rules for combining random variables from Lesson 7.3 tell us exactly how $\bar{x}$ behaves.

Center. Means always add, so the sum $X_1 + \cdots + X_n$ has mean $n\mu$. Dividing by n gives $\mu_{\bar{x}} = \dfrac{n\mu}{n} = \mu$. The sample mean is an unbiased estimator of the population mean.

Spread. For independent random variables, variances add. Each draw has variance σ^2, so the sum has variance $n\sigma^2$ and standard deviation $\sigma\sqrt{n}$. Dividing by n divides the standard deviation by n: $\sigma_{\bar{x}} = \dfrac{\sigma\sqrt{n}}{n} = \dfrac{\sigma}{\sqrt{n}}$. Here is the key: the spread of the total grows, but only like $\sqrt{n}$, while we divide by the full n. So the average gets steadier as n grows. Independence needs a random sample, and when we sample without replacement, the 10% condition ($n \le 0.10N$).

Why do averages vary less than single values? A single egg can be unusually light or heavy. In a carton, a heavy egg is likely to sit beside a light one, and in the average they partly cancel. The two dotplots below come from a simulation with eggs whose weights have mean 58 grams: the first shows 12 single eggs, the second shows the means of 12 random samples of 4 eggs. Both center at 58 grams, but the means are packed about twice as tightly, just as $\dfrac{\sigma}{\sqrt{4}} = \dfrac{\sigma}{2}$ predicts.



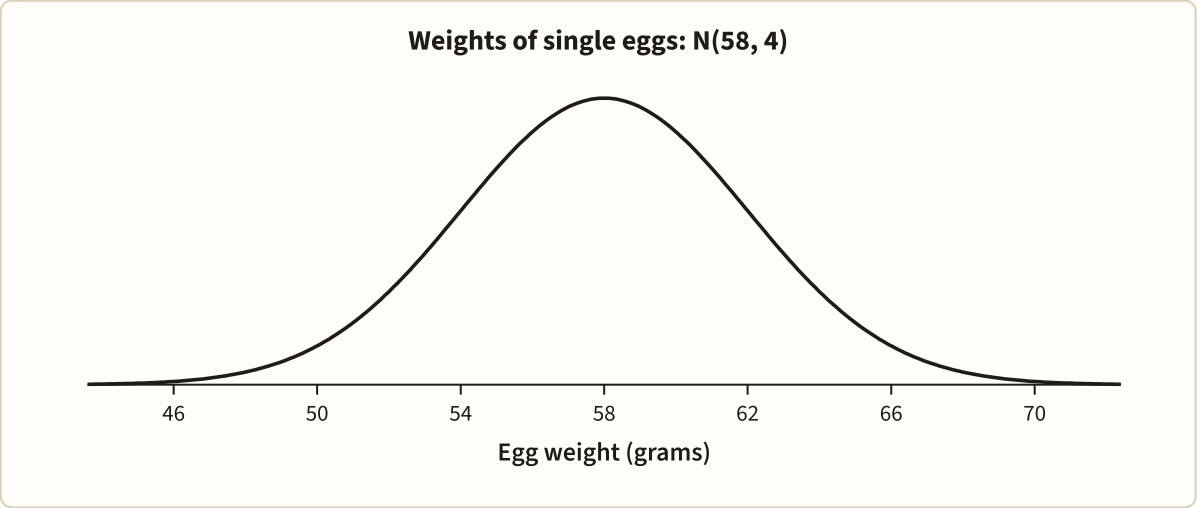


As with proportions, the spread shrinks like $\dfrac{1}{\sqrt{n}}$: to cut $\sigma_{\bar{x}}$ in half, you need four times as many observations. The table shows this for a population with $\sigma = 4$ grams.

Standard deviation of the sample mean when sigma = 4 grams

Sample size n	Standard deviation of x-bar (grams)
1	4
4	2
16	1
25	0.8
64	0.5
100	0.4

Shape. If the population itself is normal, then $\bar{x}$ is **exactly** normal for any sample size, because a sum of independent normal random variables is normal (Lesson 7.4), and dividing by n keeps it normal. Suppose egg weights from a large flock are $N(58, 4)$ grams, as in the curve below. For one egg, $P(X > 59.2)$ has $z = \dfrac{59.2 - 58}{4} = 0.30$, so the probability is $1 - 0.6179 = 0.3821$. For the mean of 25 eggs, $\sigma_{\bar{x}} = \dfrac{4}{\sqrt{25}} = 0.8$ and $z = \dfrac{59.2 - 58}{0.8} = 1.50$, so $P(\bar{x} > 59.2) = 0.0668$. Over a third of single eggs are that heavy, but only about 7% of cartons average that much.



One caution: $\dfrac{\sigma}{\sqrt{n}}$ describes how much **sample means** vary from sample to sample. It does not say the eggs inside one carton are less variable; the individual eggs still vary with standard deviation about σ. Always ask whether a question is about one value or about the mean of several.

KEY FORMULAS

$$\mu_{\bar{x}} = \mu$$

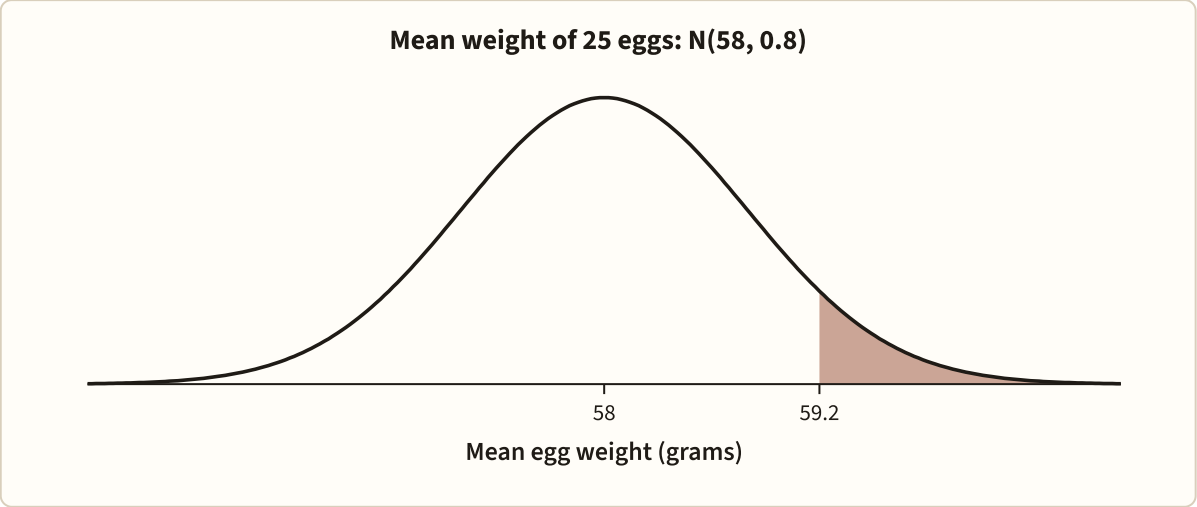
$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}} \quad (10\% \text{ condition: } n \leq 0.10N)$$

$$z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}}$$

$$\text{Population } N(\mu, \sigma) \implies \bar{x} \text{ is } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \text{ for every } n$$

Practice habit. Before standardizing, ask: "Is this about ONE value or about the MEAN of n values?" One value uses σ; a mean uses $\frac{\sigma}{\sqrt{n}}$.

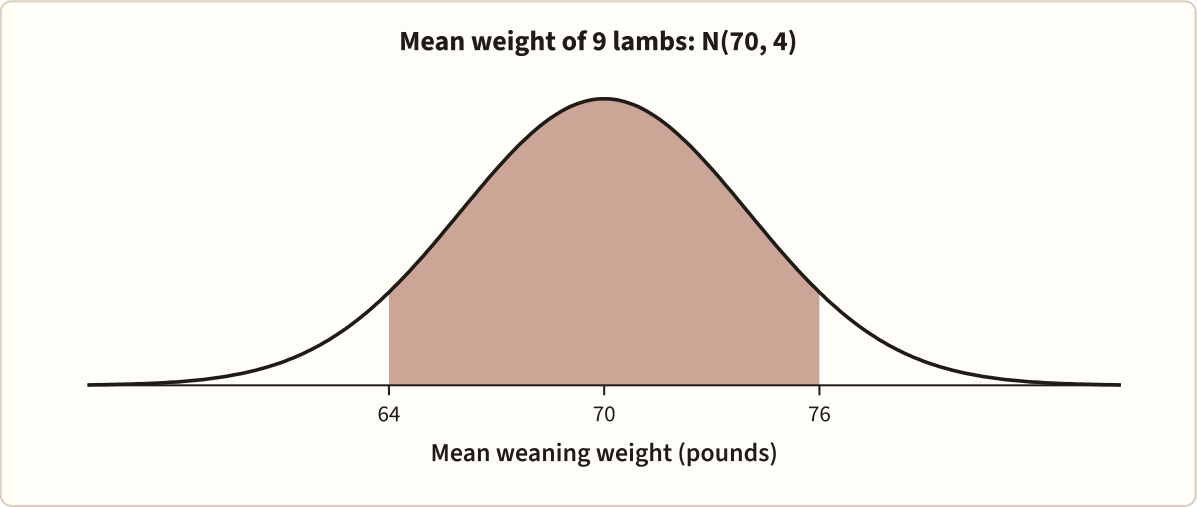
Example 1. The weights of eggs from a large flock are normally distributed with mean 58 grams and standard deviation 4 grams. A carton holds a random sample of 25 eggs. (a) Find the mean and standard deviation of the sampling distribution of $\bar{x}$, the mean weight of the eggs in a carton. (b) Find the probability that the mean weight of the carton's eggs is more than 59.2 grams. Round to 4 decimal places.



1. Center: $\mu_{\bar{x}} = \mu = 58$ grams.
2. 10% condition: 25 eggs is far less than 10% of a large flock's production.
3. Spread: $\sigma_{\bar{x}} = \dfrac{4}{\sqrt{25}} = \dfrac{4}{5} = 0.8$ gram.
4. Shape: the population is normal, so $\bar{x}$ is exactly normal: $N(58, 0.8)$.
5. Standardize: $z = \dfrac{59.2 - 58}{0.8} = \dfrac{1.2}{0.8} = 1.50$.
6. Area to the right: $P(\bar{x} > 59.2) = 1 - 0.9332 = 0.0668$.
7. Conclusion: only about 7% of cartons of 25 eggs would have a mean weight above 59.2 grams, even though over a third of single eggs weigh that much.

Answer: (a) Mean 58 grams, standard deviation 0.8 gram. (b) $z = 1.50$, $P(\bar{x} > 59.2) = 0.0668$.

Example 2. Weaning weights of lambs on a large sheep ranch are normally distributed with mean 70 pounds and standard deviation 12 pounds. (a) Find the probability that one randomly chosen lamb weighs between 64 and 76 pounds. (b) Find the probability that the mean weight of a random sample of 9 lambs is between 64 and 76 pounds. Round both to 4 decimal places, and explain why they differ.



1. (a) One lamb: use $\sigma = 12$. $z = \dfrac{64 - 70}{12} = -0.50$ and $z = \dfrac{76 - 70}{12} = 0.50$.
2. $P(-0.50 < Z < 0.50) = 0.3829$ (table: $0.6915 - 0.3085 = 0.3830$).
3. (b) Mean of 9 lambs: $\sigma_{\bar{x}} = \dfrac{12}{\sqrt{9}} = 4$ pounds, and $\bar{x}$ is normal because the population is normal.
4. $z = \dfrac{64 - 70}{4} = -1.50$ and $z = \dfrac{76 - 70}{4} = 1.50$.
5. $P(-1.50 < Z < 1.50) = 0.9332 - 0.0668 = 0.8664$.
6. Conclusion: only about 38% of single lambs weigh within 6 pounds of 70, but about 87% of samples of 9 lambs have a mean within 6 pounds of 70. Heavy and light lambs in a sample partly cancel, so the mean stays closer to μ.

Answer: (a) 0.3829 (table: 0.3830). (b) $\sigma_{\bar{x}} = 4$, $z = \pm 1.50$, probability $= 0.8664$. The mean of 9 lambs varies only one third as much as a single lamb.

PRACTICE

1. EASY Corn yields on the many test plots of a seed farm have mean 180 bushels per acre and standard deviation 20 bushels per acre. Find the mean and standard deviation of the sampling distribution of $\bar{x}$, the mean yield of a random sample of 16 plots.
2. EASY A honey packer's jars hold amounts with standard deviation 0.3 ounce. Find the standard deviation of the mean fill of a random sample of (a) 9 jars and (b) 36 jars.

3. EASY A sawmill cuts boards whose lengths are normal with mean 96 inches and standard deviation 0.5 inch. Which is more likely: that one randomly chosen board is longer than 97 inches, or that the mean length of 10 randomly chosen boards is longer than 97 inches? Explain without calculating.

4. EASY A farmer wants to use $\frac{\sigma}{\sqrt{n}}$ for the mean weight of a random sample of 30 calves from a herd of 200. Is the 10% condition met?

5. MEDIUM In Problem 1, the plot yields are normally distributed. Find the probability that the mean yield of 16 randomly chosen plots is less than 170 bushels per acre. Round to 4 decimal places.

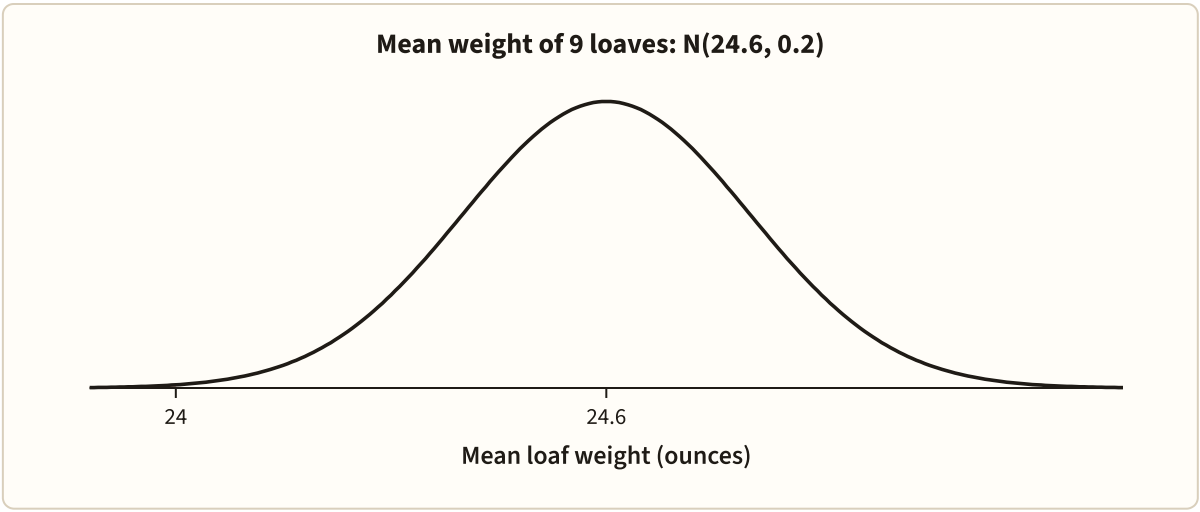
6. MEDIUM Ruth measures the brightness of a star through her telescope. Her measurements are normally distributed around the true value 4.20 (on the magnitude scale) with standard deviation 0.10. She averages 4 independent measurements. Find the probability that her average is within 0.05 of the true value. Round to 4 decimal places.

7. MEDIUM Samuel's golf drives are normally distributed with mean 220 yards and standard deviation 16 yards. Find the probability that the mean of 4 of his drives is more than 230 yards. Round to 4 decimal places.

8. MEDIUM In Problem 1, the standard deviation of $\bar{x}$ is 5 bushels per acre. Interpret this number in context, and explain why it is smaller than 20.

9. MEDIUM Lamb weights have standard deviation 12 pounds. How many lambs must be in a random sample for the standard deviation of $\bar{x}$ to be (a) 3 pounds and (b) 1.5 pounds?

10. HARD A bakery's loaves are labeled 24 ounces. Their weights are normal with mean 24.6 ounces and standard deviation 0.6 ounce. (a) Find the probability that one loaf weighs less than 24 ounces. (b) An inspector weighs a random sample of 9 loaves. Find the probability that their mean weight is less than 24 ounces. Round both to 4 decimal places.



11. HARD Egg weights from a large flock are normal with mean 58 grams and standard deviation 4 grams. A box holds a random sample of 16 eggs. Only 10% of boxes have a mean egg weight above what value? Round to 2 decimal places.

12. **HARD** Hay bales from a farm have weights that are normal with mean 50 pounds and standard deviation 5 pounds. A wagon is loaded with a random sample of 25 bales and can safely carry 1,275 pounds of hay. Find the probability that the load exceeds 1,275 pounds. Round to 4 decimal places.

13. **CHALLENGE** Egg weights have standard deviation 4 grams and are normally distributed. What is the smallest number of eggs in a random sample that makes the probability at least 0.95 that $\bar{x}$ is within 1 gram of μ? Explain.

COMMON MISTAKES

- Using $\frac{\sigma}{\sqrt{n}}$ for a question about one individual. A single egg, board or lamb varies with standard deviation σ; only a sample mean uses $\frac{\sigma}{\sqrt{n}}$.
- Dividing by n instead of $\sqrt{n}$. The standard deviation of $\bar{x}$ is $\frac{\sigma}{\sqrt{n}}$, because variances (not standard deviations) add.
- Thinking a larger sample makes the data in the sample less spread out. It makes the sample mean steadier from sample to sample; the individual values still vary with standard deviation about σ.
- Forgetting the shape. In this lesson $\bar{x}$ is normal because the population is normal; for a non-normal population with a small sample, normal calculations for $\bar{x}$ are not justified (see Lesson 9.4).

ANSWERS TO PRACTICE

1. Mean $= 180$ bushels per acre; standard deviation $= \frac{20}{\sqrt{16}} = 5$ bushels per acre.
2. (a) $\frac{0.3}{\sqrt{9}} = 0.1$ ounce. (b) $\frac{0.3}{\sqrt{36}} = 0.05$ ounce.
3. One board. The mean of 10 boards has a standard deviation of only $\frac{0.5}{\sqrt{10}} \approx 0.16$ inch, so it rarely strays an inch from 96; long and short boards in the sample cancel out. A single board has no such balancing.
4. No. 10% of 200 is 20, and 30 is more than that.
5. $\sigma_{\bar{x}} = 5$; $z = \frac{170 - 180}{5} = -2.00$; $P(\bar{x} < 170) = 0.0228$.
6. $\sigma_{\bar{x}} = \frac{0.10}{\sqrt{4}} = 0.05$; $z = \pm 1.00$; probability $= 0.6827$ (table: 0.6826).
7. $\sigma_{\bar{x}} = \frac{16}{\sqrt{4}} = 8$; $z = \frac{230 - 220}{8} = 1.25$; $P(\bar{x} > 230) = 1 - 0.8944 = 0.1056$.
8. If we took many random samples of 16 plots, the mean yield of a sample would typically differ from the true mean of 180 by about 5 bushels per acre. It is smaller than the 20 for single plots because high-yield and low-yield plots in the same sample partly cancel when we average.
9. (a) $\frac{12}{\sqrt{n}} = 3$ gives $\sqrt{n} = 4$, $n = 16$. (b) $\sqrt{n} = 8$, $n = 64$: half the spread takes four times the lambs.

10. (a) $z = \dfrac{24 - 24.6}{0.6} = -1.00$; probability $= 0.1587$. (b) $\sigma_{\bar{x}} = \dfrac{0.6}{3} = 0.2$; $z = \dfrac{24 - 24.6}{0.2} = -3.00$; probability $= 0.0013$. About 1 loaf in 6 is light, but a sample of 9 almost never averages under 24 ounces.
11. $\sigma_{\bar{x}} = \dfrac{4}{\sqrt{16}} = 1$ gram. The top 10% has $z = 1.28$ (technology 1.2816), so the cutoff is $58 + 1.2816(1) \approx 59.28$ grams.
12. Total $> 1{,}275$ means $\bar{x} > \dfrac{1{,}275}{25} = 51$ pounds. $\sigma_{\bar{x}} = \dfrac{5}{\sqrt{25}} = 1$; $z = \dfrac{51 - 50}{1} = 1.00$; probability $= 1 - 0.8413 = 0.1587$.
13. We need $1.96\,\dfrac{4}{\sqrt{n}} \le 1$, so $\sqrt{n} \ge 7.84$ and $n \ge 61.47$. Round up: $n = 62$ eggs (with 61 eggs the probability would fall just short of 0.95).

LESSON 9.4 · 2 SESSIONS OF 45 MINUTES

The Central Limit Theorem

OBJECTIVES

- State the central limit theorem and explain what it does and does not say.
- Decide whether the sampling distribution of $\bar{x}$ is approximately normal, using the shape of the population and the sample size.
- Use the central limit theorem to compute and interpret probabilities about a sample mean or a total from a non-normal population.

VIDEO

WATCH

Introduction to the Central Limit Theorem

jbstatistics · 13:13

When to watch: Before Example 1: the central limit theorem.

https://www.youtube.com/watch?v=Pujol1yC1_A

BACKUP

But what is the Central Limit Theorem?

3Blue1Brown · 31:14

When to watch: Enrichment: a visual look at why the central limit theorem works (long).

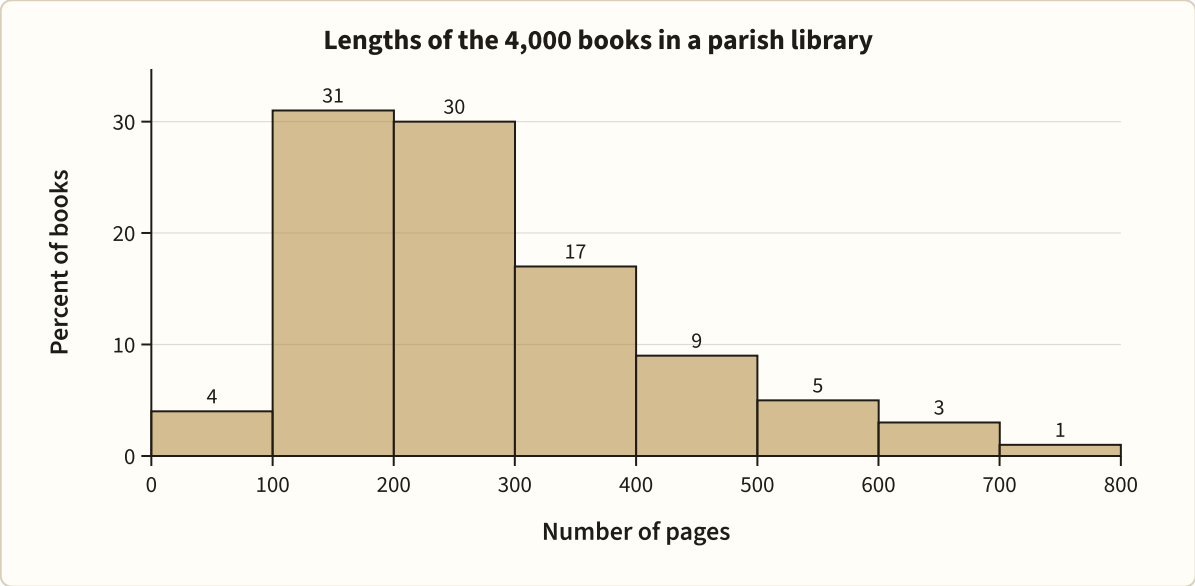
https://www.youtube.com/watch?v=zeJD6dqJ5lo

KEY TERMS

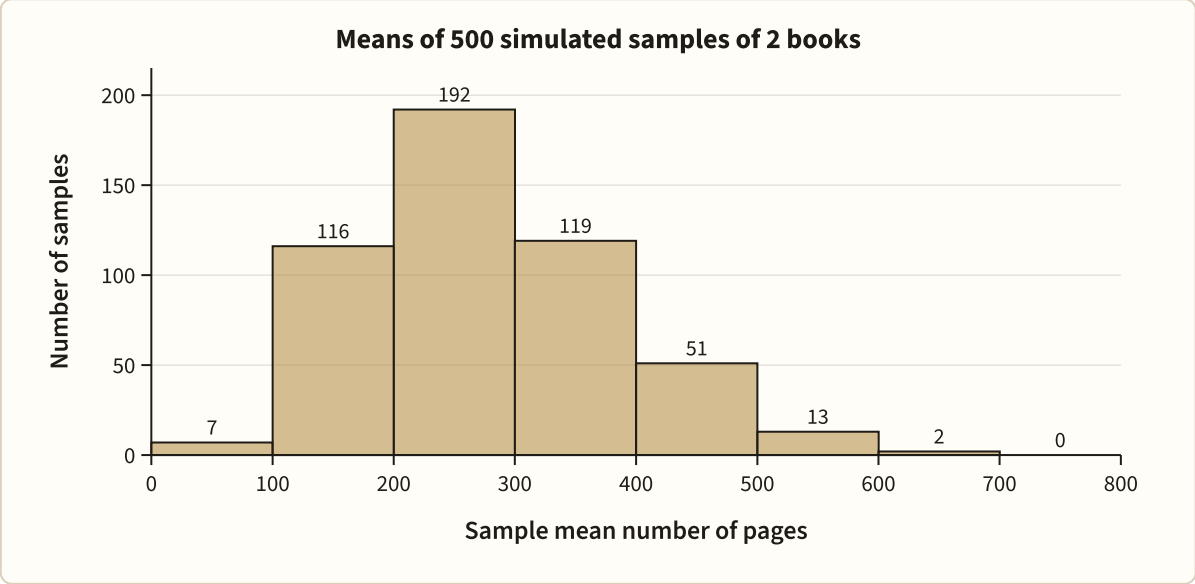
Central limit theorem (CLT)	When n is large, the sampling distribution of $\bar{x}$ is approximately normal, with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, whatever the shape of the population.
Large sample rule of thumb	In this course, $n \geq 30$ is usually large enough for the CLT, even for a strongly skewed population. Extremely skewed populations or populations with extreme outliers may need more.
Normal/Large Sample condition	We may use normal calculations for $\bar{x}$ if the population is normal (any n) or if $n \geq 30$ (any population shape).

CORE IDEAS

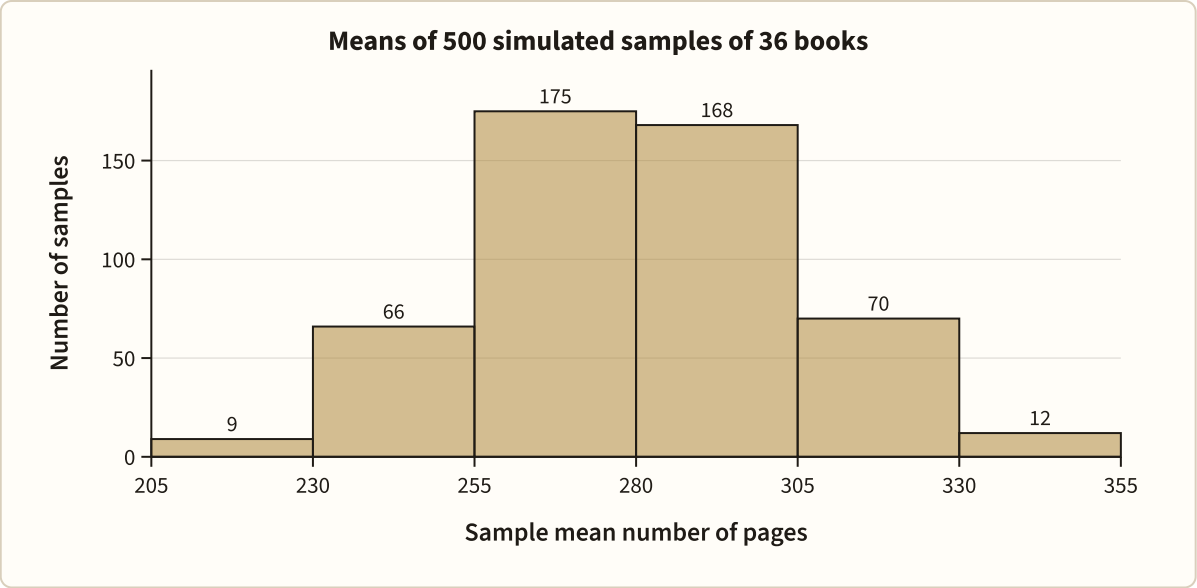
Lesson 9.3 needed a normal population. But many populations are far from normal: the lengths of books, the rain from a storm, the amount customers spend at a farm stand. These are skewed right, with many modest values and a few very large ones. What happens to the sample mean then? A simulation answers the question. The histogram below shows the lengths of the 4,000 books in a parish library: mean $\mu = 280$ pages, standard deviation $\sigma = 150$ pages, strongly skewed right.



A computer drew 500 random samples of 2 books and recorded $\bar{x}$ for each. The sample means are still skewed right, but less so than single books: a long book is often paired with an ordinary one, and the average pulls it back.



Then the computer drew 500 random samples of 36 books. Now the histogram of $\bar{x}$ is close to a symmetric bell, centered at about 280 pages with a spread close to $\dfrac{150}{\sqrt{36}} = 25$ pages. The skewness of the population has nearly vanished.



This is the **central limit theorem**: when the sample is large, the sampling distribution of $\bar{x}$ is approximately normal, with mean μ and standard deviation $\frac{\sigma}{\sqrt{n}}$, **whatever the shape of the population**. Why does it happen? A sample mean is built from many independent pieces. To get an extreme mean, many pieces would have to be extreme in the same direction at once, which is rare; most samples mix high and low values, which pile up near μ and taper off evenly on both sides. A long right tail in the population shows up as an occasional large value, but in a sample of 36 one large value is diluted by 35 others.

How large is large? It depends on the population. If the population is normal, any n works (Lesson 9.3). If it is roughly symmetric, even $n = 10$ is usually fine. For skewed populations, even strongly skewed ones, our rule of thumb is $n \geq 30$. For extreme skewness or extreme outliers, a larger sample may be needed. Remember also what the CLT does **not** say: it does not make the population normal, and it does not make the data in your sample look normal. A large sample of book lengths still has a skewed histogram. Only the distribution of $\bar{x}$ over many samples becomes normal, so you still cannot use the normal curve for one single book.

The CLT also explains Lesson 9.2. Record each seed as 1 if it sprouts and 0 if not; then $\hat{p}$ is just the mean of those 0s and 1s. The Large Counts condition is the version of "large enough n" for this population of 0s and 1s, which is very skewed when p is near 0 or 1, so those cases need a larger n. In Units 10 to 12 the CLT is what lets us use normal and t methods for means even when we do not know the shape of the population.

KEY FORMULAS

$$n \geq 30 \implies \bar{x} \text{ is approximately } N\left(\mu, \frac{\sigma}{\sqrt{n}}\right) \text{ for any population shape}$$

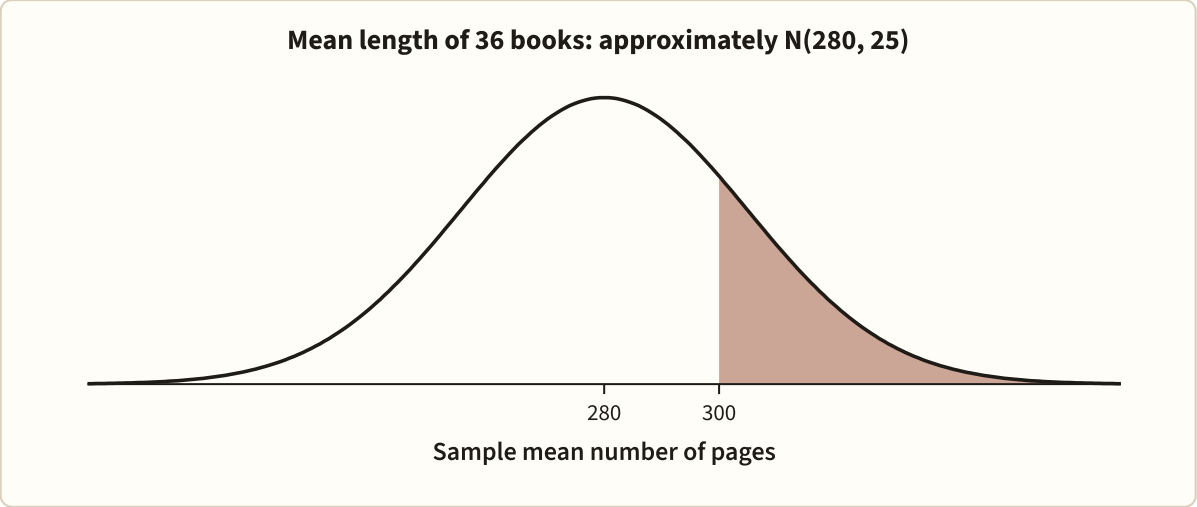
$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$$

$$\text{total} = n\bar{x}, \text{ so } P(\text{total} > T) = P\left(\bar{x} > \frac{T}{n}\right)$$

From history. Pierre-Simon Laplace, the French mathematician and astronomer, proved an early form of the central limit theorem in 1810. It extended Abraham de Moivre's discovery (1733) that binomial counts follow a bell-shaped curve to averages of many kinds of measurements, and it helped explain why the normal curve fits errors in astronomical observations so well.

Practice habit. Before any normal calculation for $\bar{x}$, say which reason applies: "the population is normal" or "$n \geq 30$, so the CLT applies." If neither is true, stop.

Example 1. The 4,000 books in a parish library have lengths with mean 280 pages and standard deviation 150 pages, and the distribution is strongly skewed right. The librarian chooses a random sample of 36 books. (a) Describe the shape, mean and standard deviation of the sampling distribution of $\bar{x}$. (b) Find the probability that the mean length of the 36 books is more than 300 pages. Round to 4 decimal places.



1. 10% condition: $36 \le 0.10(4{,}000) = 400$, so we may use $\dfrac{\sigma}{\sqrt{n}}$.
2. Shape: the population is skewed, but $n = 36 \ge 30$, so by the CLT $\bar{x}$ is approximately normal.
3. Mean $\mu_{\bar{x}} = 280$ pages; standard deviation $\sigma_{\bar{x}} = \dfrac{150}{\sqrt{36}} = \dfrac{150}{6} = 25$ pages.
4. Standardize: $z = \dfrac{300 - 280}{25} = 0.80$.
5. Area to the right: $P(\bar{x} > 300) = 1 - 0.7881 = 0.2119$.
6. Conclusion: about 21% of random samples of 36 books would have a mean length over 300 pages.

Answer: (a) Approximately normal (CLT, $n = 36$), mean 280 pages, standard deviation 25 pages. (b) $z = 0.80$, $P(\bar{x} > 300) = 0.2119$.

Example 2. At a farm stand, the amounts customers spend are skewed right with mean \$18 and standard deviation \$16. On a busy Saturday, 64 customers come; treat them as a random sample of all the stand's customers. (a) Find the probability that the Saturday's total sales are more than \$1,312. Round to 4 decimal places. (b) Can you find the probability that a single customer spends more than \$20.50 the same way? Explain.

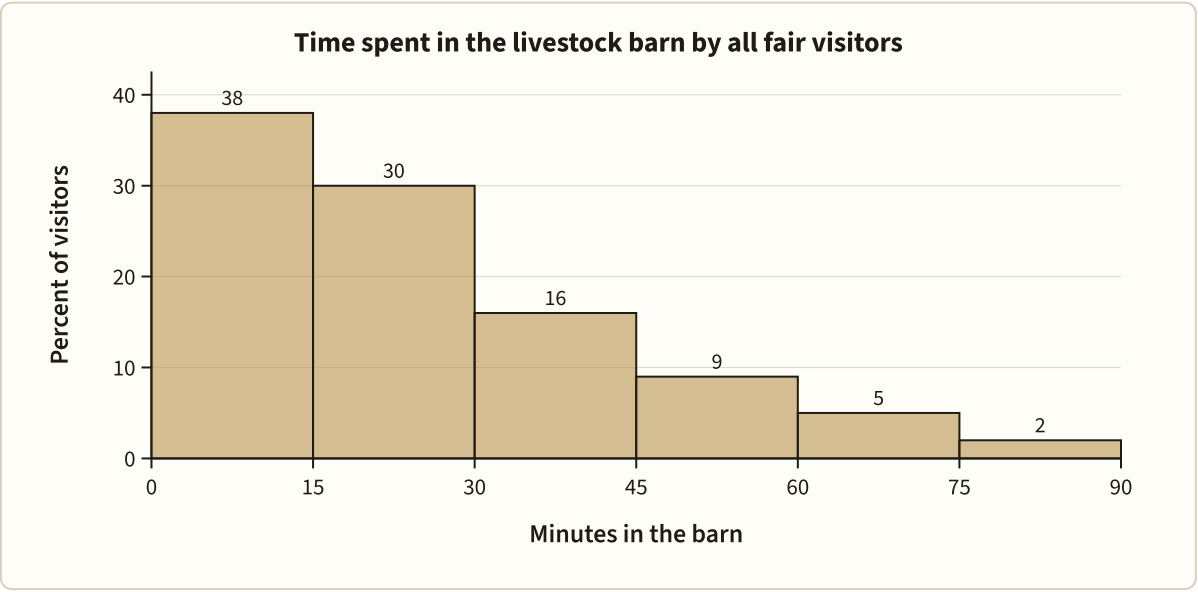
1. Turn the total into a mean: total $> \$1{,}312$ means $\bar{x} > \dfrac{1{,}312}{64} = \$20.50$.
2. Shape: the population is skewed right, but $n = 64 \geq 30$, so by the CLT $\bar{x}$ is approximately normal. The stand has many more than 640 customers over a season, so the 10% condition holds.
3. $\mu_{\bar{x}} = \$18$ and $\sigma_{\bar{x}} = \dfrac{16}{\sqrt{64}} = \dfrac{16}{8} = \$2$.
4. Standardize: $z = \dfrac{20.50 - 18}{2} = 1.25$.
5. $P(\bar{x} > 20.50) = 1 - 0.8944 = 0.1056$.
6. (b) No. One customer's spending follows the skewed population distribution, not a normal curve, and we do not know that distribution well enough to find areas. The CLT is about means of large samples, not single values.
7. Conclusion: on about 1 Saturday in 10 with 64 customers, sales would top \$1,312.

Answer: (a) $\sigma_{\bar{x}} = \$2$, $z = 1.25$, probability $= 0.1056$. (b) No: a single amount comes from a skewed distribution, and the CLT applies only to the mean of a large sample.

PRACTICE

1. EASY Can we use normal calculations for $\bar{x}$ in each case? Answer yes or no. (a) Board lengths are normally distributed; $n = 5$ boards. (b) Rainfall amounts from single storms are strongly skewed right; $n = 5$ storms. (c) The same rainfall amounts; $n = 60$ storms.

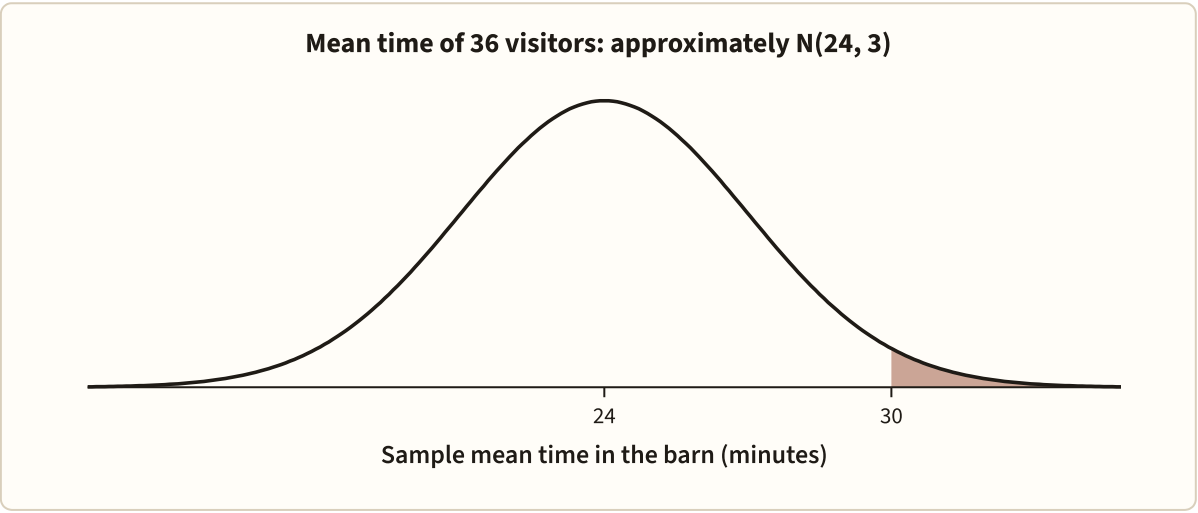
2. **EASY** The histogram shows the time thousands of visitors spent in the livestock barn at a county fair: mean 24 minutes, standard deviation 18 minutes, skewed right. Describe the sampling distribution of $\bar{x}$ for random samples of 36 visitors (shape, mean and standard deviation).



3. **EASY** Henry says, "The central limit theorem says that if I take a big enough sample, my data will be normally distributed." Explain what is wrong with his statement.

4. **EASY** The library books in the lesson have $\sigma = 150$ pages. How many books must be in a random sample for the standard deviation of $\bar{x}$ to be 15 pages?

5. **MEDIUM** For the fair visitors in Problem 2, find the probability that a random sample of 36 visitors spends a mean of more than 30 minutes in the barn. Round to 4 decimal places.



6. **MEDIUM** The 4,000 library books have mean length 280 pages and standard deviation 150 pages (skewed right). Find the probability that a random sample of 100 books has a mean length between 265 and 295 pages. Round to 4 decimal places.

7. MEDIUM At a weather station, the rainfall on rainy days is strongly skewed right, with mean 0.30 inch and standard deviation 0.40 inch. A student picks a random sample of 64 rainy days from many years of records. Find the probability that the sample mean rainfall is more than 0.36 inch. Round to 4 decimal places.

8. MEDIUM In Problem 7, explain why we cannot use the normal curve to find the probability that a single rainy day has more than 0.36 inch of rain.

9. MEDIUM Peach crates from an orchard have weights with mean 45 pounds and standard deviation 6 pounds; the distribution is not normal. A truck carries a random sample of 36 crates and is rated for 1,656 pounds of load. Find the probability that the load is over 1,656 pounds. Round to 4 decimal places.

10. HARD For random samples of 36 fair visitors (Problem 2), find the interval centered at 24 minutes that contains the middle 95% of sample means. Round to 2 decimal places.

11. HARD In a parish of 1,500 families, the number of children per family is skewed right with mean 2.1 and standard deviation 1.2. A random sample of 100 families is chosen for a survey about the summer picnic. (a) Find the probability that the sample mean is 2.4 children or more. Round to 4 decimal places. (b) Would a sample mean of 2.4 be surprising?

12. HARD On ordinary nights, the number of meteors Samuel counts per hour is skewed right with mean 6 and standard deviation 3. (a) Is a sample of 9 nights large enough to treat $\bar{x}$ as normal? (b) For a random sample of 36 nights, find the probability that the mean count is less than 5.25 meteors per hour. Round to 4 decimal places.

13. CHALLENGE In a large seed lot, 80% of seeds sprout. Record each seed as 1 if it sprouts and 0 if it does not, so the population is made of 1s and 0s. (a) Find the population mean μ. (b) Find the population standard deviation σ, using the formula for the standard deviation of a random variable from Lesson 7.2. (c) Find $\frac{\sigma}{\sqrt{n}}$ for $n = 100$ and compare it with $\sqrt{\frac{p(1-p)}{n}}$. (d) Explain how this connects the CLT to the Large Counts condition.

COMMON MISTAKES

- Thinking the CLT makes the population or the sample data normal. It is only about the distribution of $\bar{x}$ over many samples.
- Using the normal curve for a single value from a skewed population. The CLT does not apply to one individual; you need the population's own distribution.
- Applying the CLT to a small sample from a skewed population. Check: normal population (any n) or $n \geq 30$; otherwise do not use normal calculations.
- Forgetting to convert a total into a mean. Divide the total by n first (or use mean $n\mu$ and standard deviation $\sigma\sqrt{n}$ for the total).

ANSWERS TO PRACTICE

1. (a) Yes, a normal population gives a normal $\bar{x}$ for any n. (b) No, the population is skewed and the sample is small. (c) Yes, $n = 60 \geq 30$, so the CLT applies.
2. Approximately normal by the CLT ($n = 36 \geq 30$), with mean 24 minutes and standard deviation $\dfrac{18}{\sqrt{36}} = 3$ minutes.
3. The CLT is about the sampling distribution of $\bar{x}$, the pattern of sample means over many samples. A large sample's data look like the population; if the population is skewed, a large sample's histogram is skewed too.
4. $\dfrac{150}{\sqrt{n}} = 15$ gives $\sqrt{n} = 10$, so $n = 100$ books.
5. $\sigma_{\bar{x}} = 3$; $z = \dfrac{30 - 24}{3} = 2.00$; $P(\bar{x} > 30) = 1 - 0.9772 = 0.0228$.
6. $100 \leq 400$; $n \geq 30$, so the CLT applies. $\sigma_{\bar{x}} = \dfrac{150}{10} = 15$; $z = \pm 1.00$; probability $= 0.6827$ (table: 0.6826).
7. CLT: $n = 64 \geq 30$. $\sigma_{\bar{x}} = \dfrac{0.40}{8} = 0.05$; $z = \dfrac{0.36 - 0.30}{0.05} = 1.20$; $P(\bar{x} > 0.36) = 1 - 0.8849 = 0.1151$.
8. Single days follow the population distribution, which is strongly skewed right, not normal. The CLT makes the mean of 64 days approximately normal, but it says nothing about one day.
9. Total $> 1{,}656$ means $\bar{x} > \dfrac{1{,}656}{36} = 46$. CLT with $n = 36$: $\sigma_{\bar{x}} = \dfrac{6}{6} = 1$; $z = 1.00$; probability $= 1 - 0.8413 = 0.1587$.
10. $24 \pm 1.96(3) = 24 \pm 5.88$, so from 18.12 to 29.88 minutes. About 95% of samples of 36 visitors have a mean time in this range.
11. $100 \leq 150$; CLT since $n \geq 30$. $\sigma_{\bar{x}} = \dfrac{1.2}{10} = 0.12$; $z = \dfrac{2.4 - 2.1}{0.12} = 2.50$. (a) $P(\bar{x} \geq 2.4) = 0.0062$. (b) Yes; only about 6 in 1,000 random samples would average that many children.
12. (a) No: the population is skewed and $n = 9$ is small. (b) CLT with $n = 36$: $\sigma_{\bar{x}} = \dfrac{3}{6} = 0.5$; $z = \dfrac{5.25 - 6}{0.5} = -1.50$; probability $= 0.0668$.
13. (a) $\mu = 1(0.8) + 0(0.2) = 0.8$. (b) $\sigma^2 = (1 - 0.8)^2(0.8) + (0 - 0.8)^2(0.2) = 0.032 + 0.128 = 0.16$, so $\sigma = 0.4$. (c) $\dfrac{0.4}{10} = 0.04$, and $\sqrt{\dfrac{0.8(0.2)}{100}} = 0.04$: the same. (d) $\hat{p}$ is the sample mean of the 0s and 1s, so the CLT says it becomes normal for large n. The Large Counts condition is the rule for how large n must be for a population of 0s and 1s, which is skewed when p is far from 0.5.

Unit 9 test

You have 60 minutes. You may use a scientific or graphing calculator and the standard normal table at the back of the course. Show your work: state the mean and standard deviation you use, check the conditions (10%, Large Counts, or normal population / central limit theorem) before any normal calculation, and draw and shade a sketch. Round probabilities to 4 decimal places unless a problem says otherwise; table answers may differ from technology in the last digit.

1. **6 PTS** A dairy cooperative has 1,800 member farms. Its records show that the mean herd size of all 1,800 farms is 85 cows. An inspector tests milk from a random sample of 50 member farms and finds that 6% of the sampled farms have a high bacteria count. (a) Is 85 cows a parameter or a statistic? Give its symbol. (b) Is 6% a parameter or a statistic? Give its symbol.

2. **7 PTS** The table shows the bushels of beans from the four rows of a family garden; treat the four rows as the population. (a) Find μ. (b) Find the mean of the sample means of all 6 samples of 2 rows. (c) How many of the 6 samples have a sample mean exactly equal to μ?

   Bean harvest from four garden rows

Row	Bushels of beans
A	1
B	3
C	5
D	7

3. **7 PTS** A forester wants to estimate the mean trunk diameter of the trees in a 500-acre forest. Method A measures 20 randomly chosen trees, Method B measures 80 randomly chosen trees, and Method C measures the 200 trees nearest the road. Which methods are unbiased, and which method is best? Explain.

4. **7 PTS** In a large cornfield, 36% of the plants bear two ears. A farmer checks a random sample of 100 plants. Find the mean and standard deviation of the sampling distribution of $\hat{p}$, the sample proportion of plants with two ears.

5. **7 PTS** A supplier's pumpkin seeds sprout 80% of the time. A 4-H club plants a random sample of 400 seeds from a large lot. Find the probability that fewer than 77% of the seeds sprout.

6. **6 PTS** At a large sugarhouse, 4% of the jars of maple syrup are slightly underfilled. An inspector checks a random sample of 150 jars. Is the sampling distribution of $\hat{p}$ approximately normal? Explain.

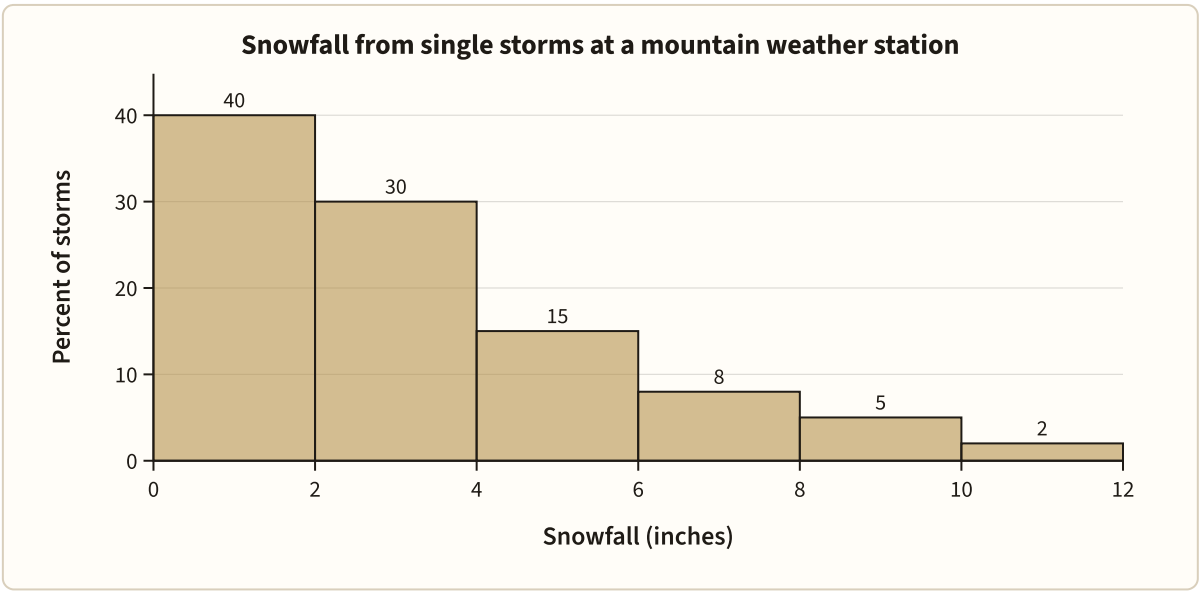
7. **6 PTS** When $p = 0.5$ and $n = 100$, the standard deviation of $\hat{p}$ is 0.05. (a) What sample size makes it 0.025? (b) What is it when $n = 25$?

8. **7 PTS** Sacks of feed labeled 50 pounds have weights that are normally distributed with mean 50.5 pounds and standard deviation 0.8 pound. A store manager weighs a random sample of 16 sacks. Find the probability that their mean weight is less than 50.3 pounds.

9. **8 PTS** Watermelons from a farm have weights that are normally distributed with mean 20 pounds and standard deviation 3 pounds. (a) Find the probability that one watermelon weighs more than 21.5 pounds. (b) Find the probability that the mean weight of a random sample of 16 watermelons is more than 21.5 pounds.

10. **7 PTS** Single nails from a hardware store bin vary in weight with standard deviation 0.20 gram, so the mean weight of a random sample of 25 nails has standard deviation 0.04 gram. Interpret 0.04 gram in context, and explain why it is smaller than 0.20 gram.

11. **7 PTS** The histogram shows that snowfall from single storms at a mountain weather station is skewed right; the mean is 3.2 inches and the standard deviation is 2.4 inches. A student chooses a random sample of 36 storms from many years of records. Find the probability that the sample mean snowfall is more than 3.7 inches.



12. **7 PTS** Can we use normal calculations for $\bar{x}$ in each case? Answer yes or no. (a) Daily visitor counts at a church library are skewed right; $n = 8$ days. (b) Heights of corn plants are normally distributed; $n = 8$ plants. (c) Cords of firewood burned per winter by households are skewed right; $n = 50$ households.

13. **8 PTS** The 3,000 trees in a pecan orchard have yields that are skewed, with mean 60 pounds and standard deviation 20 pounds per tree. The owner harvests a random sample of 100 trees first. Find the probability that the total harvest from those 100 trees is more than 6,350 pounds.

14. **10 PTS** About 10% of the trees in a region's forests show storm damage. A survey team wants the standard deviation of $\hat{p}$ to be at most 0.01. (a) How many randomly chosen trees must the team examine, using $p = 0.1$? (b) A team member says a region with ten times as many trees would need a sample ten times as large. Is this right? Explain.

UNIT 9 TEST: ANSWER KEY

1. (a) A parameter, $\mu = 85$ cows: it describes all 1,800 farms. (b) A statistic, $\hat{p} = 0.06$: it comes from the sample of 50 farms.
2. (a) $\mu = \dfrac{16}{4} = 4$ bushels. (b) Sample means 2, 3, 4, 4, 5, 6; their mean is $\dfrac{24}{6} = 4$ bushels. (c) 2 samples (rows A and D, rows B and C).
3. A and B are unbiased because random sampling lets chance pick the trees. C is likely biased: trees by the road get more light and may grow differently, and its large size does not fix that. B is best: unbiased, and with 80 trees its estimate has less variability than A's.
4. Mean $= 0.36$; $\sigma_{\hat{p}} = \sqrt{\dfrac{0.36(0.64)}{100}} = \sqrt{0.002304} = 0.048$ (the field has thousands of plants, so the 10% condition holds).
5. $\sigma_{\hat{p}} = \sqrt{\dfrac{0.80(0.20)}{400}} = 0.02$; Large Counts: 320 and 80. $z = \dfrac{0.77 - 0.80}{0.02} = -1.50$; $P(\hat{p} < 0.77) = 0.0668$.
6. No. $np = 150(0.04) = 6 < 10$, so the Large Counts condition fails and the distribution of $\hat{p}$ is skewed right.
7. (a) Half the spread needs four times the sample: $n = 400$. (b) One quarter of the sample doubles the spread: $\sqrt{\dfrac{0.25}{25}} = 0.1$.
8. Population normal, so $\bar{x}$ is normal. $\sigma_{\bar{x}} = \dfrac{0.8}{4} = 0.2$; $z = \dfrac{50.3 - 50.5}{0.2} = -1.00$; probability $= 0.1587$.
9. (a) $z = \dfrac{21.5 - 20}{3} = 0.50$; probability $= 1 - 0.6915 = 0.3085$. (b) $\sigma_{\bar{x}} = \dfrac{3}{4} = 0.75$; $z = \dfrac{1.5}{0.75} = 2.00$; probability $= 1 - 0.9772 = 0.0228$.
10. If we took many random samples of 25 nails, the sample mean weight would typically differ from the true mean weight of all the nails by about 0.04 gram. It is smaller because heavy and light nails in the same sample partly cancel when averaged; the spread of the mean is $\dfrac{0.20}{\sqrt{25}}$.
11. Skewed population, but $n = 36 \geq 30$, so by the CLT $\bar{x}$ is approximately normal. $\sigma_{\bar{x}} = \dfrac{2.4}{6} = 0.4$; $z = \dfrac{3.7 - 3.2}{0.4} = 1.25$; probability $= 1 - 0.8944 = 0.1056$.
12. (a) No: skewed population and small sample. (b) Yes: a normal population gives a normal $\bar{x}$ for any n. (c) Yes: $n = 50 \geq 30$, so the CLT applies.
13. 10%: $100 \leq 300$; CLT: $n = 100 \geq 30$. Total $> 6{,}350$ means $\bar{x} > 63.5$. $\sigma_{\bar{x}} = \dfrac{20}{10} = 2$; $z = \dfrac{63.5 - 60}{2} = 1.75$; probability $= 1 - 0.9599 = 0.0401$.

14. (a) $\sqrt{\frac{0.1(0.9)}{n}} \le 0.01$ gives $\frac{0.09}{n} \le 0.0001$, so $n \ge 900$ trees. (b) No. As long as the population is at least 10 times the sample, the standard deviation $\sqrt{\frac{p(1-p)}{n}}$ depends on the sample size, not the population size, so 900 trees would serve either region.

UNIT 10

Confidence Intervals

In Unit 9 you learned how a statistic such as $\hat{p}$ or $\bar{x}$ varies from sample to sample when we know the population. This unit turns that knowledge around. In real life we have one sample and do not know the population, and we want to say something honest about it: not just "the germination rate is 86%" but "we are 95% confident the germination rate of the whole lot is between 81% and 91%." That is a confidence interval: an estimate together with a margin of error that tells how far off it could reasonably be.

You will learn what "95% confident" really means (a statement about how often the method succeeds), build intervals for a population proportion with z and for a population mean with the t distributions, check the conditions that make each method trustworthy, and plan how large a sample to take. From now on every inference problem follows four steps: State, Plan, Do, Conclude, ending with a careful sentence in context.

Confidence intervals are the first of the two main tools of statistical inference. Unit 11 introduces the second, significance tests, and Units 12 and 13 extend both to comparing groups and to categorical data. The reasoning you learn here (estimate ± critical value × standard error) appears in every one of those lessons.

PREREQUISITES

- Compute and interpret the sample standard deviation s (Lesson 2.4).
- Find areas and critical values under the normal curve with the table or technology (Lessons 3.4 and 3.5).
- Recognize random samples and sources of bias such as convenience, voluntary response and nonresponse (Lessons 5.1–5.3).
- Use the sampling distributions of $\hat{p}$ and $\bar{x}$: their means, standard deviations and the conditions for normality (Lessons 9.2–9.4).
- Make and read dotplots and stemplots and spot skewness and outliers (Lessons 1.4 and 1.5).

MASTERY CHECKLIST

☐ I can find the point estimate and margin of error from a confidence interval, and build an interval from them.
☐ I can explain what a confidence level means and why "95% probability that μ is in this interval" is the wrong reading.
☐ I can explain why a higher confidence level gives a wider interval and a larger sample gives a narrower one.
☐ I can recall $z^* = 1.645$, 1.960 and 2.576 for 90%, 95% and 99% confidence.
☐ I can check the Random, 10% and Large Counts conditions and construct a one-sample z interval for a proportion.
☐ I can explain why s in place of σ leads to the t distributions and read t^* from the t table with $df = n - 1$.

☐ I can check the Normal/Large Sample condition with a graph and construct a one-sample t interval for a mean.
☐ I can carry out the four-step process (State, Plan, Do, Conclude) and write the conclusion in context.
☐ I can use an interval to judge whether a claimed value of a parameter is plausible.
☐ I can explain what the margin of error does and does not cover, and compute the sample size needed for a given margin of error, rounding up.

LESSON 10.1 · 2 SESSIONS OF 45 MINUTES

The Idea of a Confidence Interval

OBJECTIVES

- Explain why an interval of the form estimate ± margin of error is more honest and more useful than a single point estimate.
- Compute a confidence interval for a population mean when σ is known, using $z^* = 1.645$, 1.960 or 2.576.
- Interpret a confidence interval and a confidence level correctly in context.
- Explain how the confidence level and the sample size affect the width of a confidence interval.

VIDEO

WATCH

Intro to Confidence Intervals for One Mean (Sigma Known)

jbstatistics · 10:36

When to watch: Before Example 2: a confidence interval for a mean when σ is known.

https://www.youtube.com/watch?v=KG921rfbTDw

BACKUP

Confidence Intervals for One Mean: Interpreting the Interval

jbstatistics · 7:13

When to watch: Before Example 1: what "95% confident" does and does not mean.

https://www.youtube.com/watch?v=MkEFOsPe37A

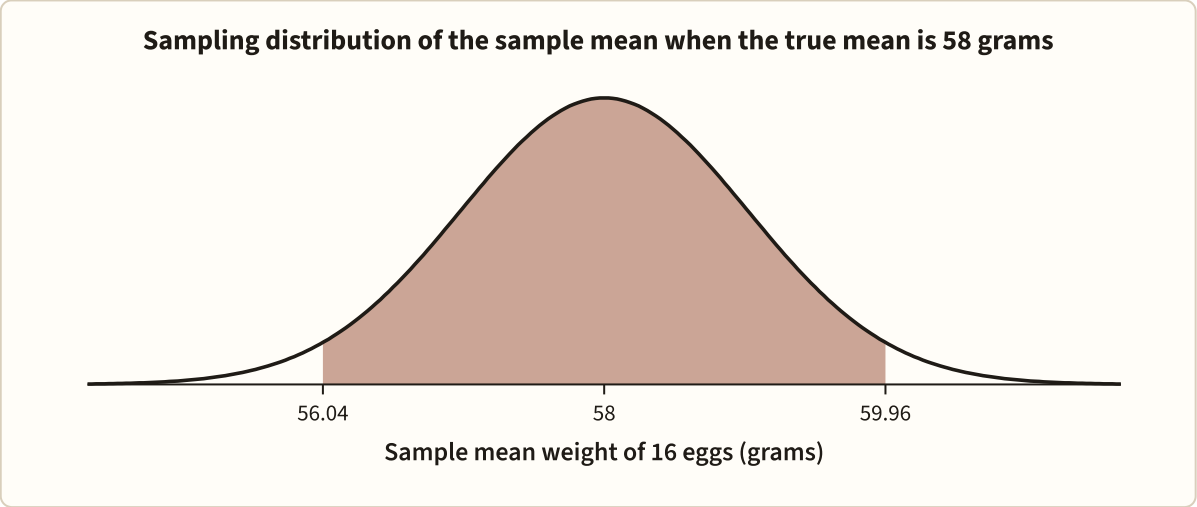
KEY TERMS

point estimate	A single statistic used as our best guess for a parameter, such as $\bar{x}$ for μ or $\hat{p}$ for p.
margin of error	How far the estimate could reasonably be from the parameter because of random sampling variation alone; it is half the width of the interval.
confidence interval	An interval, estimate ± margin of error, computed from sample data, that gives a range of plausible values for a parameter.
confidence level	The percent of all possible samples for which the method gives an interval that captures the true parameter (for example 95%).
critical value z^*	The number of standard deviations needed so that the middle C% of the standard normal curve lies between $-z^*$ and z^*.

CORE IDEAS

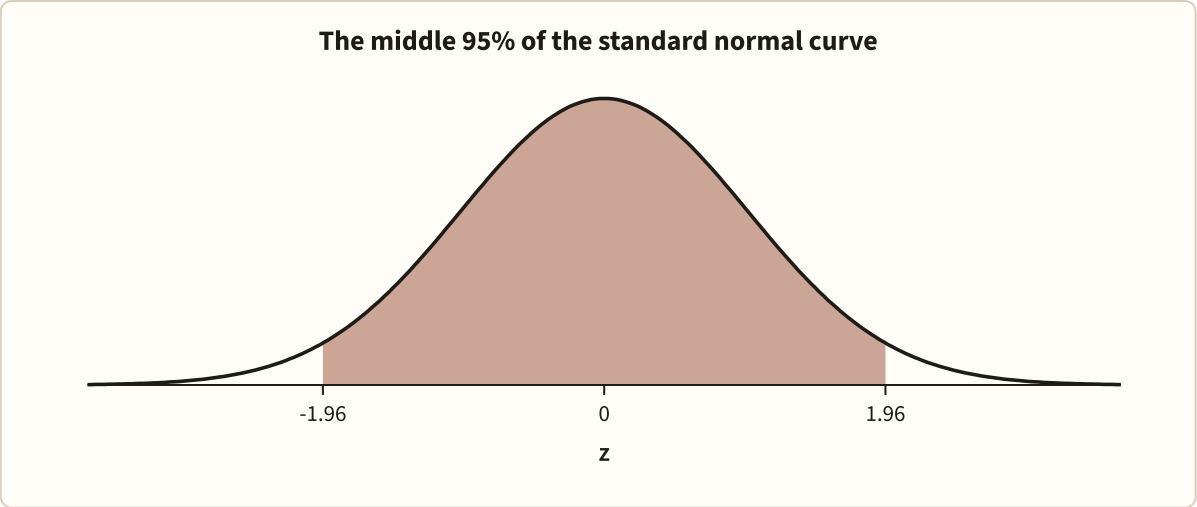
In Unit 9 you saw that a statistic changes from sample to sample. Suppose Samuel weighs a random sample of 16 eggs from his family's large flock and finds $\bar{x} = 57.3$ grams. That number is a **point estimate** of μ, the mean weight of all the eggs the flock lays. But another sample would give a different $\bar{x}$, and almost no sample hits μ exactly. An honest report says how far off the estimate might be: *estimate ± margin of error*. Such a range is called a **confidence interval**.

Here is why the method works. Suppose the egg weights have standard deviation $\sigma = 4$ grams. From Lesson 9.3, the sample mean of $n = 16$ eggs has mean μ and standard deviation $\dfrac{\sigma}{\sqrt{n}} = \dfrac{4}{\sqrt{16}} = 1$ gram, and, if the egg weights themselves are roughly normal (as egg weights usually are), $\bar{x}$ is approximately normal as well. The middle 95% of any normal distribution lies within 1.96 standard deviations of its mean. So in 95% of all samples, $\bar{x}$ lands within $1.96 \times 1 = 1.96$ grams of μ. The picture shows this for a flock whose true mean happens to be 58 grams.



Now turn that sentence around. If $\bar{x}$ is within 1.96 grams of μ, then μ is within 1.96 grams of $\bar{x}$. So the interval $\bar{x} \pm 1.96$ captures μ for exactly those samples whose $\bar{x}$ landed in the shaded middle, and that is 95% of all samples. Notice that we never need to know μ to build the interval; we only need to know how much $\bar{x}$ varies. Every confidence interval in this course has the same shape: statistic ± (critical value) × (standard deviation of the statistic).

The **critical value** sets the confidence level. For confidence level C, z^* is the number that puts the middle C of the standard normal curve between $-z^*$ and z^*. For 95%, the two tails hold 2.5% each, so z^* is the z with area 0.975 to its left: $z^* = 1.960$. In the same way, $z^* = 1.645$ for 90% (area 0.95 to its left) and $z^* = 2.576$ for 99% (area 0.995 to its left).



What does "95% confident" mean? It is a statement about the *method*: if we took sample after sample and built an interval from each one, about 95% of those intervals would capture μ. The table shows ten simulated samples of 16 eggs from a flock with $\mu = 58$ grams. Nine of the ten intervals capture 58; sample 6 misses. In real life you see only one interval, and you cannot tell whether it is one of the hits or one of the misses. So do NOT say "there is a 95% probability that μ is between 55.34 and 59.26." The mean μ is a fixed number; it either is or is not in that interval. Also, the interval is about the mean μ, not about individual eggs.

Ten simulated 95% intervals (true mean 58 grams)

Sample	$\bar{x}$ (grams)	Interval $\bar{x} \pm 1.96$ (grams)	Captures 58?
1	57.3	55.34 to 59.26	yes
2	58.9	56.94 to 60.86	yes
3	56.4	54.44 to 58.36	yes
4	59.1	57.14 to 61.06	yes
5	58.2	56.24 to 60.16	yes
6	60.3	58.34 to 62.26	no
7	57.8	55.84 to 59.76	yes
8	58.6	56.64 to 60.56	yes
9	56.9	54.94 to 58.86	yes
10	57.5	55.54 to 59.46	yes

Two things control the width. A higher confidence level needs a larger z^*, so the interval gets wider: to be more sure of catching μ, you must cast a wider net. A larger sample makes $\frac{\sigma}{\sqrt{n}}$ smaller, so the interval gets narrower. Because of the square root, you need four times as many observations to cut

the margin of error in half. Finally, every interval in this unit rests on conditions you will check each time: the data come from a random sample, the sample is at most 10% of the population when sampling without replacement, and the sampling distribution is approximately normal. In this lesson we also pretend that σ is known, which is rare; Lesson 10.3 shows what to do when it is not.

KEY FORMULAS

$$\text{estimate} \pm \text{margin of error}$$

$$\text{statistic} \pm (\text{critical value}) \times (\text{standard deviation of the statistic})$$

$$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}} \quad (\sigma \text{ known})$$

$$z^* = 1.645 \ (90\%), \qquad z^* = 1.960 \ (95\%), \qquad z^* = 2.576 \ (99\%)$$

From history. The statistician Jerzy Neyman set out the theory of confidence intervals in a paper published in 1937. His key idea is the one in this lesson: the confidence level describes how often the method succeeds over many possible samples, not the chance that one particular interval is right.

Practice habit. Know the three critical values by heart: $z^* = 1.645$ for 90%, 1.960 for 95% and 2.576 for 99%. Whenever you report an interval, name the parameter in context and give the units.

WORKED EXAMPLES

Example 1. A beekeeper takes a random sample of hives from a large apiary and reports a 95% confidence interval for the mean honey harvest per hive this season: 38.2 to 45.8 pounds. Find the point estimate and the margin of error, and interpret the interval.

1. The point estimate is the center of the interval: $\bar{x} = \dfrac{38.2 + 45.8}{2} = \dfrac{84.0}{2} = 42.0$ pounds.
2. The margin of error is half the width: $ME = \dfrac{45.8 - 38.2}{2} = \dfrac{7.6}{2} = 3.8$ pounds.
3. Check: $42.0 - 3.8 = 38.2$ and $42.0 + 3.8 = 45.8$, which matches the reported interval.
4. Interpret: We are 95% confident that the mean honey harvest per hive for all hives in this apiary this season is between 38.2 and 45.8 pounds.

Answer: Point estimate $\bar{x} = 42.0$ pounds; margin of error 3.8 pounds. We are 95% confident that the mean harvest per hive for all hives in the apiary is between 38.2 and 45.8 pounds.

Example 2. A sawmill's cutting machine has run for years with a standard deviation of $\sigma = 0.12$ inch in board length. After a blade change, the foreman measures a simple random sample of 36 boards from the day's output of several thousand and finds $\bar{x} = 96.03$ inches. (a) Construct a 95% confidence interval for the mean length μ of all boards cut after the blade change. (b) Construct a 99% confidence interval. Round endpoints to 3 decimal places. (c) The machine is supposed to average 96 inches. Do the intervals give reason to doubt that?

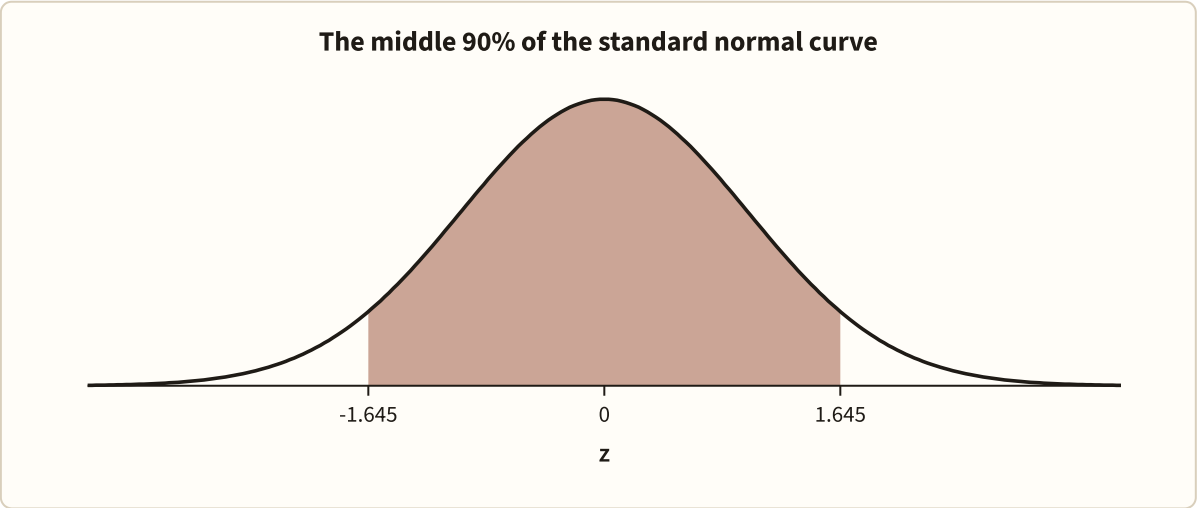
1. Conditions: the boards are an SRS; 36 boards is less than 10% of several thousand; $n = 36 \geq 30$, so by the central limit theorem $\bar{x}$ is approximately normal.
2. Standard deviation of $\bar{x}$: $\dfrac{\sigma}{\sqrt{n}} = \dfrac{0.12}{\sqrt{36}} = \dfrac{0.12}{6} = 0.02$ inch.
3. (a) 95%: $ME = 1.960 \times 0.02 = 0.0392$. Interval: 96.03 ± 0.0392, from 95.9908 to 96.0692, or 95.991 to 96.069 inches.
4. (b) 99%: $ME = 2.576 \times 0.02 = 0.05152$. Interval: 96.03 ± 0.05152, from 95.97848 to 96.08152, or 95.978 to 96.082 inches.
5. The 99% interval is wider: to be more confident of capturing μ we need a larger critical value.
6. (c) Both intervals contain 96, so 96 inches is a plausible value of μ. These data give no reason to doubt that the machine still averages 96 inches.
7. Conclusion: We are 95% confident that the mean length of all boards cut after the blade change is between 95.991 and 96.069 inches.

Answer: (a) 95.991 to 96.069 inches. (b) 95.978 to 96.082 inches. (c) No; 96 lies inside both intervals, so it is a plausible value for the mean length.

PRACTICE

1. EASY A county extension agent reports a 95% confidence interval for the mean wheat yield of farms in the county: 52.4 to 57.6 bushels per acre. Find the point estimate and the margin of error.
2. EASY From a random sample of beef calves on a large ranch, the point estimate of the mean weaning weight is 512 pounds with a margin of error of 14 pounds. Write the confidence interval.

3. **EASY** Use the picture of the standard normal curve. What critical value z^* puts the middle 90% of the area between $-z^*$ and z^*? How much area is in each tail?



4. **EASY** Using a random sample of July days from many years of records, a weather station computes a 95% confidence interval for the mean daily high temperature in July: 86.2°F to 89.0°F. Interpret this interval in context.

5. **MEDIUM** Over many seasons, the distances of Henry's golf drives have been approximately normal with standard deviation $\sigma = 15$ yards. In a random sample of 25 of his drives this season, $\bar{x} = 236$ yards. (a) Find the margin of error for a 95% confidence interval for his mean drive distance μ this season. (b) Find the interval.

6. **MEDIUM** Ruth computed a 95% confidence interval for the mean weight of the apples in her family's orchard this year: 182 to 194 grams. She wrote, "There is a 95% probability that the true mean is between 182 and 194 grams." Explain what is wrong, and write a correct interpretation of the confidence level.

7. **MEDIUM** The weights of a variety of Roma tomato are approximately normal with $\sigma = 8$ grams. A random sample of 16 tomatoes from a large garden has $\bar{x} = 70$ grams. Construct (a) a 90% and (b) a 99% confidence interval for the mean weight of all the tomatoes. Round to 2 decimal places.

8. **MEDIUM** A 95% confidence interval for the mean height of sunflowers in a large field, based on a random sample of 25 plants with σ known, has a margin of error of 4.2 centimeters. If the farmer had measured 100 plants instead (same σ, same confidence level), what would the margin of error be?

9. **MEDIUM** Fifty students in a statistics course each take their own random sample of eggs from the same large hatchery and each builds a 95% confidence interval for the mean egg weight. How many of the 50 intervals would you expect to capture the true mean?

10. HARD A teacher has a computer take ten random samples from a population of corn plants whose true mean height is known to be 50 inches. Each sample gives a 90% confidence interval, shown in the table. (a) How many of the ten intervals capture the true mean of 50 inches? (b) Grace says, "That is not 90%, so the method must be wrong." Explain why she is mistaken.

Ten simulated 90% intervals for the mean height of corn plants

Sample	$\bar{x}$ (inches)	90% interval (inches)
1	48.2	45.2 to 51.2
2	51.7	48.7 to 54.7
3	53.4	50.4 to 56.4
4	49	46.0 to 52.0
5	46.5	43.5 to 49.5
6	50.8	47.8 to 53.8
7	52.1	49.1 to 55.1
8	47.9	44.9 to 50.9
9	50.3	47.3 to 53.3
10	54	51.0 to 57.0

11. HARD A 95% confidence interval for the mean height of sunflower plants in a large field, based on a random sample of 36 plants with σ known, is 61.04 to 64.96 inches. Find $\bar{x}$ and σ.

12. HARD From a random sample of Sundays, a 95% confidence interval for the mean time a church van takes to complete its Sunday route is 26.08 to 33.92 minutes (σ known). Using the same data, construct a 99% confidence interval. Round to 2 decimal places.

13. CHALLENGE A farmer builds three separate 95% confidence intervals from three independent random samples: one for the mean egg weight, one for the mean daily milk yield and one for the mean apple diameter. (a) What is the probability that all three intervals capture their parameters? (b) What is the probability that at least one of them misses? Round to 4 decimal places.

COMMON MISTAKES

- Saying "there is a 95% chance that μ is in this interval." Once the sample is taken, the interval either contains μ or it does not. Speak about the method: 95% of all such intervals capture μ.
- Thinking the interval describes individual values ("95% of the eggs weigh between ..."). A confidence interval for μ is about the population mean, not about single eggs.
- Confusing the margin of error with the width. The width is twice the margin of error; find the margin of error as half the width and the estimate as the midpoint.

- Believing a higher confidence level gives a narrower interval. More confidence needs a larger z^* and so a wider interval; only a larger sample narrows the interval without giving up confidence.

ANSWERS TO PRACTICE

1. Point estimate $= \dfrac{52.4 + 57.6}{2} = 55.0$ bushels per acre; margin of error $= \dfrac{57.6 - 52.4}{2} = 2.6$ bushels per acre.
2. 512 ± 14: from 498 to 526 pounds.
3. $z^* = 1.645$; each tail has area 0.05, so the area to the left of 1.645 is 0.95.
4. We are 95% confident that the mean daily high temperature for all July days in this station's records is between 86.2°F and 89.0°F.
5. (a) $\dfrac{\sigma}{\sqrt{n}} = \dfrac{15}{5} = 3$; $ME = 1.96 \times 3 = 5.88$ yards. (b) 236 ± 5.88: from 230.12 to 241.88 yards.
6. The true mean is a fixed number, so this particular interval either contains it or does not; there is no 95% chance about it. The 95% describes the method: if she took many random samples and built an interval from each, about 95% of those intervals would capture the true mean apple weight. Correct statement: We are 95% confident that the mean weight of all the apples in the orchard this year is between 182 and 194 grams.
7. $\dfrac{\sigma}{\sqrt{n}} = \dfrac{8}{4} = 2$. (a) $70 \pm 1.645 \times 2 = 70 \pm 3.29$: 66.71 to 73.29 grams. (b) $70 \pm 2.576 \times 2 = 70 \pm 5.152$: 64.85 to 75.15 grams. The 99% interval is wider.
8. The margin of error is proportional to $\dfrac{1}{\sqrt{n}}$. Going from 25 to 100 multiplies $\sqrt{n}$ by $\dfrac{10}{5} = 2$, so the margin of error is halved: $\dfrac{4.2}{2} = 2.1$ centimeters.
9. $0.95 \times 50 = 47.5$, so about 47 or 48 of the intervals. No one can tell which ones missed.
10. (a) 7 intervals capture 50; samples 3 (50.4 to 56.4), 5 (43.5 to 49.5) and 10 (51.0 to 57.0) miss. (b) The 90% describes the long-run success rate over very many samples. With only ten intervals, chance variation can easily give 7, 8, 9 or 10 hits, just as ten coin tosses need not give exactly five heads.
11. $\bar{x} = \dfrac{61.04 + 64.96}{2} = 63.00$ inches. $ME = 1.96 = 1.96 \times \dfrac{\sigma}{\sqrt{36}}$, so $\dfrac{\sigma}{6} = 1$ and $\sigma = 6$ inches.
12. $\bar{x} = 30$ and $ME = 3.92 = 1.96 \times \dfrac{\sigma}{\sqrt{n}}$, so $\dfrac{\sigma}{\sqrt{n}} = 2$. At 99%: $ME = 2.576 \times 2 = 5.152$; interval 30 ± 5.152: 24.85 to 35.15 minutes.
13. (a) The samples are independent, so $P(\text{all capture}) = 0.95^3 = 0.857375 \approx 0.8574$. (b) $P(\text{at least one misses}) = 1 - 0.857375 = 0.142625 \approx 0.1426$. Even with good methods, the more intervals you build, the more likely one of them misses.

LESSON 10.2 · 2 SESSIONS OF 45 MINUTES

Estimating a Population Proportion

OBJECTIVES

- Explain why the standard error $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$ is used in place of the unknown standard deviation of $\hat{p}$.
- Check the Random, 10% and Large Counts conditions for a one-sample z interval for a proportion.
- Construct a one-sample z interval for p using the four-step process (State, Plan, Do, Conclude).
- Interpret the interval in context and use it to judge whether a claimed value of p is plausible.

VIDEO

WATCH

An Introduction to Inference for a Proportion

jbstatistics · 10:26

When to watch: Before Example 1: the one-proportion z interval and its conditions.

https://www.youtube.com/watch?v=owYtDtmrCoE

BACKUP

1 Sample Proportion z Confidence Interval

ProfRobBob · 11:25

When to watch: Extra practice: a one-proportion z interval, worked through.

https://www.youtube.com/watch?v=NNCMPukJMFM

> **Note:** One example in the main video is a presidential approval poll; no party or person is named.

KEY TERMS

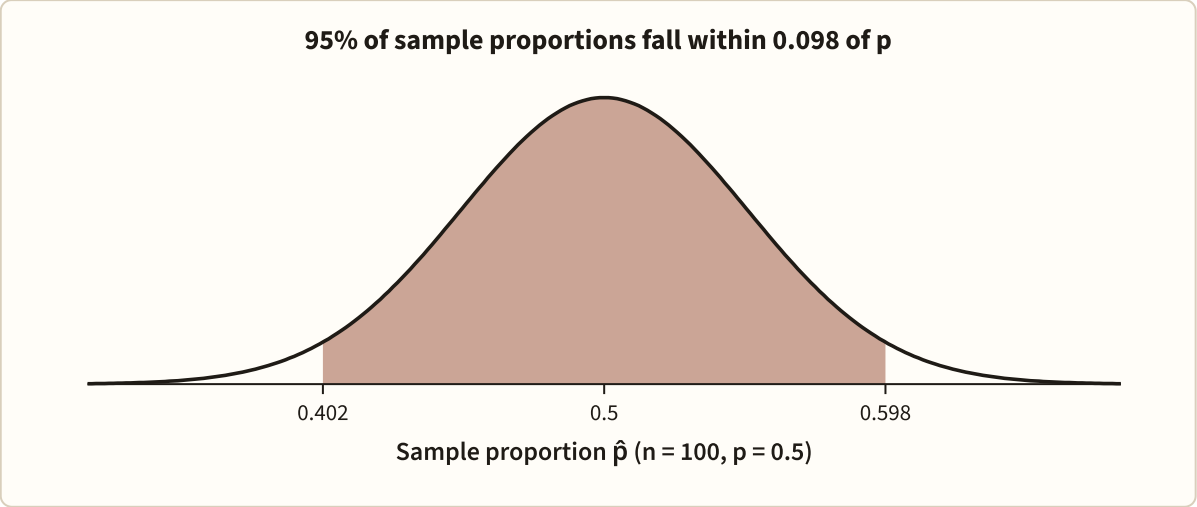
sample proportion $\hat{p}$	The fraction of the sample with the trait of interest: $\hat{p} = \frac{X}{n}$, where X is the number of successes.
standard error	An estimate of the standard deviation of a statistic, computed from the sample. For $\hat{p}$ it is $\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$.
10% condition	When sampling without replacement, the sample size n is at most 10% of the population size, so the draws are nearly independent.
Large Counts condition	The sample has at least 10 successes and at least 10 failures ($n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$), so the sampling distribution of $\hat{p}$ is approximately normal.

four-step process	State the parameter and confidence level; Plan by naming the procedure and checking conditions; Do the calculations; Conclude in context.

CORE IDEAS

Many practical questions ask *what fraction*: what fraction of the seeds in a lot will germinate, what fraction of the eggs in an incubator will hatch, what fraction of an apple harvest has blemishes. The parameter is the population proportion p. Our point estimate is the sample proportion $\hat{p} = \dfrac{X}{n}$, the number of successes divided by the sample size.

Lesson 9.2 told us how $\hat{p}$ behaves. Its sampling distribution has mean p and standard deviation $\sqrt{\dfrac{p(1-p)}{n}}$, and it is approximately normal when the counts are large. For example, if $p = 0.5$ and $n = 100$, the standard deviation is $\sqrt{\dfrac{0.5 \times 0.5}{100}} = 0.05$, so 95% of all samples give a $\hat{p}$ within $1.96 \times 0.05 = 0.098$ of 0.5.



There is a catch. The formula for the standard deviation contains p, the very number we are trying to estimate. The way out is to put $\hat{p}$ in place of p. The result, $\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}}$, is called the **standard error** of $\hat{p}$: it estimates how far $\hat{p}$ typically falls from p. With a reasonably large sample, $\hat{p}$ is close to p, so the standard error is close to the true standard deviation and the interval $\hat{p} \pm z^* \times SE$ captures p in very nearly C% of all samples. This is the **one-sample z interval for a proportion**.

The method works only when three conditions hold, and each has a reason. Notice that Large Counts uses the observed numbers of successes and failures, $n\hat{p}$ and $n(1-\hat{p})$, because we do not know p.

Conditions for a one-sample z interval for p

Condition	What to check	Why it matters
Random	The data come from a random sample or a randomized experiment	Without chance selection the sampling distribution does not apply and the estimate may be biased
10%	n is at most 10% of the population when sampling without replacement	Keeps the draws nearly independent, so the standard deviation formula is accurate
Large Counts	At least 10 successes and at least 10 failures in the sample	Makes the sampling distribution of p̂ approximately normal, so z* is the right critical value

From now on, write every inference problem in four steps. **State**: name the parameter in words and give the confidence level. **Plan**: name the procedure ("one-sample z interval for p") and check each condition with numbers. **Do**: find $\hat{p}$, the standard error, the margin of error and the interval. **Conclude**: interpret the interval in context: "We are 95% confident that the interval from ... to ... captures the true proportion of ..."

An interval also helps judge a claim. Values inside the interval are plausible values of p; values outside it are not plausible at that confidence level. If a seed company claims 90% germination and your whole interval lies below 0.90, the data give good reason to doubt the claim. An interval that contains the claimed value does not prove the claim; it only says the data are consistent with it.

KEY FORMULAS

$$\hat{p} = \frac{X}{n}$$

$$SE_{\hat{p}} = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}$$

Large Counts: $n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$

Practice habit. Check the conditions before calculating. Write Random, 10% and Large Counts as three separate lines, each with the numbers that show the condition is met.

WORKED EXAMPLES

Example 1. A seed company tests a random sample of 200 seeds from a large lot of sweet corn seed (tens of thousands of seeds), and 172 of them germinate. Construct a 95% confidence interval for p, the proportion of all seeds in the lot that germinate. Round to 3 decimal places.

1. $\hat{p} = \dfrac{172}{200} = 0.86$.
2. Conditions: Random sample, yes. 10%: 200 seeds is far less than 10% of tens of thousands. Large Counts: 172 successes and $200 - 172 = 28$ failures, both at least 10.
3. Standard error: $SE = \sqrt{\dfrac{0.86 \times 0.14}{200}} = \sqrt{\dfrac{0.1204}{200}} = \sqrt{0.000602} \approx 0.024536$.
4. Margin of error: $ME = 1.96 \times 0.024536 \approx 0.04809$.
5. Interval: 0.86 ± 0.04809, from 0.81191 to 0.90809, or 0.812 to 0.908.
6. Conclusion: We are 95% confident that between 81.2% and 90.8% of all the seeds in this lot will germinate.

Answer: 0.812 to 0.908. We are 95% confident that between 81.2% and 90.8% of all seeds in the lot germinate.

Example 2. An orchard's harvest this year is about 20,000 apples. The owner inspects a random sample of 150 apples and finds 33 with blemishes. (a) Use the four-step process to construct a 90% confidence interval for the proportion of all apples in the harvest with blemishes. Round to 3 decimal places. (b) A buyer pays full price only if at most 15% of the harvest is blemished. Is 0.15 a plausible value for the true proportion?

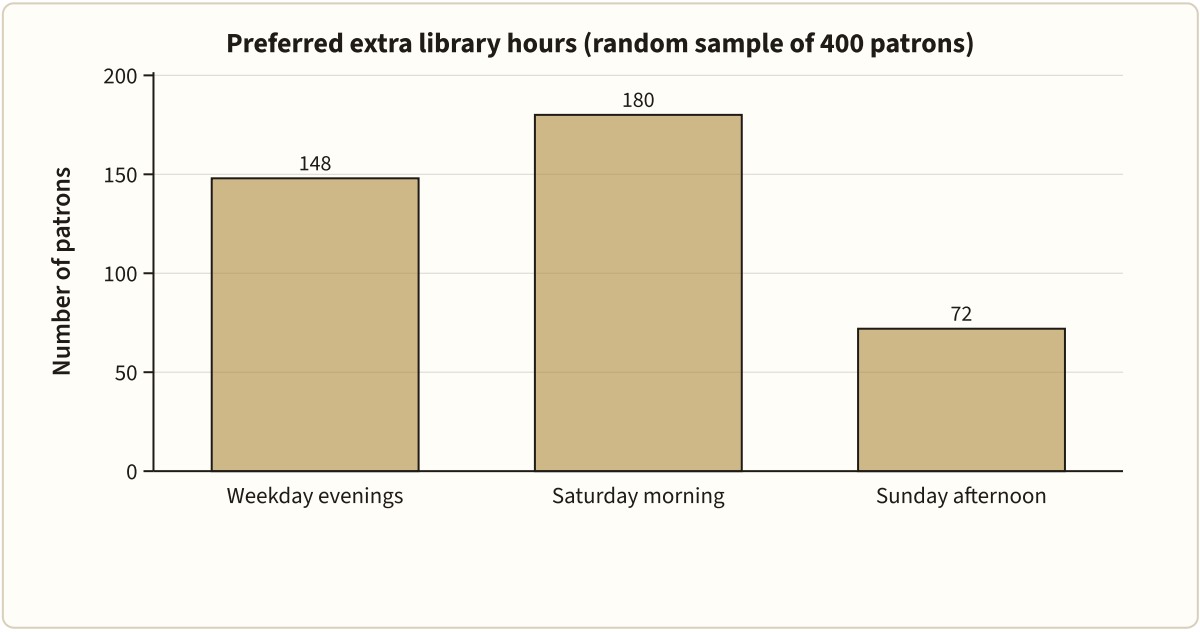
1. State: p = the proportion of all apples in this year's harvest that have blemishes. We want a 90% confidence interval.
2. Plan: one-sample z interval for p. Random: a random sample of apples. 10%: $150 \le 0.10 \times 20{,}000 = 2{,}000$. Large Counts: 33 blemished and $150 - 33 = 117$ unblemished, both at least 10.
3. Do: $\hat{p} = \dfrac{33}{150} = 0.22$. $SE = \sqrt{\dfrac{0.22 \times 0.78}{150}} = \sqrt{\dfrac{0.1716}{150}} = \sqrt{0.001144} \approx 0.033823$.
4. $ME = 1.645 \times 0.033823 \approx 0.05564$. Interval: 0.22 ± 0.05564, from 0.16436 to 0.27564, or 0.164 to 0.276.
5. Conclude: We are 90% confident that the interval from 0.164 to 0.276 captures the true proportion of blemished apples in this year's harvest; that is, between about 16.4% and 27.6%.
6. (b) The whole interval lies above 0.15, so 0.15 is not a plausible value. The data give good reason to believe more than 15% of the harvest is blemished.

Answer: (a) 0.164 to 0.276: we are 90% confident that between 16.4% and 27.6% of all the apples in the harvest have blemishes. (b) No; 0.15 is below the entire interval.

PRACTICE

1. EASY In a random sample of 250 households in a large rural county, 85 keep a vegetable garden. Find the sample proportion $\hat{p}$.
2. EASY For the sample in Problem 1 ($\hat{p} = 0.34$, $n = 250$), find the standard error of $\hat{p}$. Round to 4 decimal places.
3. EASY Henry cracks a random sample of 40 pecans from a large harvest and finds 37 good nuts and 3 bad ones. Is the Large Counts condition met for a confidence interval for the proportion of good pecans?
4. EASY Using a random sample of members of a large regional 4-H program, the leaders find a 95% confidence interval of 0.62 to 0.70 for the proportion of members who raise an animal project. Interpret the interval in context.
5. MEDIUM Use the garden data from Problems 1 and 2 (85 of 250 households) to construct a 95% confidence interval for the proportion of all households in the county that keep a vegetable garden. The county has about 12,000 households. Check the conditions and round to 3 decimal places.

6. MEDIUM A county library with about 9,000 registered patrons asked a random sample of 400 patrons which extra hours they would most like; the bar graph shows the results. (a) Construct a 95% confidence interval for the proportion of all patrons who prefer Saturday morning. Round to 3 decimal places. (b) Does the interval suggest that a majority of all patrons prefer Saturday morning?



7. MEDIUM A large hatchery sets thousands of eggs each week. The table shows what happened to a random sample of 200 eggs. Construct a 99% confidence interval for the proportion of all eggs set that hatch. Round to 3 decimal places.

Random sample of eggs from a large hatchery

Outcome	Number of eggs
Hatched	168
Did not hatch	32
Total	200

8. MEDIUM Explain why the Large Counts condition for a confidence interval is checked with the observed counts $n\hat{p}$ and $n(1-\hat{p})$ instead of np and $n(1-p)$, and why the condition is needed at all.

9. MEDIUM A parish has 500 members. The parish council plans to ask a random sample of 80 members whether they would help with a spring service project and then build a confidence interval for the proportion of all members who would help. Is the 10% condition met?

10. HARD A weather station has 50 years of June records, 1,500 June days in all. A student picks a random sample of 120 of those days and finds that 42 had measurable rain. Use the four-step process to construct a 90% confidence interval for the proportion of all June days in the record with measurable rain. Round to 3 decimal places.

11. **HARD** A seed company claims that 90% of its bean seeds germinate. Grace plants a random sample of 250 seeds from a large shipment, and 210 germinate. (a) Construct a 95% confidence interval for the proportion of all seeds in the shipment that germinate. Round to 3 decimal places. (b) Is 0.90 a plausible value for the germination rate?

12. **HARD** A Christmas-tree farm has 4,000 trees. The owner checks a random sample of 200 trees and finds 26 with deer damage. (a) Construct a 95% confidence interval for the proportion of all the trees with deer damage (3 decimal places). (b) Convert the interval into an interval for the number of damaged trees on the farm, using the unrounded endpoints and rounding to whole trees.

13. **CHALLENGE** A newsletter reports that, based on a random sample of dairy farms in a large region, a 95% confidence interval for the proportion of farms that graze their herds on pasture all summer is 0.56 to 0.64. About how many farms were in the sample?

COMMON MISTAKES

- Checking Large Counts with only the successes. You need at least 10 successes AND at least 10 failures; compute $n - X$ as well as X.
- Using a count instead of a proportion in the formula. Convert first: $\hat{p} = \dfrac{X}{n}$ is a number between 0 and 1.
- Forgetting the square root, or putting n outside it. The standard error is $\sqrt{\hat{p}(1-\hat{p})/n}$, with n under the root.
- Writing a conclusion without context ("we are 95% confident p is between 0.81 and 0.91"). Say what p is: the proportion of ALL seeds in the lot that germinate.

ANSWERS TO PRACTICE

1. $\hat{p} = \dfrac{85}{250} = 0.34$.
2. $SE = \sqrt{\dfrac{0.34 \times 0.66}{250}} = \sqrt{\dfrac{0.2244}{250}} = \sqrt{0.0008976} \approx 0.0300$.
3. No. There are 37 successes but only 3 failures, and 3 is less than 10, so the sampling distribution of $\hat{p}$ may not be close to normal.
4. We are 95% confident that between 62% and 70% of all members of this 4-H program raise an animal project. (The point estimate is 0.66 and the margin of error is 0.04.)
5. Random sample; $250 \leq 1{,}200$; 85 and 165 are both at least 10. $ME = 1.96 \times 0.029960 \approx 0.0587$. Interval 0.34 ± 0.0587: 0.281 to 0.399.
6. (a) $\hat{p} = \dfrac{180}{400} = 0.45$; conditions: random, $400 \leq 900$, 180 and 220 at least 10. $SE = \sqrt{\dfrac{0.45 \times 0.55}{400}} \approx 0.024875$; $ME = 1.96 \times 0.024875 \approx 0.0488$; interval 0.401 to 0.499. (b) No; the entire interval is below 0.5, so a majority is not plausible.

7. $\hat{p} = \dfrac{168}{200} = 0.84$; 168 and 32 are both at least 10; 200 is less than 10% of the eggs set. $SE = \sqrt{\dfrac{0.84 \times 0.16}{200}} = \sqrt{0.000672} \approx 0.025923$; $ME = 2.576 \times 0.025923 \approx 0.0668$; interval 0.773 to 0.907.
8. We do not know p; that is what we are estimating, so we use the counts of successes and failures we actually observed. The condition is needed because the critical value z^* comes from the normal curve, and the sampling distribution of $\hat{p}$ is close to normal only when there are enough successes and failures.
9. No. 10% of 500 is 50, and the sample of 80 is larger than that, so the draws are not close to independent and the usual standard error would be too large.
10. State: p = proportion of all 1,500 June days with measurable rain; 90% confidence. Plan: one-sample z interval for p; random sample; $120 \le 150$; 42 and 78 at least 10. Do: $\hat{p} = 0.35$, $SE = \sqrt{\dfrac{0.35 \times 0.65}{120}} \approx 0.043541$, $ME = 1.645 \times 0.043541 \approx 0.0716$; interval 0.278 to 0.422. Conclude: We are 90% confident that between 27.8% and 42.2% of June days in this record had measurable rain.
11. (a) $\hat{p} = 0.84$; 210 and 40 are at least 10. $SE = \sqrt{\dfrac{0.84 \times 0.16}{250}} = \sqrt{0.0005376} \approx 0.023186$; $ME = 1.96 \times 0.023186 \approx 0.0454$; interval 0.795 to 0.885. (b) No; 0.90 lies above the whole interval, so the data give good reason to doubt the company's claim.
12. (a) $\hat{p} = 0.13$; 26 and 174 at least 10; $200 \le 400$. $SE = \sqrt{\dfrac{0.13 \times 0.87}{200}} = \sqrt{0.0005655} \approx 0.023780$; $ME = 1.96 \times 0.023780 \approx 0.046609$; interval 0.0833906 to 0.1766094, or 0.083 to 0.177. (b) $4{,}000 \times 0.0833906 \approx 334$ and $4{,}000 \times 0.1766094 \approx 706$: about 334 to 706 trees.
13. $\hat{p} = 0.60$ and $ME = 0.04$. Solve $1.96\sqrt{\dfrac{0.60 \times 0.40}{n}} = 0.04$: $\dfrac{0.24}{n} = \left(\dfrac{0.04}{1.96}\right)^2$, so $n = 0.24 \times \left(\dfrac{1.96}{0.04}\right)^2 = 0.24 \times 2401 = 576.24$. About 576 farms.

LESSON 10.3 · 2 SESSIONS OF 45 MINUTES

The t Distributions

OBJECTIVES

- Explain why replacing σ by s adds variability and why the t distributions have heavier tails than the standard normal.
- Compute the standard error $\frac{s}{\sqrt{n}}$ and the degrees of freedom $n - 1$ for a sample mean.
- Read critical values t^* from the t table, including when the exact df is not listed.
- Decide whether a situation calls for z or for t.

VIDEO

WATCH

Introduction to the t Distribution (non-technical)

jbstatistics · 8:54

When to watch: Before Example 1: why we use t when σ is unknown.

https://www.youtube.com/watch?v=Uv6nGIgZMVw

BACKUP

Using the t Table to Find Areas and Percentiles

jbstatistics · 7:55

When to watch: Before Example 1: reading the t table.

https://www.youtube.com/watch?v=qH7QZoMbB80

KEY TERMS

standard error of the mean	The estimate $\frac{s}{\sqrt{n}}$ of the standard deviation of $\bar{x}$, used when σ is unknown.
t distribution	A symmetric, bell-shaped distribution centered at 0 with heavier tails than the standard normal; it describes $t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$ for samples from a normal population.
degrees of freedom (df)	The number that picks out one t distribution from the family; for one sample, $df = n - 1$, the number of deviations that are free to vary.
heavier tails	More area far from the center than the normal curve has, so a larger critical value is needed to hold the middle 95%.

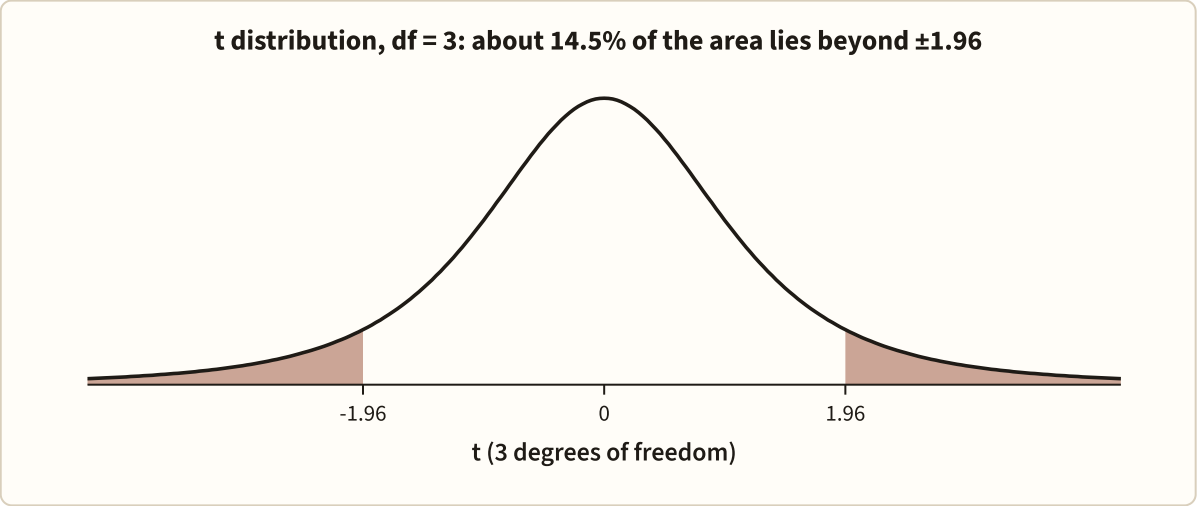
critical value t^*	The number such that the middle C% of the t distribution with the given df lies between $-t^*$ and t^*.

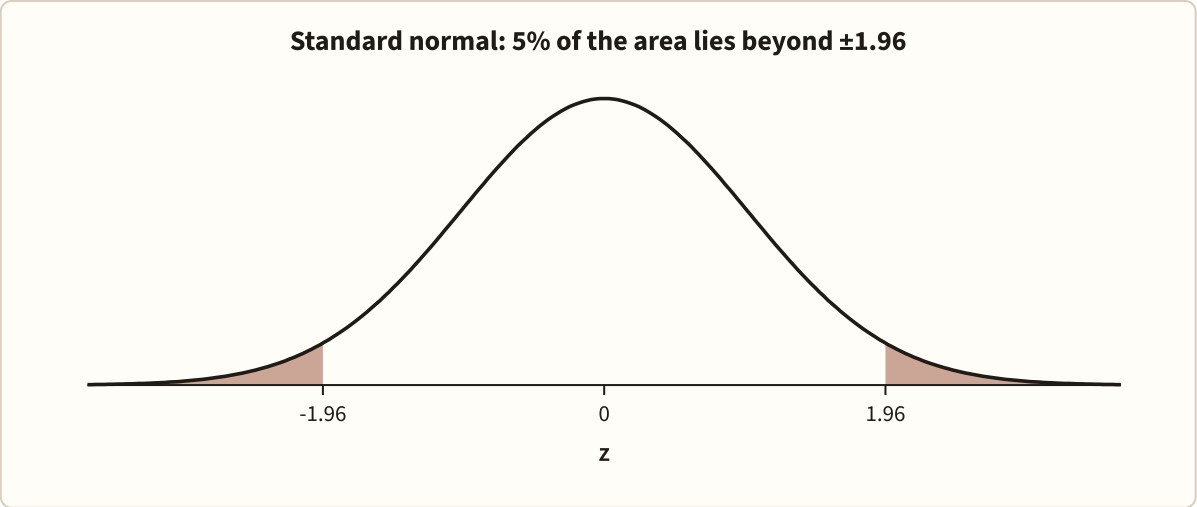
CORE IDEAS

In Lesson 10.1 we pretended to know σ, the population standard deviation. In real life we almost never do. If you weigh a random sample of lambs, you do not know the standard deviation of all the lambs' weights; you know only the sample standard deviation s. The natural fix is to use s in place of σ. The result, $\frac{s}{\sqrt{n}}$, is the **standard error** of $\bar{x}$: our estimate of how far $\bar{x}$ typically falls from μ.

But this fix has a cost. When σ is known, the standardized value $z = \dfrac{\bar{x} - \mu}{\sigma/\sqrt{n}}$ has only one part that changes from sample to sample, the top. With s in its place, $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$ has two parts that change: $\bar{x}$ on top and s on the bottom. In some samples s happens to come out smaller than σ, and dividing by a too-small number makes t larger. So t values are more spread out than z values, with more of them far from 0. If we kept using $z^* = 1.96$, our "95%" intervals would capture μ less than 95% of the time, especially with small samples.

The two curves below show how much this matters for a sample of only $n = 4$. Beyond ± 1.96 the standard normal curve has 5% of its area, but the t distribution with 3 degrees of freedom has about 14.5%. A t interval must reach farther to hold the middle 95%.





The **t distributions** are a whole family, one for each number of **degrees of freedom**. For a single sample, $df = n - 1$: the same $n - 1$ you divided by when you computed s in Lesson 2.4, because the n deviations from $\bar{x}$ must add to 0, so only $n - 1$ of them are free. Every t distribution is symmetric about 0 and bell-shaped, with heavier tails than the normal. The fewer the degrees of freedom, the less information s carries about σ, and the heavier the tails. As df grows, s becomes a reliable estimate of σ, and the t distribution comes closer and closer to the standard normal.

To find t^*, use the t table: go to the row for your df and the column for your confidence level. If your df is not in the table, use the next smaller df that is listed; that gives a slightly larger t^* and a slightly wider interval, so you stay on the safe side. Technology gives the exact value. Look down any column: t^* shrinks as df grows, and the bottom row is just z^*. For example, with $n = 10$ and 95% confidence, $df = 9$ and $t^* = 2.262$, compared with $z^* = 1.960$.

Critical values t* (excerpt from the t table)

df	90%	95%	99%
1	6.314	12.706	63.657
2	2.920	4.303	9.925
3	2.353	3.182	5.841
4	2.132	2.776	4.604
5	2.015	2.571	4.032
9	1.833	2.262	3.250
10	1.812	2.228	3.169
11	1.796	2.201	3.106
14	1.761	2.145	2.977
15	1.753	2.131	2.947
19	1.729	2.093	2.861
24	1.711	2.064	2.797
29	1.699	2.045	2.756
40	1.684	2.021	2.704
60	1.671	2.000	2.660
100	1.660	1.984	2.626
1000	1.646	1.962	2.581
z*	1.645	1.960	2.576

When do you use which? For proportions, use z (Lesson 10.2): there is no separate σ to estimate, because the standard deviation of $\hat{p}$ depends only on p. For means, use z only in the rare case that σ is truly known (for example, a measuring instrument with a long, well-documented record). Whenever you estimate σ with s, use t with $df = n - 1$. Lesson 10.4 puts this to work in the one-sample t interval.

KEY FORMULAS

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

$$t = \frac{x - \mu}{s/\sqrt{n}}, \qquad df = n - 1$$

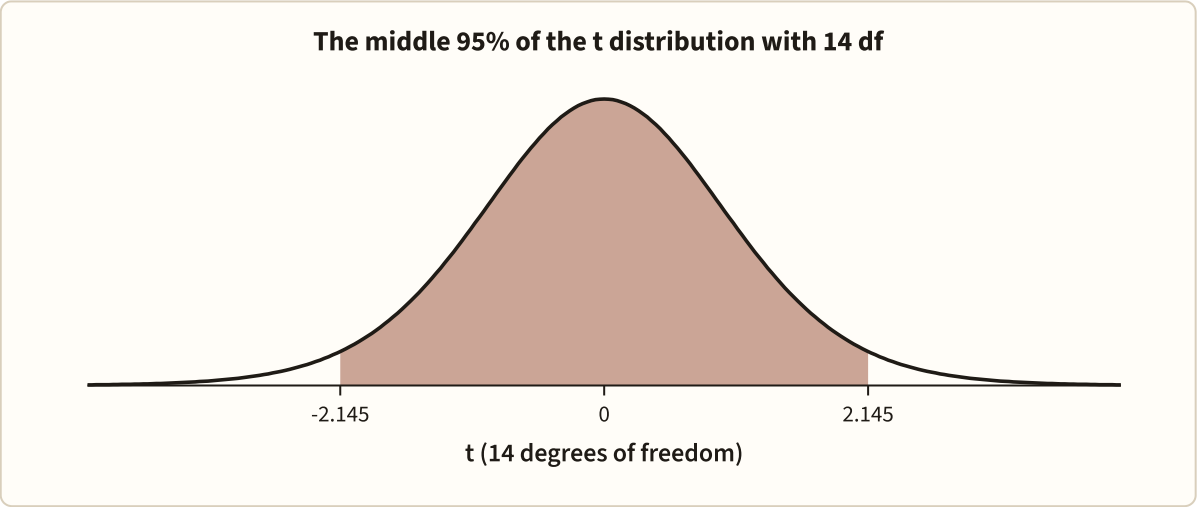
$t^* > z^*$ for every df, and t^* gets closer to z^* as df grows

From history. William Sealy Gosset was a chemist at the Guinness brewery in Dublin who studied small samples of barley and hops, where using the normal curve with s gave intervals that were too narrow. He worked out the correct distribution and published it in 1908 under the pen name "Student", which is why the t distribution is still often called Student's t.

Practice habit. Write $df = n - 1$ before you open the table, and if your df is not listed, use the next smaller df.

WORKED EXAMPLES

Example 1. Find the critical value t^* for (a) a 95% confidence interval from a sample of $n = 15$ and (b) a 99% confidence interval from a sample of $n = 25$.



1. (a) $df = n - 1 = 15 - 1 = 14$.
2. In the t table, row 14, column 95%: $t^* = 2.145$. The middle 95% of the t distribution with 14 df lies between -2.145 and 2.145 (see the curve).
3. (b) $df = 25 - 1 = 24$.
4. Row 24, column 99%: $t^* = 2.797$.
5. Both are larger than the matching z^* (1.960 and 2.576), as they must be, because estimating σ with s adds variability.

Answer: (a) $t^* = 2.145$ (df = 14). (b) $t^* = 2.797$ (df = 24).

Example 2. Ruth weighs a random sample of 10 lambs from her family's large flock at 30 days old. The sample standard deviation is $s = 4.2$ pounds. For a 95% confidence interval for the mean weight of all the flock's 30-day-old lambs, find (a) the standard error of $\bar{x}$, (b) t^* and (c) the margin of error. Round to 3 decimal places. (d) How much larger is this margin of error than the one you would get by wrongly using z^*?

1. (a) $SE = \dfrac{s}{\sqrt{n}} = \dfrac{4.2}{\sqrt{10}} = \dfrac{4.2}{3.16228} \approx 1.328$ pounds.
2. (b) $df = 10 - 1 = 9$; the table, row 9, column 95%, gives $t^* = 2.262$.
3. (c) $ME = t^* \times SE = 2.262 \times 1.328157 \approx 3.004$ pounds.
4. (d) With z^*: $1.960 \times 1.328157 \approx 2.603$ pounds. The ratio is $\dfrac{2.262}{1.960} \approx 1.15$, so the correct margin of error is about 15% larger.
5. Conclusion: Because Ruth estimates σ from only 10 lambs, her interval must be about 15% wider than a z interval would be; using z^* would give an interval that captures the true mean lamb weight less than 95% of the time.

Answer: (a) $SE \approx 1.328$ pounds. (b) $t^* = 2.262$. (c) $ME \approx 3.004$ pounds, about 15% larger than the (incorrect) z margin of 2.603 pounds.

PRACTICE

1. EASY A baker weighs a random sample of 20 loaves from a week's output. How many degrees of freedom does the t distribution for this sample have?
2. EASY A random sample of 12 calves will be used for a 90% confidence interval for their mean daily weight gain. Give the df and t^*.
3. EASY A random sample of 16 boards from a sawmill has a standard deviation of $s = 0.24$ inch in length. Find the standard error of the sample mean length.
4. EASY For a sample of 6 observations, the 95% critical value is $t^* = 2.571$, while $z^* = 1.960$. Explain in one or two sentences why t^* must be larger.
5. MEDIUM An astronomy club counts meteors per hour on a random sample of 30 clear nights. Find t^* for a 99% confidence interval for the mean hourly count.
6. MEDIUM Samuel has a random sample of $n = 45$ fence posts and wants a 95% confidence interval for their mean length. His t table lists df = 40 and df = 60 but not df = 44. Using the table, what t^* should he use, and why?
7. MEDIUM A hay farmer's square bales are supposed to average 50 pounds. A random sample of 16 bales has $\bar{x} = 52.5$ pounds and $s = 6$ pounds. How many standard errors above 50 pounds is $\bar{x}$? Round to 2 decimal places.

8. MEDIUM Should each interval use z^* or t^*? Explain briefly. (a) The proportion of a large flock's eggs that are double-yolked, from a random sample of 500 eggs. (b) The mean brightness of a star in repeated measurements by an instrument whose standard deviation σ is known from years of testing. (c) The mean weight of pumpkins in a patch, from a random sample of 15 pumpkins with sample standard deviation s.

9. MEDIUM A random sample of 25 of Grace's practice drives at the golf range has a standard deviation of $s = 12$ yards. For a 95% confidence interval for her mean drive distance, find (a) the standard error, (b) t^* and (c) the margin of error, to 2 decimal places.

10. HARD Henry says, "The t distribution is just the normal curve drawn wider. Why should the amount of widening depend on the sample size?" Write an explanation he would understand.

11. HARD A beekeeper weighs the honey from a random sample of 5 hives; the sample standard deviation is $s = 10$ pounds. (a) Find the margin of error for a 95% confidence interval for the mean harvest per hive. (b) Find the margin of error you would get by wrongly using $z^* = 1.960$. Round both to 2 decimal places.

12. HARD A rancher reports that a 95% confidence interval for the mean weight gain of a random sample of 16 steers had a margin of error of 4.262 pounds. What was the sample standard deviation s?

13. CHALLENGE A seed catalog says a watermelon variety averages 13 pounds. Grace weighs a random sample of 10 of her melons of this variety: $\bar{x} = 14.2$ pounds and $s = 1.5$ pounds. (a) How many standard errors is $\bar{x}$ from 13? Round to 2 decimal places. (b) Is $\bar{x}$ more than t^* standard errors from 13, using 95% confidence? (c) What does your answer to (b) say about whether 13 pounds would fall inside her 95% confidence interval?

COMMON MISTAKES

- Using z^* when σ has been estimated by s. If you computed s from the sample, use t^*.
- Using $df = n$ instead of $df = n - 1$. Subtract 1 first, then look up the row.
- Reading the wrong column. Make sure the column you use is labeled with your confidence level (some tables label columns by the area in one tail; for 95% confidence that is 0.025).
- Rounding df up when it is not in the table (using 60 for df = 44). Always use the next SMALLER df listed, so the interval is a little wider, not narrower.

ANSWERS TO PRACTICE

1. $df = n - 1 = 20 - 1 = 19$.
2. $df = 11$; the table, row 11, column 90%, gives $t^* = 1.796$.
3. $SE = \dfrac{0.24}{\sqrt{16}} = \dfrac{0.24}{4} = 0.06$ inch.

4. With only 6 observations, s may be well above or below σ, and when it is too small the t statistic comes out large. The t distribution with 5 df therefore has heavier tails, so we must go out farther (2.571 instead of 1.960) to hold the middle 95%.
5. $df = 29$; row 29, column 99%: $t^* = 2.756$.
6. $df = 44$ is not listed, so use the next smaller df, 40: $t^* = 2.021$. This is slightly larger than the exact value (technology gives 2.015 for df = 44), so the interval is slightly wider and errs on the safe side.
7. $SE = \dfrac{6}{\sqrt{16}} = 1.5$; $t = \dfrac{52.5 - 50}{1.5} = \dfrac{2.5}{1.5} \approx 1.67$ standard errors.
8. (a) z^*: an interval for a proportion uses the normal curve. (b) z^*: σ is truly known, so nothing extra is estimated. (c) t^* with df = 14: σ is unknown and estimated by s.
9. (a) $SE = \dfrac{12}{5} = 2.4$ yards. (b) $df = 24$, $t^* = 2.064$. (c) $ME = 2.064 \times 2.4 = 4.9536 \approx 4.95$ yards.
10. The extra spread comes from using s in place of σ. With a small sample, s can be far from σ, so t values vary a lot and the tails are heavy. With a large sample, s is almost always close to σ, so t behaves almost exactly like z. That is why the t distribution depends on df = n − 1 and gets closer to the normal curve as n grows.
11. $SE = \dfrac{10}{\sqrt{5}} \approx 4.47214$; df = 4. (a) $ME = 2.776 \times 4.47214 \approx 12.41$ pounds. (b) $1.960 \times 4.47214 \approx 8.77$ pounds. With only 5 hives the correct margin is about 42% larger.
12. $df = 15$, $t^* = 2.131$. $4.262 = 2.131 \times \dfrac{s}{\sqrt{16}}$, so $\dfrac{s}{4} = \dfrac{4.262}{2.131} = 2$ and $s = 8$ pounds.
13. (a) $SE = \dfrac{1.5}{\sqrt{10}} \approx 0.474342$; $t = \dfrac{14.2 - 13}{0.474342} \approx 2.53$. (b) Yes: df = 9, $t^* = 2.262$, and $2.53 > 2.262$. (c) The interval reaches only t^* standard errors on each side of 14.2, so 13 lies outside it; 13 pounds is not a plausible mean for Grace's melons.

LESSON 10.4 · 2 SESSIONS OF 45 MINUTES

Estimating a Population Mean

OBJECTIVES

- Explain why the one-sample t interval $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$ captures μ in C% of all samples.
- Check the Random, 10% and Normal/Large Sample conditions, using a graph of the data for small samples.
- Construct a one-sample t interval for μ from raw data or summary statistics using the four-step process.
- Interpret the interval in context and avoid treating it as a range for individual values.

VIDEO

WATCH

Confidence Intervals for One Mean: Sigma Not Known (t Method)

jbstatistics · 9:46

When to watch: Before Example 1: the one-sample t interval.

https://www.youtube.com/watch?v=bFefxSE5bmo

BACKUP

Confidence Intervals for One Mean: Sigma Not Known (t Method)

jbstatistics · 6:03

When to watch: Extra practice: a one-sample t interval with the t table.

https://www.youtube.com/watch?v=_NGYJxrUGgQ

KEY TERMS

one-sample t interval	The confidence interval $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$ for a population mean μ, with t^* from the t distribution with $n - 1$ degrees of freedom.
Normal/Large Sample condition	The population is normal, or $n \geq 30$, or (for a small sample) a dotplot, stemplot or boxplot of the data shows no strong skewness and no outliers.
robust	A procedure is robust if its confidence level stays close to the stated level even when a condition, such as normality, is not exactly met.

plausible value	A value of the parameter that lies inside the confidence interval and so is consistent with the data at that confidence level.

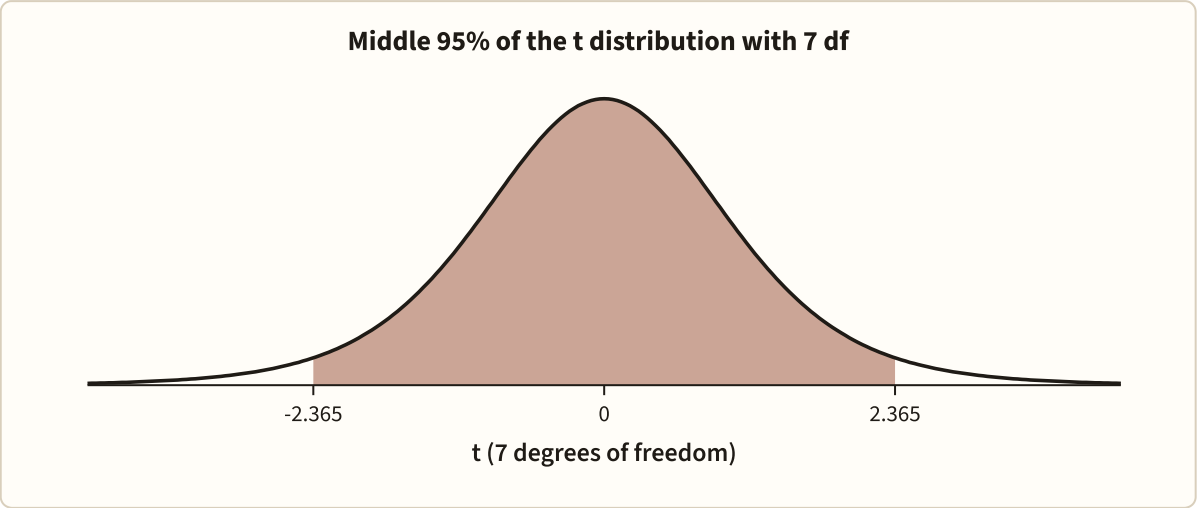
CORE IDEAS

Now the pieces fit together. To estimate a population mean μ when σ is unknown, use the **one-sample t interval**

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}, \qquad df = n - 1.$$

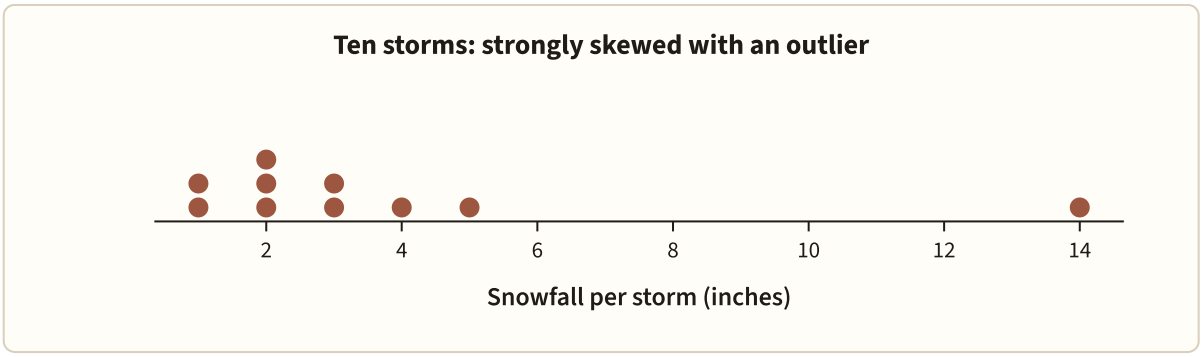
It has the same shape as every interval in this unit: statistic ± (critical value) × (standard error).

Why does it work? When the conditions hold, the statistic $t = \dfrac{\bar{x} - \mu}{s/\sqrt{n}}$ follows a t distribution with $n - 1$ df, so in 95% of all samples it lands between $-t^*$ and t^*. That says $\bar{x}$ is within t^* standard errors of μ. Turn it around, exactly as in Lesson 10.1: μ is then within t^* standard errors of $\bar{x}$, so the interval captures μ in 95% of all samples. The curve shows the middle 95% for a sample of 8 (df = 7).



The conditions, and why each matters: **Random**: a random sample (or randomized experiment), so the sampling distribution applies and the estimate is not biased. **10%**: when sampling without replacement, n is at most 10% of the population. **Normal/Large Sample**: the t distribution is exact when the population is normal; by the central limit theorem it is close enough when $n \geq 30$. For a small sample we cannot see the population, so we look at the sample: make a dotplot or stemplot and check for strong skewness or outliers.

t procedures are **robust**: moderate skewness in the data changes the true confidence level only a little, and less and less as n grows. Outliers are another matter. Both $\bar{x}$ and s are not resistant, so one wild value can drag the center of the interval and stretch its width. The snowfall data below come from 10 storms; the single 14-inch storm pulls $\bar{x}$ up and inflates s. With a sample this small and this lopsided, a t interval is not safe. When you find an outlier, first check whether it is a recording mistake; if it is a real value, do not hide it, and do not trust a t interval from a small sample that contains it.



Use the four steps every time. **State:** "μ = the mean … of all …; 95% confidence." **Plan:** "one-sample t interval for μ," then check Random, 10% and Normal/Large Sample with evidence. **Do:** $\bar{x}$, s, df, t^*, standard error, margin of error, interval. **Conclude:** "We are 95% confident that the interval from … to … captures the true mean … of all …" Give units.

Remember what the interval is about. It estimates the *mean* of the population, not the values of individuals. In Worked Example 2, the interval for the mean weight of the lambs is only about 5 pounds wide, but individual lambs in the sample range from 47 to 56 pounds, and many lambs in the flock weigh less or more than anything in the interval. Means vary much less than individual values (Lesson 9.3), which is exactly why the interval is narrow.

KEY FORMULAS

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}, \qquad df = n - 1$$

$$SE_{\bar{x}} = \frac{s}{\sqrt{n}}$$

Normal/Large Sample: population normal, or $n \geq 30$, or a graph of the data shows no strong skewnes

Practice habit. Always plot the data first. Before you compute a t interval from a small sample, make a quick dotplot or stemplot and look for strong skewness and outliers.

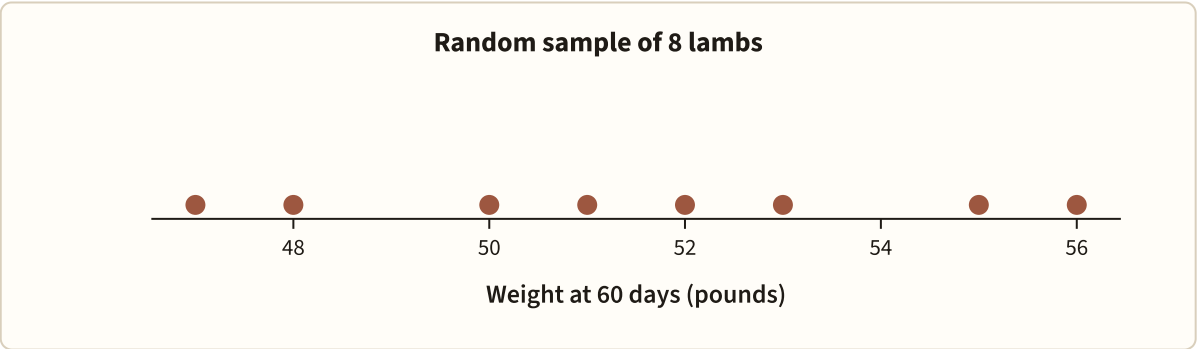
WORKED EXAMPLES

Example 1. A sawmill foreman measures a simple random sample of 16 boards from a day's output of several thousand. The lengths are roughly symmetric with no outliers; $\bar{x} = 96.05$ inches and $s = 0.20$ inch. Construct a 95% confidence interval for the mean length of all boards cut that day. Round to 2 decimal places.

1. Conditions: SRS; 16 boards is less than 10% of several thousand; $n = 16 < 30$, but the data show no strong skewness or outliers.
2. $df = 16 - 1 = 15$, so $t^* = 2.131$.
3. $SE = \dfrac{s}{\sqrt{n}} = \dfrac{0.20}{\sqrt{16}} = \dfrac{0.20}{4} = 0.05$ inch.
4. $ME = 2.131 \times 0.05 = 0.10655$ inch.
5. Interval: 96.05 ± 0.10655, from 95.94345 to 96.15655, or 95.94 to 96.16 inches.
6. Conclusion: We are 95% confident that the mean length of all boards cut that day is between 95.94 and 96.16 inches.

Answer: 95.94 to 96.16 inches. We are 95% confident that the mean length of all boards cut that day is in this interval.

Example 2. A sheep ranch has about 400 lambs born this spring. The rancher weighs a random sample of 8 of them at 60 days: 52, 48, 55, 50, 47, 53, 51, 56 pounds. Use the four-step process to construct a 95% confidence interval for the mean 60-day weight of all the spring lambs. Round to 1 decimal place.



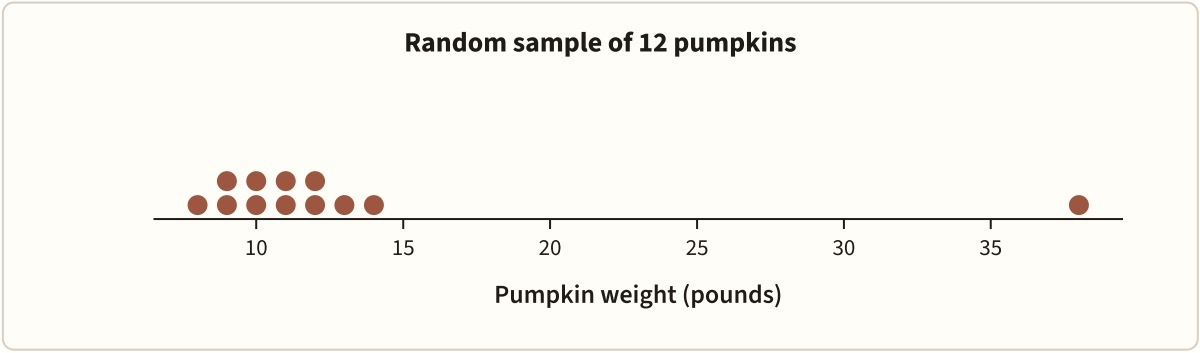
1. State: μ = the mean 60-day weight of all lambs born on this ranch this spring. We want a 95% confidence interval.
2. Plan: one-sample t interval for μ. Random: a random sample of lambs. 10%: $8 \le 0.10 \times 400 = 40$. Normal/Large Sample: $n = 8$ is small, but the dotplot shows no strong skewness and no outliers.
3. Do: $\bar{x} = \dfrac{412}{8} = 51.5$ pounds. Deviations: 0.5, −3.5, 3.5, −1.5, −4.5, 1.5, −0.5, 4.5; their squares add to $0.25 + 12.25 + 12.25 + 2.25 + 20.25 + 2.25 + 0.25 + 20.25 = 70$.
4. $s = \sqrt{\dfrac{70}{8-1}} = \sqrt{10} \approx 3.162$ pounds.
5. $df = 7$, $t^* = 2.365$. $SE = \dfrac{3.162}{\sqrt{8}} = \sqrt{\dfrac{10}{8}} = \sqrt{1.25} \approx 1.118$ pounds.
6. $ME = 2.365 \times 1.118034 \approx 2.644$. Interval: 51.5 ± 2.644, from 48.856 to 54.144, or 48.9 to 54.1 pounds.
7. Conclude: We are 95% confident that the interval from 48.9 to 54.1 pounds captures the true mean 60-day weight of all the lambs born on this ranch this spring.

Answer: 48.9 to 54.1 pounds. We are 95% confident that the mean 60-day weight of all this spring's lambs is between 48.9 and 54.1 pounds.

PRACTICE

1. EASY A bakery weighs a random sample of 25 loaves from a week's output of about 3,000 loaves. A dotplot shows no strong skewness or outliers; $\bar{x} = 20.4$ ounces and $s = 0.5$ ounce. Construct a 95% confidence interval for the mean weight of all the week's loaves. Round to 2 decimal places.
2. EASY From a random sample of eggs laid by a large flock, a farmer computes a 90% confidence interval for the mean egg weight: 57.8 to 60.2 grams. Interpret the interval in context.

3. **EASY** At a church pumpkin patch, a volunteer weighs a random sample of 12 pumpkins (dotplot shown). Is it safe to use a one-sample t interval for the mean weight of all the pumpkins in the patch?



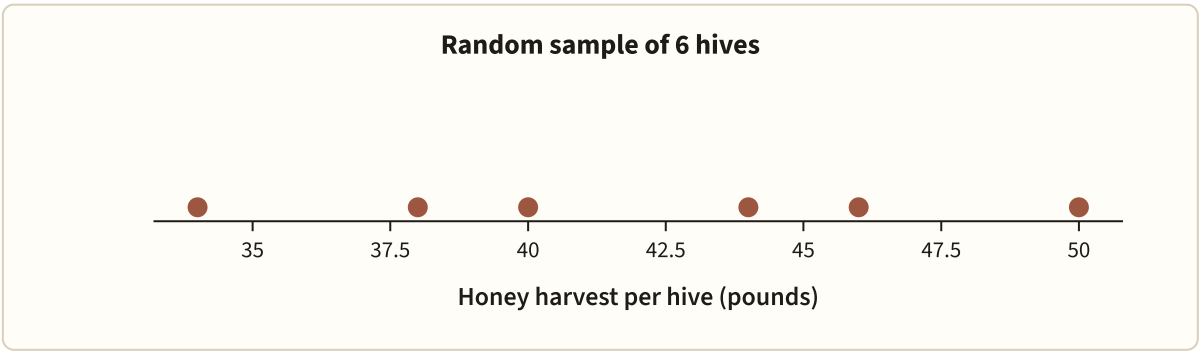
4. **EASY** A weather station has records of thousands of rainstorms. A random sample of 40 storms gives rainfall amounts that are skewed right with no extreme outliers. Is the Normal/Large Sample condition met for a t interval for the mean rainfall per storm?

5. **MEDIUM** The computer output describes the daily milk yield of a random sample of 20 cows from a large dairy herd; a dotplot of the data is roughly symmetric. (a) Check the SE Mean value. (b) Construct a 90% confidence interval for the mean daily yield of all cows in the herd. Round to 2 decimal places.

Computer output: random sample of cows

Variable	N	Mean	StDev	SE Mean
Milk yield (lb/day)	20	64.3	5.8	1.297

6. **MEDIUM** A beekeeper with 80 hives weighs the honey from a random sample of 6: 40, 46, 38, 44, 50, 34 pounds (dotplot shown). (a) Find $\bar{x}$ and s (1 decimal place). (b) Construct a 95% confidence interval for the mean harvest per hive for all 80 hives. Round to 1 decimal place.



7. **MEDIUM** After reading Worked Example 2, Grace says, "So 95% of the lambs on the ranch weigh between 48.9 and 54.1 pounds." Explain her mistake.

8. **MEDIUM** Honey jars are labeled 16 ounces. From a random sample of jars filled by a small apiary's machine, a 95% confidence interval for the mean fill is 15.82 to 15.96 ounces. Does the interval give reason to believe the machine's mean fill is below 16 ounces? Explain.

9. MEDIUM A weather station has 130 years of April rainfall records. A student picks a random sample of 12 Aprils; a dotplot shows no strong skewness or outliers. The sample has $\bar{x} = 3.42$ inches and $s = 0.90$ inch. Construct a 90% confidence interval for the mean April rainfall over all 130 years. Round to 2 decimal places.

10. HARD A large ranch weans about 900 calves. A random sample of 36 calves has mean weaning weight $\bar{x} = 486$ pounds and $s = 42$ pounds. Use the four-step process to construct a 99% confidence interval for the mean weaning weight of all the calves. Round to the nearest pound.

11. HARD Two farmers estimate the mean weight of potatoes from the same large harvest. Both samples have $s = 6$ ounces. Samuel weighs 9 potatoes; Ruth weighs 25. (a) Find the margin of error for each 95% interval, to 2 decimal places. (b) The square root of Ruth's sample size is $\sqrt{25}/\sqrt{9} = 5/3$ times that of Samuel's, yet her margin of error is less than $3/5$ of his. Explain why.

12. HARD An astronomy club counts meteors on a random sample of 16 clear nights during an annual shower. Their 95% confidence interval for the mean number of meteors per hour is 31.6 to 36.4. Find $\bar{x}$ and s (2 decimal places).

13. CHALLENGE A random sample of 25 watermelons from a large field has $\bar{x} = 20$ pounds and $s = 4$ pounds. A report gives a confidence interval of 18.6312 to 21.3688 pounds from these data but forgets to state the confidence level. What confidence level (as a percent) was used?

COMMON MISTAKES

- Skipping the graph for a small sample. With $n < 30$, plot the data and check for strong skewness and outliers before using t.
- Using z^* instead of t^*, or $df = n$ instead of $n - 1$. When s estimates σ, use t^* with $df = n - 1$.
- Interpreting the interval as a range for individual values. It estimates the population mean, so say "the mean ... of all ...".
- Dividing s by n instead of $\sqrt{n}$. The standard error is $\dfrac{s}{\sqrt{n}}$.

ANSWERS TO PRACTICE

1. $df = 24$, $t^* = 2.064$; $SE = \dfrac{0.5}{5} = 0.1$; $ME = 0.2064$. Interval 20.4 ± 0.2064: 20.19 to 20.61 ounces.
2. We are 90% confident that the interval from 57.8 to 60.2 grams captures the true mean weight of all the eggs this flock lays.
3. No. The sample is small ($n = 12 < 30$) and the 38-pound pumpkin is a clear outlier, which would pull $\bar{x}$ up and inflate s.
4. Yes. $n = 40 \geq 30$, so by the central limit theorem $\bar{x}$ is approximately normal despite the moderate skewness.

5. (a) $\frac{5.8}{\sqrt{20}} = \frac{5.8}{4.47214} \approx 1.297$. (b) $df = 19$, $t^* = 1.729$; $ME = 1.729 \times 1.29692 \approx 2.2424$; interval 64.3 ± 2.2424: 62.06 to 66.54 pounds per day.
6. (a) $\bar{x} = \frac{252}{6} = 42.0$; squared deviations $4 + 16 + 16 + 4 + 64 + 64 = 168$; $s = \sqrt{\frac{168}{5}} = \sqrt{33.6} \approx 5.8$ pounds. (b) $6 \leq 8$ (10% of 80); no outliers or strong skew. $df = 5$, $t^* = 2.571$; $SE = \sqrt{\frac{33.6}{6}} = \sqrt{5.6} \approx 2.3664$; $ME \approx 6.084$; interval 35.9 to 48.1 pounds.
7. The interval estimates the mean weight μ of all the lambs, not the weights of individual lambs. Individual weights vary much more than means do; even the 8 lambs in the sample ranged from 47 to 56 pounds, so far fewer than 95% of the lambs fall in this narrow interval.
8. Yes. Every value in the interval is below 16, so 16 ounces is not a plausible value for the mean fill; the data suggest the machine is under-filling on average.
9. $12 \leq 13$ (10% of 130). $df = 11$, $t^* = 1.796$; $SE = \frac{0.90}{\sqrt{12}} \approx 0.25981$; $ME \approx 0.4666$; interval 3.42 ± 0.4666: 2.95 to 3.89 inches.
10. State: μ = mean weaning weight of all calves on the ranch; 99%. Plan: one-sample t interval; random sample; $36 \leq 90$; $n = 36 \geq 30$. Do: $SE = \frac{42}{6} = 7$; $df = 35$, $t^* = 2.724$ (table with df 30: 2.750); $ME = 2.724 \times 7 \approx 19.07$; interval 466.93 to 505.07, or 467 to 505 pounds (the table value gives the same rounded interval). Conclude: We are 99% confident that the mean weaning weight of all the ranch's calves is between 467 and 505 pounds.
11. (a) Samuel: $SE = 2$, $df = 8$, $t^* = 2.306$, $ME = 4.61$ ounces. Ruth: $SE = 1.2$, $df = 24$, $t^* = 2.064$, $ME = 2.48$ ounces. (b) A larger sample shrinks the standard error AND gives more degrees of freedom, so t^* is smaller too. Both effects narrow the interval.
12. $\bar{x} = \frac{31.6 + 36.4}{2} = 34.00$; $ME = 2.4$. With $df = 15$, $t^* = 2.131$: $2.4 = 2.131 \times \frac{s}{4}$, so $s = \frac{9.6}{2.131} \approx 4.50$ meteors per hour.
13. $ME = 21.3688 - 20 = 1.3688$ and $SE = \frac{4}{5} = 0.8$, so $t^* = \frac{1.3688}{0.8} = 1.711$. In the row for df = 24, 1.711 is in the 90% column. The level is 90%.

LESSON 10.5 · 2 SESSIONS OF 45 MINUTES

Margin of Error and Sample Size

OBJECTIVES

- Explain what the margin of error covers (random sampling variation) and what it does not cover (bias from poor sampling, nonresponse or question wording).
- Compute the sample size needed for a desired margin of error for a proportion, using $p^* = 0.5$ when there is no prior guess, and explain why 0.5 is the safe choice.
- Compute the sample size needed for a desired margin of error for a mean, using a planning value of σ.
- Estimate the 95% margin of error for a proportion quickly with $\frac{1}{\sqrt{n}}$.

VIDEO

WATCH

Inference for One Proportion: Determining the Required Sample Size

jbstatistics · 6:46

When to watch: Before Example 1: sample size for a proportion.

https://www.youtube.com/watch?v=PCL3g1RLUNU

BACKUP

Confidence Intervals for One Mean: Determining the Required Sample Size

jbstatistics · 5:15

When to watch: Before Example 2: sample size for a mean.

https://www.youtube.com/watch?v=7zcbVaVz_P8

KEY TERMS

sampling error	The random difference between a statistic and the parameter that comes from which individuals happened to be chosen; the margin of error measures this.
nonsampling error	Error from the way the data were gathered, such as undercoverage, nonresponse, voluntary response or misleading question wording; the margin of error does not include it.
planning value p^*	A guess at the population proportion used only to choose a sample size; use $p^* = 0.5$ when you have no good guess.

planning value of σ A reasonable value for the population standard deviation, from past records or a small pilot study, used to choose a sample size for estimating a mean.

CORE IDEAS

The margin of error is (critical value) × (standard error). It measures one thing only: how much the estimate could vary because of *which* individuals chance put in the sample. It says nothing about errors in how the sample was gathered or how the questions were asked (Lesson 5.3). If those other errors are present, the true error can be far larger than the stated margin of error.

What the margin of error covers

Source of error	Example	Covered by the margin of error?
Random sampling variation	A different random sample of apples gives a slightly different $\hat{p}$	Yes
Undercoverage	Sampling only the trees near the road	No
Nonresponse	Many chosen families never return the form	No
Voluntary response	Only people who feel strongly answer a newsletter poll	No
Wording or response bias	A leading question about the library's hours	No

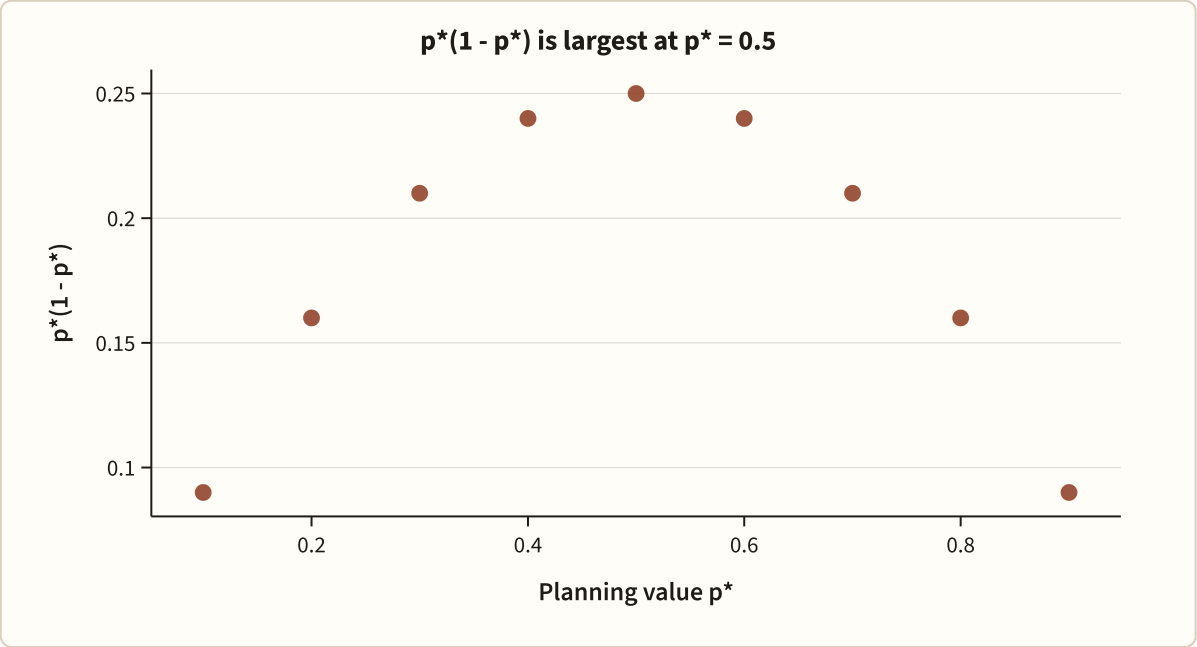
A larger sample shrinks the margin of error, but it never fixes bias. A voluntary-response survey of 5,000 people can report a tiny margin of error and still be badly wrong, because the people who chose to answer are not like the population. A biased method with a huge sample gives a very precise estimate of the wrong number. Good design comes first; sample size comes second.

With a good design in hand, three things control the margin of error: the confidence level (a lower level gives a smaller z^*, but we are less sure), the variability in the population (sometimes reduced by measuring more carefully), and the sample size n. The sample size is the one we usually choose, and we choose it *before* collecting data. For a proportion, we want $z^*\sqrt{\dfrac{p^*(1-p^*)}{n}} \le ME$. Square both sides and solve for n:

$$n \ge \left(\frac{z^*}{ME}\right)^2 p^*(1-p^*).$$

But p^* is a guess at the very proportion we have not yet measured. Look at how $p(1-p)$ behaves: it is largest when $p = 0.5$, where it equals 0.25, and it shrinks toward 0 as p moves toward 0 or 1. So using $p^* = 0.5$ gives the largest n we could ever need: whatever $\hat{p}$ turns out to be, the margin of error will be

no larger than planned. If past experience gives a guess far from 0.5 (for example, about 20% of seeds defective), using it saves effort.



For a mean, we want $z^* \frac{\sigma}{\sqrt{n}} \le ME$, which gives $n \ge \left(\frac{z^* \sigma}{ME} \right)^2$. We need a planning value for σ from past records or a small pilot study. We use z^* rather than t^* because we cannot know df until we know n; when n is not tiny, the two give nearly the same answer. In both formulas, **always round n UP** to the next whole number: rounding down would give a margin of error slightly larger than the one you promised.

A useful mental check: at 95% confidence with $p^* = 0.5$, the margin of error is $1.96\sqrt{\frac{0.25}{n}} = \frac{0.98}{\sqrt{n}}$, which is almost exactly $\frac{1}{\sqrt{n}}$. A sample of 400 gives about $\frac{1}{20} = 0.05$, or 5 percentage points. Because of the square root, halving the margin of error takes four times as many observations, as the table shows.

Largest 95% margin of error for a proportion

Sample size n	0.98 / √n	Quick estimate 1 / √n
100	0.098	0.1
400	0.049	0.05
1000	0.031	0.032
2500	0.02	0.02

KEY FORMULAS

$$n = \left(\frac{z^*}{ME}\right)^2 p^*(1-p^*) \quad \text{(proportion; use } p^* = 0.5 \text{ with no prior guess)}$$

$$n = \left(\frac{z^*\sigma}{ME}\right)^2 \quad \text{(mean; planning value of } \sigma\text{)}$$

Always round n UP to the next whole number.

$$ME \approx \frac{1}{\sqrt{n}} \quad \text{(95\% confidence, proportion near 0.5)}$$

Practice habit. Carry the quick estimate in your head: at 95% confidence, a proportion's margin of error is about $\frac{1}{\sqrt{n}}$ (400 people, about 5 points). And always round a sample size UP.

WORKED EXAMPLES

Example 1. A state forester wants to estimate the proportion of trees in a large state forest that show beetle damage, within 0.04 at 95% confidence. There is no earlier survey to give a guess. How many randomly chosen trees must be inspected?

1. No prior guess, so use $p^* = 0.5$. At 95%, $z^* = 1.96$; $ME = 0.04$.
2. $n = \left(\frac{1.96}{0.04}\right)^2 (0.5)(0.5) = (49)^2 \times 0.25$.
3. $49^2 = 2401$, and $2401 \times 0.25 = 600.25$.
4. Round UP: $n = 601$. (600 trees would give a margin of error just over 0.04.)
5. Conclusion: The forester should inspect a random sample of 601 trees; then, whatever proportion is damaged, the 95% margin of error will be at most 0.04.

Answer: $n = 601$ trees.

Example 2. An apiary's jar-filling machine has had a standard deviation of about 0.3 ounce in past records. The owner wants to estimate the mean fill of the jars. (a) How many jars must be weighed to estimate the mean within 0.05 ounce at 99% confidence? (b) How many if a margin of error of 0.10 ounce is good enough?

1. (a) $z^* = 2.576$, planning $\sigma = 0.3$, $ME = 0.05$.
2. $n = \left(\dfrac{2.576 \times 0.3}{0.05}\right)^2 = \left(\dfrac{0.7728}{0.05}\right)^2 = (15.456)^2 \approx 238.89$.
3. Round UP: $n = 239$ jars.
4. (b) $n = \left(\dfrac{0.7728}{0.10}\right)^2 = (7.728)^2 \approx 59.72$; round UP to $n = 60$ jars.
5. Doubling the margin of error cut the needed sample to about one quarter: $\dfrac{238.89}{4} \approx 59.72$.
6. Conclusion: Weigh a random sample of 239 jars for a 99% margin of error of 0.05 ounce, or 60 jars for a margin of 0.10 ounce.

Answer: (a) 239 jars. (b) 60 jars.

PRACTICE

1. EASY A county fair board printed a survey in its newsletter, and 600 readers mailed it back. The board reported, "About 70% want a longer fair, with a margin of error of 4 percentage points." Explain why the margin of error does not tell the whole story.
2. EASY A seed cooperative wants to estimate the proportion of its members who save their own seed, within 0.05 at 95% confidence, with no prior guess. How large a random sample is needed?
3. EASY Samuel's sample-size calculation gives $n = 150.06$. How many must he sample, and why not 150?
4. EASY Use the quick rule to estimate the 95% margin of error, in percentage points, for a proportion from a random sample of 400 church members.
5. MEDIUM In past years about 20% of the seeds in a supplier's lots had cracked coats. (a) How many seeds must be examined to estimate this year's proportion within 0.03 at 95% confidence, using $p^* = 0.2$? (b) How many would be needed using $p^* = 0.5$?
6. MEDIUM A rancher wants to estimate the mean weaning weight of this year's calves within 3 pounds at 95% confidence. Past records suggest $\sigma \approx 12$ pounds. How many calves should be weighed?
7. MEDIUM An orchard keeper wants to estimate the mean diameter of this year's apples within 0.1 inch at 90% confidence. A small pilot sample suggests $\sigma \approx 0.4$ inch. How many apples should be measured?

8. MEDIUM Explain why using $p^* = 0.5$ is called the "safe" choice when you have no idea what the population proportion is.

9. MEDIUM A beekeepers' association found that a random sample of 150 hives gave a 95% margin of error of 0.08 for the proportion of hives that survived the winter. Next year they plan to check 600 hives. Estimate the new margin of error (complete the table).

95% margins of error for a proportion near 0.5

Sample size n	Margin of error
150	0.08
600	?

10. HARD To estimate the proportion of families who would use a new parish library, a volunteer interviews the first 300 families leaving the 11 o'clock service. She computes a margin of error of 5.7 percentage points. Name two kinds of error her margin of error does not cover, and explain how each could make her estimate wrong.

11. HARD A state beekeepers' association wants to estimate the proportion of hives in the state that survived the winter, within 0.02 at 99% confidence, with no prior guess. How many hives must be in the random sample?

12. HARD A random sample of 625 4-H members gives $\hat{p} = 0.50$ for the proportion who show animals at the county fair. (a) Compute the 95% margin of error exactly. (b) Compute the quick estimate $\frac{1}{\sqrt{n}}$.

13. CHALLENGE A farmer wants to estimate the mean soil pH across a large farm. Each laboratory soil test costs \$18, and the budget is \$1,500. A planning value is $\sigma \approx 0.6$. (a) What is the largest number of tests the farmer can afford? (b) What 95% margin of error can the farmer expect with that many tests? Round to 3 decimal places.

COMMON MISTAKES

- Rounding the sample size to the nearest whole number. Always round UP, even when the decimal is small (150.06 becomes 151).
- Thinking a large sample removes bias. The margin of error covers only random sampling error; a convenience or voluntary-response sample stays biased however large it is.
- Using the sample proportion from the study you have not done yet. Before collecting data, use a prior estimate p^*, or $p^* = 0.5$ if you have none.
- Forgetting to square. In $n = \left(\frac{z^*}{ME}\right)^2 p^*(1-p^*)$ the fraction is squared before multiplying by $p^*(1-p^*)$.

ANSWERS TO PRACTICE

1. The margin of error covers only random sampling variation. These responses are a voluntary response sample: people who care most about the fair chose to answer, so the result may be biased, and the margin of error says nothing about that bias.
2. $n = \left(\frac{1.96}{0.05}\right)^2 (0.25) = (39.2)^2(0.25) = 1536.64 \times 0.25 = 384.16$; round up to 385 members.
3. 151. Rounding down to 150 would make the margin of error slightly larger than he wanted; rounding up guarantees it is no larger.
4. $\frac{1}{\sqrt{400}} = \frac{1}{20} = 0.05$, about 5 percentage points.
5. (a) $\left(\frac{1.96}{0.03}\right)^2 = 4268.44$; $4268.44 \times 0.16 \approx 682.95$; round up to 683 seeds. (b) $4268.44 \times 0.25 \approx 1067.11$; round up to 1068 seeds. A good prior guess saves almost 400 seeds.
6. $n = \left(\frac{1.96 \times 12}{3}\right)^2 = (7.84)^2 = 61.4656$; round up to 62 calves.
7. $n = \left(\frac{1.645 \times 0.4}{0.1}\right)^2 = (6.58)^2 = 43.2964$; round up to 44 apples.
8. The product $p^*(1 - p^*)$ is largest at $p^* = 0.5$ (it equals 0.25 there and is smaller for any other value). Using 0.5 therefore gives the largest sample size that could be needed, so the margin of error will be no bigger than planned no matter what $\hat{p}$ turns out to be.
9. Four times the sample size divides the margin of error by $\sqrt{4} = 2$: $\frac{0.08}{2} = 0.04$.
10. Undercoverage: families who attend other services or who rarely come are left out, and they may use a library less (or more) than 11 o'clock families. Response bias: talking face to face with a volunteer, families may say "yes" to be polite. The margin of error covers only random sampling variation, and her sample is not even random.
11. $n = \left(\frac{2.576}{0.02}\right)^2 (0.25) = (128.8)^2(0.25) = 16589.44 \times 0.25 = 4147.36$; round up to 4,148 hives.
12. (a) $1.96\sqrt{\frac{0.5 \times 0.5}{625}} = 1.96 \times \frac{0.5}{25} = 1.96 \times 0.02 = 0.0392$. (b) $\frac{1}{\sqrt{625}} = \frac{1}{25} = 0.04$. The quick rule is very close.
13. (a) $\frac{1500}{18} \approx 83.3$, so 83 tests (84 would cost \$1,512). (b) $ME = 1.96 \times \frac{0.6}{\sqrt{83}} = \frac{1.176}{9.11043} \approx 0.129$ pH units.

Unit 10 test

Time: 60 minutes. You may use a scientific or graphing calculator and the z and t tables at the back of the course. Show your work. For full inference problems, show the four steps (State, Plan, Do, Conclude) briefly. Keep full precision until the last step and round as each problem states. Total: 100 points.

1. **5 PTS** A fisheries biologist reports a 95% confidence interval for the mean length of trout in a mountain stream: 11.6 to 13.0 inches. Find the point estimate and the margin of error.
2. **5 PTS** (a) What is the critical value z^* for 99% confidence? (b) Explain why it is larger than the critical value for 95% confidence.
3. **6 PTS** From a random sample of taps in a large sugarbush (maple grove), a family finds a 90% confidence interval for the mean amount of maple syrup produced per tap this season: 0.21 to 0.27 gallons. (a) Interpret the interval. (b) Explain what "90% confident" means here.
4. **6 PTS** An observatory's instrument measures a star's brightness with a known standard deviation of $\sigma = 0.08$ magnitude, and its measurement errors are normally distributed. Sixteen measurements of one star have mean 6.42 magnitudes. Construct a 95% confidence interval for the star's true brightness. Round to 3 decimal places.
5. **8 PTS** A farm's strawberry harvest is about 50,000 berries. The table shows the grading of a random sample of 320 berries. Check the conditions and construct a 95% confidence interval for the proportion of all berries in the harvest that are too small to sell. Round to 3 decimal places.

   Grading a random sample of strawberries

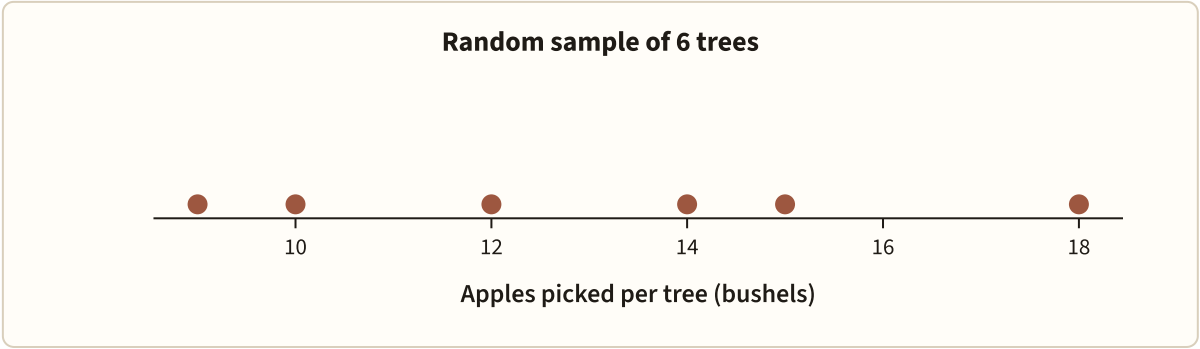
Grade	Number of berries
Marketable	272
Too small to sell	48
Total	320

6. **5 PTS** A farmer inspects a random sample of 60 ears of corn from a large field and finds 4 with earworm damage. Is the Large Counts condition met for a confidence interval for the proportion of damaged ears? Explain.
7. **10 PTS** A rural county has about 8,000 households. In a random sample of 180 households, 117 heat their homes mainly with wood. (a) Use the four-step process to construct a 90% confidence interval for the proportion of all households in the county that heat mainly with wood. Round to 3 decimal places. (b) Does the interval suggest that a majority of the county's households heat mainly with wood?

8. **5 PTS** Find t^* for (a) a 90% confidence interval from a sample of 10 and (b) a 99% confidence interval from a sample of 20.

9. **6 PTS** Explain why a confidence interval for a mean uses t^* instead of z^* when the population standard deviation is unknown, and describe how the t distribution with 4 degrees of freedom differs from the standard normal curve.

10. **10 PTS** An orchard has 300 apple trees. The owner records the bushels picked from a random sample of 6 trees: 12, 15, 9, 14, 10, 18 (dotplot shown). (a) Find $\bar{x}$ and s (1 decimal place). (b) Check the conditions and construct a 95% confidence interval for the mean bushels per tree for all 300 trees. Round to 1 decimal place.



11. **10 PTS** A large duck farm weighs a random sample of 20 eggs; a dotplot is roughly symmetric with no outliers. The sample has $\bar{x} = 69.3$ grams and $s = 3.1$ grams. (a) Use the four-step process to construct a 99% confidence interval for the mean weight of all the farm's duck eggs. Round to 2 decimal places. (b) A buyer claims the farm's eggs average 72 grams. Does the interval give reason to doubt this claim?

12. **6 PTS** After seeing the interval in Problem 11, a worker says, "So 99% of our duck eggs weigh between 67.32 and 71.28 grams." Explain what is wrong with this statement.

13. **8 PTS** A county historical society wants to estimate the proportion of the county's barns that were built before 1950, within 0.03 at 90% confidence. They have no prior estimate. How many barns must they visit in a random sample?

14. **10 PTS** A county extension office wants to estimate the mean soybean yield on the county's farms within 1.5 bushels per acre at 95% confidence. Past surveys suggest $\sigma \approx 5$ bushels per acre. (a) How many fields should be in the random sample? (b) To save driving time, an agent suggests measuring only fields that are next to the main highway. Explain why a margin of error of 1.5 bushels per acre would not protect the estimate from this problem.

UNIT 10 TEST: ANSWER KEY

1. Point estimate $\dfrac{11.6 + 13.0}{2} = 12.3$ inches; margin of error $\dfrac{13.0 - 11.6}{2} = 0.7$ inch.
2. (a) $z^* = 2.576$. (b) To capture the middle 99% of the normal curve instead of 95%, we must go farther out from the center, leaving only 0.5% in each tail; a wider net catches the parameter more often.
3. (a) We are 90% confident that the mean syrup yield per tap for all taps in the sugarbush this season is between 0.21 and 0.27 gallons. (b) If the family took many random samples of taps and built a 90% interval from each, about 90% of those intervals would capture the true mean yield per tap.
4. $\dfrac{\sigma}{\sqrt{n}} = \dfrac{0.08}{4} = 0.02$; $ME = 1.96 \times 0.02 = 0.0392$; interval 6.42 ± 0.0392: 6.381 to 6.459 magnitudes.
5. Random sample; $320 \le 5{,}000$; 48 and 272 are both at least 10. $\hat{p} = \dfrac{48}{320} = 0.15$; $SE = \sqrt{\dfrac{0.15 \times 0.85}{320}} = \sqrt{0.000398438} \approx 0.019961$; $ME = 1.96 \times 0.019961 \approx 0.0391$; interval 0.111 to 0.189.
6. No. There are only 4 successes (damaged ears), fewer than 10, so the sampling distribution of $\hat{p}$ may not be close to normal.
7. (a) State: p = proportion of all households in the county that heat mainly with wood; 90%. Plan: one-sample z interval for p; random sample; $180 \le 800$; 117 and 63 at least 10. Do: $\hat{p} = 0.65$; $SE = \sqrt{\dfrac{0.65 \times 0.35}{180}} \approx 0.035551$; $ME = 1.645 \times 0.035551 \approx 0.0585$; interval 0.592 to 0.708. Conclude: We are 90% confident that between 59.2% and 70.8% of the county's households heat mainly with wood. (b) Yes; the whole interval is above 0.5.
8. (a) $df = 9$: $t^* = 1.833$. (b) $df = 19$: $t^* = 2.861$.
9. When s replaces σ, both $\bar{x}$ and s vary from sample to sample, so the standardized statistic is more variable than z; using z^* would give intervals that capture μ less often than claimed. The t distribution with 4 df is also symmetric and bell-shaped around 0, but it has heavier tails (more area far from 0) and a lower peak, so its critical values are larger.
10. (a) $\bar{x} = \dfrac{78}{6} = 13.0$; squared deviations $1 + 4 + 16 + 1 + 9 + 25 = 56$; $s = \sqrt{\dfrac{56}{5}} = \sqrt{11.2} \approx 3.3$ bushels. (b) Random sample; $6 \le 30$; the dotplot shows no strong skewness or outliers. $df = 5$, $t^* = 2.571$; $SE = \sqrt{\dfrac{11.2}{6}} \approx 1.36626$; $ME \approx 3.513$; interval 9.5 to 16.5 bushels per tree.

11. (a) State: μ = mean weight of all eggs on the farm; 99%. Plan: one-sample t interval; random sample; 20 is far less than 10% of the farm's eggs; the dotplot shows no strong skewness or outliers. Do: $SE = \dfrac{3.1}{\sqrt{20}} \approx 0.693181$; $df = 19$, $t^* = 2.861$; $ME \approx 1.9832$; interval 67.32 to 71.28 grams. Conclude: We are 99% confident that the mean weight of all the farm's duck eggs is between 67.32 and 71.28 grams. (b) Yes; 72 grams lies above the whole interval, so it is not a plausible value for the mean.
12. The interval estimates the mean weight of all the eggs, not the weights of individual eggs. Individual eggs vary much more than the mean does (the sample standard deviation is 3.1 grams), so many eggs weigh less than 67.32 or more than 71.28 grams. The 99% describes how often the method captures the true mean, not a percent of eggs.
13. Use $p^* = 0.5$: $n = \left(\dfrac{1.645}{0.03}\right)^2 (0.25) = (54.8333)^2(0.25) \approx 3006.69 \times 0.25 \approx 751.67$; round up to 752 barns.
14. (a) $n = \left(\dfrac{1.96 \times 5}{1.5}\right)^2 = (6.5333)^2 \approx 42.68$; round up to 43 fields. (b) The margin of error covers only random sampling variation. Fields along the highway may differ systematically from other fields (soil, drainage, how they are farmed), so the sample would be biased, and that bias is not included in the margin of error no matter how many fields are measured.

UNIT 11

Significance Tests

In Unit 10 you learned to estimate a parameter with a confidence interval. This unit asks a different question: does the evidence support a claim? A seed company says 90% of its seeds germinate; a beekeeper wonders whether her machine is underfilling jars; a rancher asks whether a new feed changes how fast calves grow. A significance test answers by assuming the claim (the null hypothesis) is true and asking whether the observed result could easily happen by chance alone. The P-value measures how surprising the data would be, and students learn to read it carefully: what it means, what it does not mean, and why we "fail to reject" rather than "accept" a null hypothesis.

Students first meet the logic through simulation, then carry out the one-sample z test for a proportion and the one-sample t test for a mean, using the same four steps (State, Plan, Do, Conclude) as in Unit 10. They then study the two ways a test can go wrong (Type I and Type II errors), the power of a test to detect a real effect, and how tests and confidence intervals fit together. The unit closes with cautions every honest user of statistics needs: statistical significance is not practical importance, many tests produce false alarms, and no formula can rescue badly collected data.

Everything here rests on the sampling distributions of Unit 9 and the intervals of Unit 10. Unit 12 extends these tests to comparisons of two groups, and Unit 13 adds chi-square tests for categorical data.

PREREQUISITES

- Find areas under the standard normal curve with the table or technology (Lesson 3.4).
- Use the sampling distributions of $\hat{p}$ and $\bar{x}$: their means, standard deviations and the conditions for approximate normality (Lessons 9.2–9.4).
- Read the t table and explain why t is used when σ is unknown (Lesson 10.3).
- Construct and interpret confidence intervals for a proportion and a mean, checking the Random, 10% and Large Counts or Normal/Large Sample conditions (Lessons 10.2 and 10.4).
- Use the complement rule and binomial probabilities (Lesson 6.4 and Unit 8).

MASTERY CHECKLIST

☐ I can explain the reasoning of a significance test: assume H_0 is true and ask whether the result could easily happen by chance alone.
☐ I can state null and alternative hypotheses about a parameter, one-sided or two-sided, in symbols and in words.
☐ I can estimate a P-value from a simulation and interpret any P-value correctly in context.
☐ I can use a P-value and a significance level to reject or fail to reject H_0, and explain why we never accept H_0.
☐ I can carry out a one-sample z test for a proportion, using p_0 in the conditions and the standard deviation.

- Identify voluntary response samples, convenience samples and confounding (Unit 5).

☐ I can carry out a one-sample t test for a mean, finding the P-value with technology and bounding it with the t table.
☐ I can describe Type I and Type II errors and their consequences in context, and explain why $P(\text{Type I error}) = \alpha$.
☐ I can compute α, power and β for a stated rejection rule in a simple normal setting, and explain what increases power.
☐ I can use a confidence interval to decide a two-sided test and explain why the two agree.
☐ I can explain the difference between statistical significance and practical importance, the danger of running many tests, and why inference cannot fix badly collected data.

LESSON 11.1 · 2 SESSIONS OF 45 MINUTES

The Logic of Significance Tests

OBJECTIVES

- Explain the reasoning of a significance test: assume the null hypothesis is true and ask whether the observed result could easily happen by chance alone.
- State null and alternative hypotheses about a parameter, one-sided or two-sided, in symbols and in words.
- Estimate a P-value from a simulation and interpret it in context.
- Use a P-value and a significance level to reject or fail to reject the null hypothesis, and explain why we never accept the null hypothesis.

VIDEO

WATCH

An Introduction to Hypothesis Testing

jbstatistics · 9:54

When to watch: Before Example 1: null and alternative hypotheses and the logic of a test.

https://www.youtube.com/watch?v=tTeMYuS87oU

BACKUP

What is a p-value? (Updated and extended version)

jbstatistics · 10:50

When to watch: Before Example 2: what a P-value is.

https://www.youtube.com/watch?v=UsU-O2Z1rAs

KEY TERMS

significance test	A method that uses data from a sample or experiment to weigh the evidence against a claim about a population parameter.
null hypothesis (H_0)	The claim being tested, usually a statement of "no effect," "no difference" or "nothing has changed." It gives the parameter one specific value, such as $H_0 : p = 0.5$.
alternative hypothesis (H_a)	The claim we are looking for evidence for. It is one-sided if it says the parameter is greater than (or less than) the null value, and two-sided if it says only that the parameter is different ($\neq$).

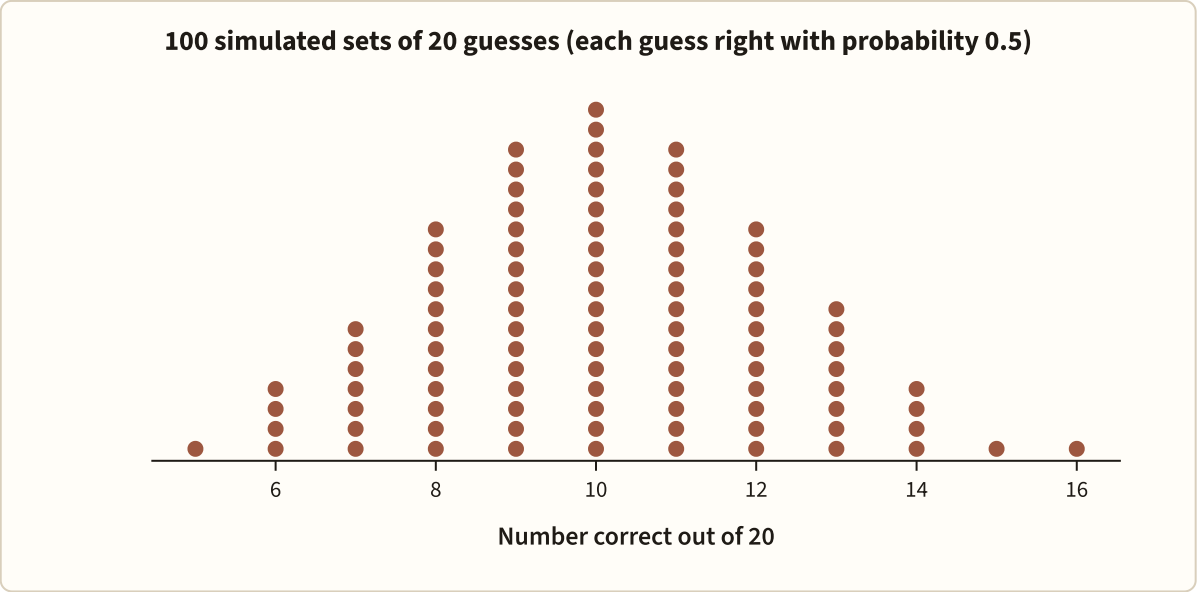
P-value	The probability, computed assuming H_0 is true, of getting a result at least as extreme as the one actually observed (extreme in the direction of H_a).
significance level (α)	A cutoff chosen before the data are collected, often 0.05. If the P-value is less than α, we reject H_0.
statistically significant	Describes a result whose P-value is less than α: a result too unlikely to be explained by chance alone if H_0 were true.

CORE IDEAS

Henry claims he can taste the difference between eggs from his family's hens and eggs from the store. To test him, his sister Ruth cooks one of each, flips a coin to decide which goes on the left, and asks Henry to pick the farm egg. They repeat this 20 times, and Henry is right 15 times. Is that impressive? Even someone who is only guessing gets about half right, and sometimes a guesser gets lucky. The question at the heart of every significance test is: **could a result like this easily happen by chance alone?**

We turn the question into two competing claims about a parameter. Let p be the proportion of all such trials in which Henry would pick the farm egg. The **null hypothesis** says nothing special is going on: $H_0 : p = 0.5$ (Henry is guessing). The **alternative hypothesis** says what we suspect: $H_a : p > 0.5$ (Henry does better than guessing). Hypotheses are always about parameters such as p or μ, never about statistics like $\hat{p} = 15/20$, because we already know the statistic; it is the population truth we are unsure of. Henry's claim is one-sided (only "better than guessing" counts as support). If we cared about a difference in either direction we would write a two-sided alternative with $\neq$. Choose the alternative from the question being asked, before looking at the data.

Next we pretend H_0 is true and ask what chance alone produces. A computer simulated 100 sets of 20 pure guesses, each guess right with probability 0.5. The dotplot shows the number correct in each set. Most guessers get between 8 and 12 right. Only 2 of the 100 simulated guessers got 15 or more right.



That fraction is an estimate of the **P-value**: the probability, assuming H_0 is true, of a result at least as extreme as the one observed. Here P-value $\approx 2/100 = 0.02$. In words: if Henry were only guessing, he would get 15 or more right in only about 2% of sets of 20 trials. A small P-value leaves two possibilities: either H_0 is true and something quite rare happened, or H_0 is false. The smaller the P-value, the harder it is to believe the first explanation, so **small P-values are evidence against** H_0. A large P-value says the result is the kind of thing chance produces often, so it gives no real evidence against H_0.

How small is small enough? Before collecting data we choose a **significance level** α, most often 0.05 (sometimes 0.10 or 0.01). If the P-value is less than α, we **reject** H_0 and say the result is statistically significant. Otherwise we **fail to reject** H_0. For Henry, $0.02 < 0.05$, so we reject H_0: there is convincing evidence that Henry can do better than guessing.

Why do we say "fail to reject" and never "accept H_0"? A significance test works like a courtroom. The defendant (H_0) is presumed innocent, and the prosecution must bring convincing evidence. A verdict of "not guilty" means the evidence was not strong enough, not that innocence was proved. In the same way, a large P-value means the data are consistent with H_0, but they may also be consistent with many other values of the parameter, especially when the sample is small. If Henry had gotten 12 of 20 right, that would not prove he is guessing; it would only mean 20 trials did not give convincing evidence that he isn't.

Two cautions about the P-value. First, it is computed assuming H_0 is true, so it is NOT the probability that H_0 is true. A P-value of 0.02 does not mean "a 2% chance Henry is guessing." Second, the conclusion should always be written in context, about the parameter, and never with the word "proves."

KEY FORMULAS

$$H_0 : \text{parameter} = \text{null value} \qquad H_a : \text{parameter} > , \ < , \text{ or } \neq \text{ null value}$$

$$\text{P-value} = P(\text{a result at least as extreme as the one observed} \mid H_0 \text{ is true})$$

$$\text{P-value} < \alpha \ \Rightarrow \ \text{reject } H_0 \qquad \text{P-value} \geq \alpha \ \Rightarrow \ \text{fail to reject } H_0$$

From history. In 1710 John Arbuthnot, a Scottish physician and mathematician, examined 82 years of London birth records and found that more boys than girls had been born in every single year. He argued that if a boy-heavy year and a girl-heavy year were equally likely, a run of 82 boy-heavy years would have probability $(1/2)^{82}$, far too small to blame on chance. His argument is often called one of the earliest significance tests.

Practice habit. Write the hypotheses in symbols AND words, and write them before you look at the results. Define the parameter in context ("p = the proportion of all ...").

WORKED EXAMPLES

Example 1. State the null and alternative hypotheses in symbols and in words. (a) A seed company claims that 85% of its pepper seeds germinate. Ruth suspects the germination rate is lower. (b) A sawmill's boards are supposed to have a mean length of 96 inches. The foreman takes a sample of boards to check whether the saw needs adjusting in either direction.

1. (a) The parameter is a proportion. Let p = the proportion of all of this company's pepper seeds that germinate.
2. The claim being tested is the company's: $H_0 : p = 0.85$ (the germination rate is 85%).
3. Ruth suspects a lower rate, so the alternative is one-sided: $H_a : p < 0.85$ (the germination rate is less than 85%).
4. (b) The parameter is a mean. Let μ = the mean length, in inches, of all boards the saw is now cutting.
5. $H_0 : \mu = 96$ (the saw is cutting correctly, on average).
6. Boards that are too long or too short both mean trouble, so the alternative is two-sided: $H_a : \mu \neq 96$ (the mean length is different from 96 inches).
7. Notice that both hypotheses are about the parameter (p or μ), not about a sample result, and the null hypothesis always contains the equals sign.

Answer: (a) $H_0 : p = 0.85$, $H_a : p < 0.85$, where p is the proportion of all the company's pepper seeds that germinate. (b) $H_0 : \mu = 96$, $H_a : \mu \neq 96$, where μ is the mean length (inches) of all boards the saw is cutting.

Example 2. Grace's coach believes she makes 80% of her free throws in the long run. During one week of practice Grace makes only 34 of 50 shots. Is this convincing evidence that her true free-throw percentage is less than 80%? The table shows the results of 200 simulated sets of 50 shots by a shooter who makes each shot with probability 0.80. Use $\alpha = 0.05$.

200 simulated sets of 50 free throws by a shooter who makes 80%

Shots made (out of 50)	Number of simulated sets
33 or fewer	3
34	3
35	9
36	14
37	19
38	23
39	25
40	26
41	24
42	20
43	15
44	10
45 or more	9
Total	200

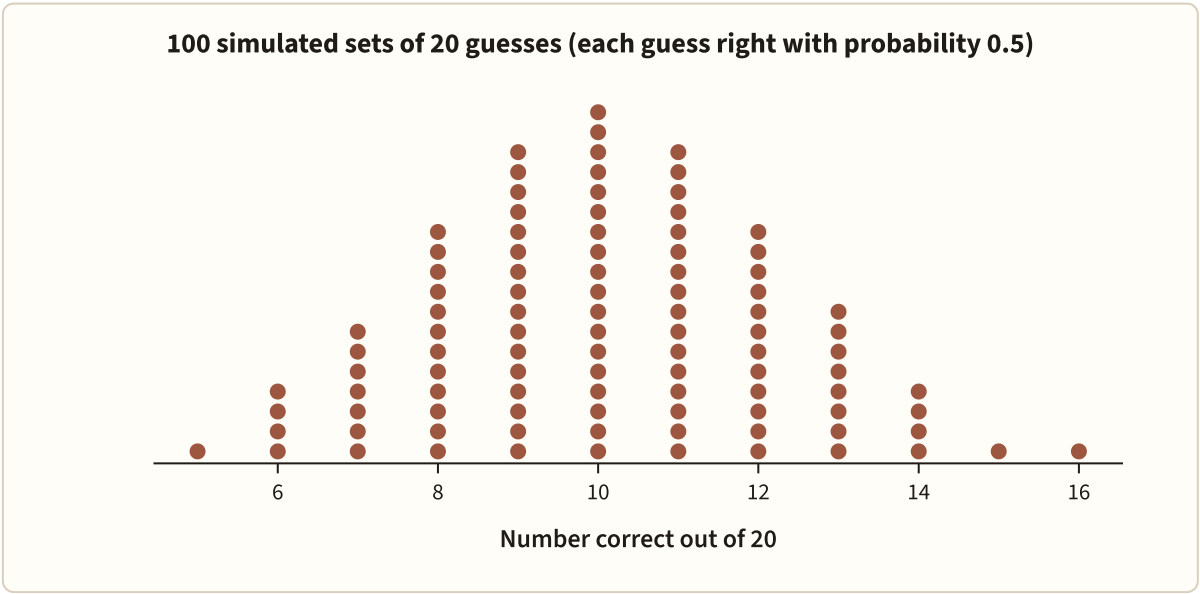
1. Parameter: p = Grace's long-run proportion of free throws made.
2. Hypotheses: $H_0 : p = 0.80$ (she makes 80%) and $H_a : p < 0.80$ (she makes less than 80%).
3. Observed result: $\hat{p} = \dfrac{34}{50} = 0.68$. A shooter who truly makes 80% would expect about $0.80 \times 50 = 40$ makes.
4. "At least as extreme" in the direction of H_a means 34 or fewer makes. From the table: $3 + 3 = 6$ of the 200 simulated sets.
5. Estimated P-value $\approx \dfrac{6}{200} = 0.03$.
6. Interpret: if Grace truly made 80% of her free throws, she would make 34 or fewer of 50 in only about 3% of weeks like this.
7. Decide: $0.03 < 0.05$, so we reject H_0.

8. Conclude: there is convincing evidence that Grace's true free-throw percentage is less than 80%.

Answer: P-value $\approx 6/200 = 0.03$. Since $0.03 < 0.05$, reject H_0: there is convincing evidence that Grace makes less than 80% of her free throws in the long run.

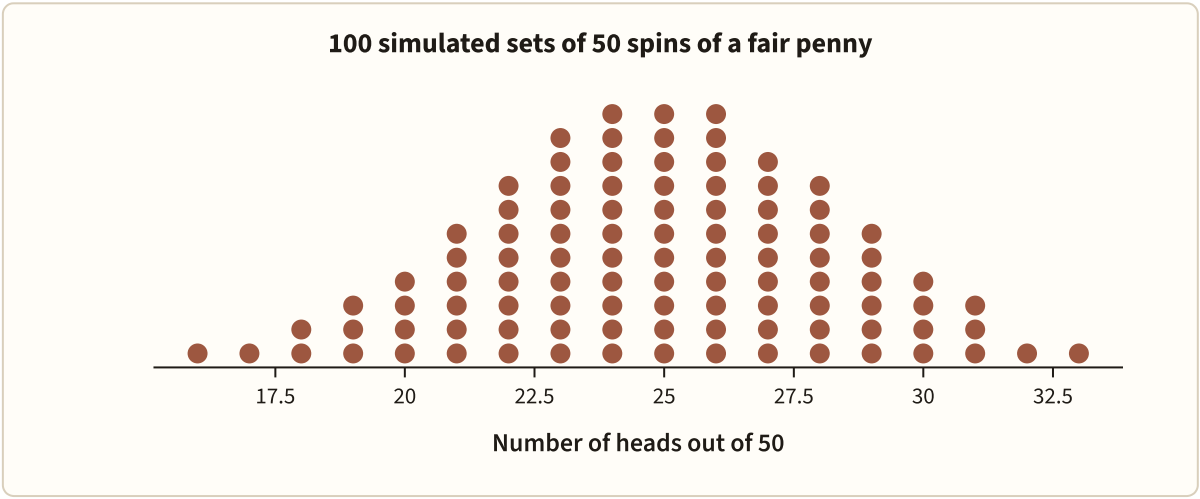
PRACTICE

1. EASY A bakery's sourdough loaves are supposed to weigh 24 ounces on average. The baker suspects the new oven is producing loaves that are too light. Define the parameter and state H_0 and H_a in symbols and words.
2. EASY A test of $H_0 : p = 0.40$ against $H_a : p > 0.40$ gives a P-value of 0.012. At the $\alpha = 0.05$ level, do you reject or fail to reject H_0?
3. EASY A test of $H_0 : \mu = 3.5$ against $H_a : \mu \neq 3.5$ gives a P-value of 0.21. At the $\alpha = 0.05$ level, do you reject or fail to reject H_0?
4. EASY In a random sample of 50 apples from the Miller orchard, 62% had no blemishes. A student writes the hypotheses $H_0 : \hat{p} = 0.62$ and $H_a : \hat{p} > 0.62$. Explain what is wrong.
5. MEDIUM Use the dotplot of 100 simulated guessers (the egg-tasting test). Suppose Henry had gotten 14 of 20 right instead of 15. (a) Estimate the P-value for $H_0 : p = 0.5$ against $H_a : p > 0.5$. (b) At $\alpha = 0.05$, do you reject or fail to reject H_0?



6. MEDIUM A gardener tests whether a new fertilizer raises the mean tomato yield above the old mean of 4.2 pounds per plant, with $H_0 : \mu = 4.2$ and $H_a : \mu > 4.2$. The P-value is 0.03. Interpret this P-value in context.

7. MEDIUM Samuel spins a penny on the kitchen table 50 times and gets 18 heads. He wonders whether a spun penny lands heads half the time. The dotplot shows 100 simulated sets of 50 spins of a penny with $P(\text{heads}) = 0.5$. (a) State the hypotheses. (b) A result "at least as extreme" as 18 heads means 18 or fewer OR 32 or more (7 or more away from 25). Estimate the P-value. (c) At $\alpha = 0.05$, do you reject or fail to reject H_0?



8. MEDIUM In Problem 7, Samuel writes, "Since I failed to reject H_0, I have shown that a spun penny lands heads exactly half the time." Explain what is wrong with his conclusion.

9. MEDIUM Two tests of the same null hypothesis give P-values of 0.04 and 0.004. Which P-value gives stronger evidence against H_0?

10. HARD A test of whether a new cold remedy shortens colds gives a P-value of 0.03. A student says, "There is a 3% chance that the remedy does nothing." Explain why this interpretation is wrong, and give a correct one.

11. **HARD** Use the table of 200 simulated sets of 50 free throws by an 80% shooter. Another player, Ruth, has a long-run rate of 80%. After a summer of practice she makes 44 of 50, and her coach wonders whether she has improved. (a) Estimate the P-value for $H_0 : p = 0.80$ against $H_a : p > 0.80$. (b) At $\alpha = 0.05$, what is your decision?

200 simulated sets of 50 free throws by a shooter who makes 80%

Shots made (out of 50)	Number of simulated sets
33 or fewer	3
34	3
35	9
36	14
37	19
38	23
39	25
40	26
41	24
42	20
43	15
44	10
45 or more	9
Total	200

12. **HARD** Arbuthnot found more boys than girls born in London in each of 82 years. Suppose each year were equally likely to have more boys or more girls, independently of other years. (a) Find the probability that all 82 years would have more boys. Give your answer in scientific notation to 3 significant figures. (b) What conclusion did Arbuthnot draw, and why?

13. **CHALLENGE** Ruth suspects that a board-game die lands on 6 more often than it should. She rolls it 30 times and gets 9 sixes. (a) Describe how to simulate one set of 30 rolls of a fair die using a random number generator. (b) In 100 simulated sets, 4 had 9 or more sixes. Estimate the P-value. (c) At $\alpha = 0.05$, what do you conclude?

COMMON MISTAKES

- Writing hypotheses about a statistic ($\hat{p}$ or $\bar{x}$). Hypotheses are always about a parameter (p or μ), and you should define that parameter in words.

- Saying "we accept H_0" or "the data prove H_0." A large P-value means only that the evidence is not convincing; say "we fail to reject H_0" and "there is not convincing evidence that ..."
- Reading the P-value as the probability that H_0 is true. The P-value assumes H_0 is true; it is the probability of data as extreme as ours under that assumption.
- Choosing a one-sided alternative after seeing which way the data went. Decide on H_a from the question being asked, before looking at the results.

ANSWERS TO PRACTICE

1. Let μ = the mean weight (ounces) of all loaves from the new oven. $H_0 : \mu = 24$ (the loaves average 24 ounces); $H_a : \mu < 24$ (the loaves average less than 24 ounces).
2. Reject H_0, because $0.012 < 0.05$.
3. Fail to reject H_0, because $0.21 \geq 0.05$.
4. Hypotheses must be about the parameter p (the proportion of ALL apples in the orchard with no blemishes), not about the statistic $\hat{p}$. We already know $\hat{p} = 0.62$ for this sample; the test is about the unknown population proportion, and the null value should come from a claim or past standard, not from the sample itself.
5. (a) Results of 14 or more: $4 + 1 + 1 = 6$ of 100, so P-value ≈ 0.06. (b) $0.06 \geq 0.05$, so fail to reject H_0; 14 of 20 is not convincing evidence that Henry does better than guessing.
6. If the new fertilizer really produced a mean yield of 4.2 pounds per plant (no improvement), there would be only about a 3% chance of getting a sample mean yield as high as or higher than the one the gardener observed.
7. (a) $H_0 : p = 0.5$, $H_a : p \neq 0.5$, where p is the long-run proportion of spins that land heads. (b) 18 or fewer: $1 + 1 + 2 = 4$; 32 or more: $1 + 1 = 2$; P-value $\approx 6/100 = 0.06$. (c) $0.06 \geq 0.05$, so fail to reject H_0: there is not convincing evidence that a spun penny lands heads in a proportion different from 0.5.
8. Failing to reject H_0 means only that 50 spins did not give convincing evidence against $p = 0.5$. The data are also consistent with other values (for example $p = 0.4$), so they do not show that p is exactly 0.5. We never accept H_0; with more spins a real difference might show up.
9. 0.004. The smaller the P-value, the less likely such a result would be by chance alone if H_0 were true, so the stronger the evidence against H_0.
10. The P-value is computed by ASSUMING H_0 is true (the remedy does nothing), so it cannot be the probability that H_0 is true. Correct: if the remedy really did nothing, there would be only a 3% chance of seeing colds shortened by as much as (or more than) in this study, just by chance.
11. (a) 44 or more: $10 + 9 = 19$ of 200, so P-value $\approx 19/200 = 0.095$. (b) $0.095 \geq 0.05$, so fail to reject H_0: there is not convincing evidence that Ruth's long-run rate is now above 80%.

12. (a) $(0.5)^{82} \approx 2.07 \times 10^{-25}$. (b) This probability (in effect a P-value) is so tiny that chance alone could not reasonably explain the pattern, so he rejected the idea that a boy-heavy year and a girl-heavy year were equally likely: births favor boys slightly.
13. (a) Generate 30 random whole numbers from 1 to 6 (each equally likely), let 6 stand for a six, and count the sixes; repeat for many sets, recording the count each time. (b) P-value $\approx 4/100 = 0.04$. (c) $0.04 < 0.05$, so reject $H_0 : p = 1/6$ in favor of $H_a : p > 1/6$: there is convincing evidence that this die lands on 6 more than one-sixth of the time.

LESSON 11.2 · 2 SESSIONS OF 45 MINUTES

Testing a Claim about a Proportion

OBJECTIVES

- Check the Random, 10% and Large Counts conditions for a one-sample z test for a proportion, using p_0.
- Compute the test statistic $z = \dfrac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}$ and explain why p_0, not $\hat{p}$, is used in the standard deviation.
- Find one-sided and two-sided P-values from the standard normal distribution.
- Carry out the four-step process (State, Plan, Do, Conclude) and write a conclusion in context.

VIDEO

WATCH

Example of 1 Sample Proportion Z-test

ProfRobBob · 16:35

When to watch: Before Example 1: a one-proportion z test, worked through.

https://www.youtube.com/watch?v=eqMfOtCMs8Q

BACKUP

Inference for a Proportion: An Example of a Confidence Interval and a Hypothesis Test

jbstatistics · 8:40

When to watch: Review: an interval and a test for one proportion.

https://www.youtube.com/watch?v=M7fUzmSbXWI

KEY TERMS

one-sample z test for a proportion	A significance test of $H_0 : p = p_0$ that uses the standard normal distribution to find the P-value for an observed sample proportion $\hat{p}$.
null value (p_0)	The specific value of the proportion stated in the null hypothesis, such as 0.90 in $H_0 : p = 0.90$.
test statistic	A number that measures how far the sample statistic is from the null value, in standard deviations: (statistic − null value) ÷ (standard deviation of the statistic).

Large Counts condition (for a test)	The check that $np_0 \geq 10$ and $n(1-p_0) \geq 10$, so that the sampling distribution of $\hat{p}$ is approximately normal when H_0 is true.
two-sided P-value	For $H_a : p \neq p_0$, the probability of a result at least as far from p_0 as the observed one in EITHER direction; it is twice the area in one tail.

CORE IDEAS

In Lesson 11.1 we found P-values by simulation. For a proportion we can do better. Unit 9 told us what the sample proportion $\hat{p}$ does over many random samples: its mean is p, its standard deviation is $\sqrt{p(1-p)/n}$, and its shape is approximately normal when the expected counts are large. If $H_0 : p = p_0$ is true, we can put p_0 in place of p and know the whole sampling distribution, so the P-value becomes a normal-curve area.

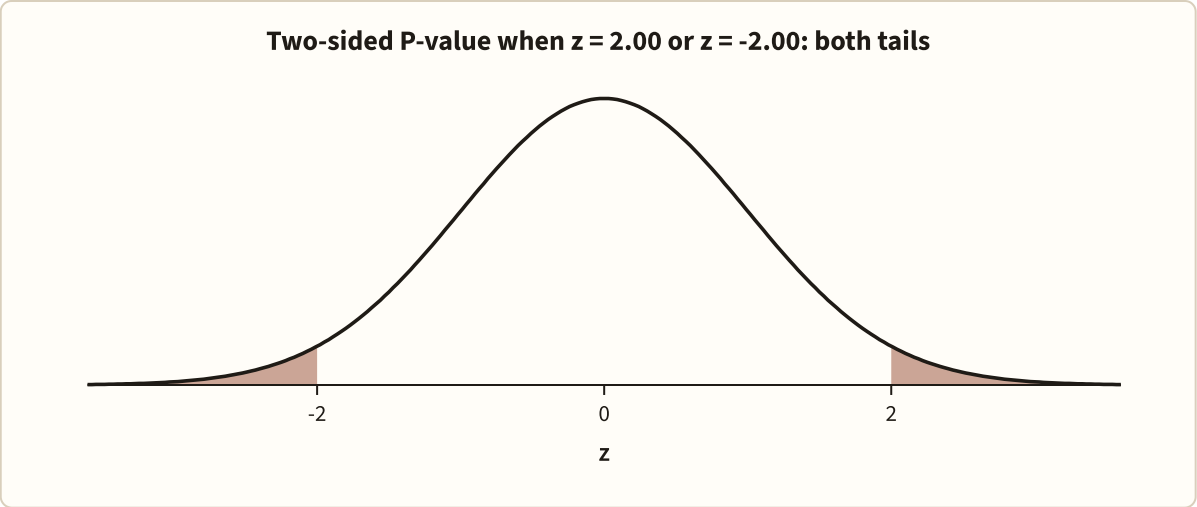
To find that area we standardize, just as with z-scores in Unit 3. The **test statistic** is

$$z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$$

It tells how many standard deviations the observed $\hat{p}$ lies from the value claimed by H_0. A seed sample with $z = -2.00$ has a germination rate two standard deviations below the company's claim, the kind of shortfall chance rarely produces.

Why p_0 and not $\hat{p}$ in the standard deviation? A P-value is calculated assuming H_0 is true. If H_0 is true, the population proportion is exactly p_0, so the standard deviation of $\hat{p}$ is exactly $\sqrt{p_0(1-p_0)/n}$; there is nothing to estimate. (In a confidence interval in Unit 10 we did not know p, so we had to use $\hat{p}$.) For the same reason the Large Counts condition uses the expected counts under H_0: $np_0 \geq 10$ and $n(1-p_0) \geq 10$.

The alternative hypothesis tells which area is the P-value. For $H_a : p > p_0$ it is the area to the right of z; for $H_a : p < p_0$, the area to the left. For $H_a : p \neq p_0$, a result counts as "at least as extreme" if it is as far from p_0 in either direction, so the P-value is the area in both tails: twice the one-tail area.



Before calculating, check three conditions. **Random:** the data come from a random sample (or a randomized experiment), so chance, not our choices, produced them. **10%:** when sampling without replacement, n is at most 10% of the population, so the observations are nearly independent and the standard deviation formula is right. **Large Counts:** $np_0 \geq 10$ and $n(1-p_0) \geq 10$, so the normal curve fits the sampling distribution. If a condition fails, the P-value from the normal table may be badly wrong.

Every test in this course is written in four steps. **State:** the hypotheses (in symbols and words), the significance level α, and the parameter in context. **Plan:** name the procedure ("one-sample z test for p ") and check the conditions. **Do:** compute $\hat{p}$, the test statistic and the P-value, with a sketch of the shaded curve. **Conclude:** compare the P-value with α, make the decision, and state the conclusion in context: "there is (or is not) convincing evidence that ..."

KEY FORMULAS

$$z = \frac{\hat{p} - p_0}{\sqrt{\dfrac{p_0(1-p_0)}{n}}}$$

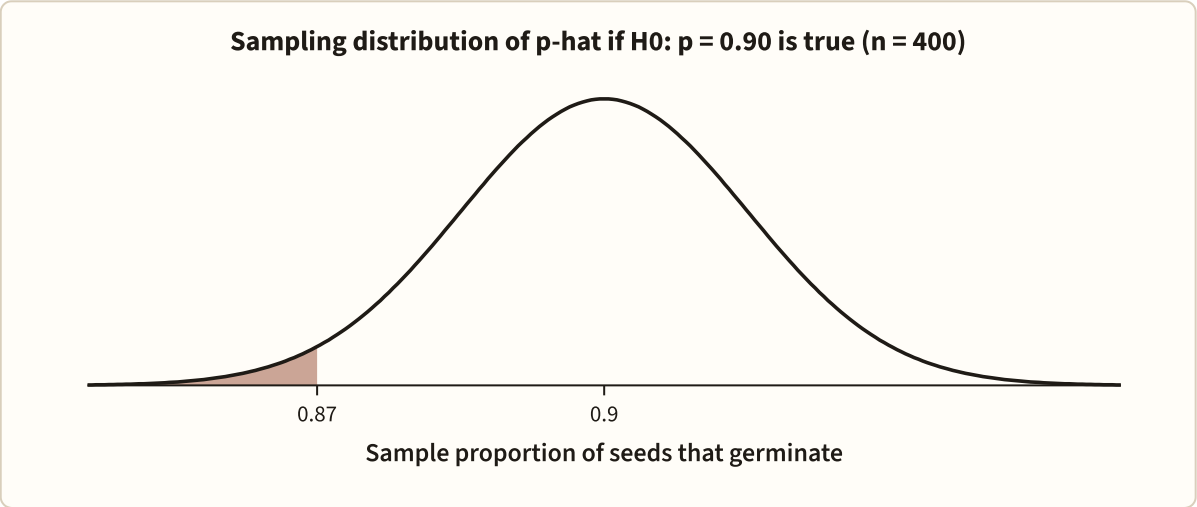
$$\text{test statistic} = \frac{\text{statistic} - \text{null value}}{\text{standard deviation of the statistic}}$$

Large Counts: $np_0 \geq 10$ and $n(1-p_0) \geq 10$ $\qquad$ 10%: $n \leq 0.10 \times$ population size

$H_a : p > p_0 : P(Z \geq z)$ $\qquad$ $H_a : p < p_0 : P(Z \leq z)$ $\qquad$ $H_a : p \neq p_0 : 2P(Z \geq |z|)$

Practice habit. Check the conditions before calculating, and use p_0 (not $\hat{p}$) both in the Large Counts check and in the standard deviation. Always sketch the normal curve and shade the P-value area.

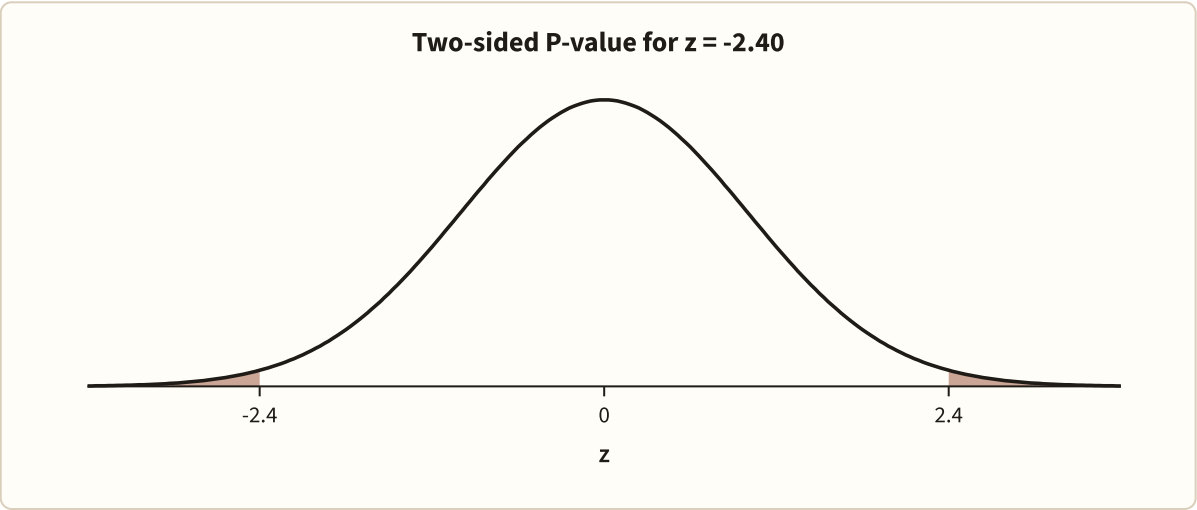
Example 1. A seed company claims that 90% of its tomato seeds germinate. A garden club plants a random sample of 400 of the company's seeds, and 348 germinate. Is there convincing evidence at the $\alpha = 0.05$ level that the true germination rate is less than 90%?



1. State: $H_0 : p = 0.90$ and $H_a : p < 0.90$, where p is the proportion of all the company's tomato seeds that germinate. Use $\alpha = 0.05$.
2. Plan: one-sample z test for p. Random: the seeds are a random sample. 10%: 400 seeds is far less than 10% of all the company's seeds. Large Counts: $400(0.90) = 360 \geq 10$ and $400(0.10) = 40 \geq 10$.
3. Do: $\hat{p} = \dfrac{348}{400} = 0.87$.
4. Standard deviation if H_0 is true: $\sqrt{\dfrac{0.90(0.10)}{400}} = \sqrt{0.000225} = 0.015$.
5. Test statistic: $z = \dfrac{0.87 - 0.90}{0.015} = -2.00$. The sample rate is 2 standard deviations below the claim.
6. P-value (area to the left, because H_a says "less than"): $P(Z \leq -2.00) = 0.0228$.
7. Conclude: $0.0228 < 0.05$, so reject H_0. There is convincing evidence that fewer than 90% of the company's tomato seeds germinate.

Answer: $z = -2.00$, P-value $= 0.0228$. Reject H_0: there is convincing evidence that the true germination rate is less than 90%.

Example 2. A genetic model predicts that 75% of the plants from a certain cross of pea plants will have purple flowers. Ruth grows a random sample of 300 plants from this cross, and 207 have purple flowers. Do the data give convincing evidence at the $\alpha = 0.05$ level that the true proportion differs from 0.75? Would the decision change at $\alpha = 0.01$?

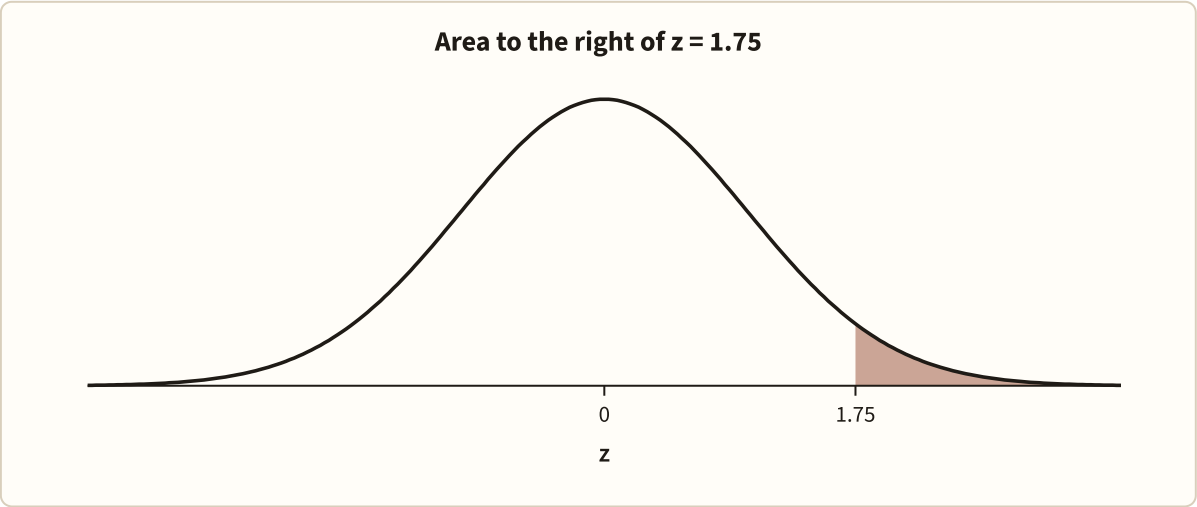


1. State: $H_0 : p = 0.75$ and $H_a : p \neq 0.75$, where p is the proportion of all plants from this cross that have purple flowers. Use $\alpha = 0.05$. The model could be wrong in either direction, so H_a is two-sided.
2. Plan: one-sample z test for p. Random: random sample of plants. 10%: 300 is less than 10% of all possible plants from this cross. Large Counts: $300(0.75) = 225 \geq 10$ and $300(0.25) = 75 \geq 10$.
3. Do: $\hat{p} = \dfrac{207}{300} = 0.69$.
4. Standard deviation if H_0 is true: $\sqrt{\dfrac{0.75(0.25)}{300}} = \sqrt{0.000625} = 0.025$.
5. $z = \dfrac{0.69 - 0.75}{0.025} = -2.40$.
6. Two-sided P-value: $2P(Z \leq -2.40) = 2(0.0082) = 0.0164$.
7. Conclude at $\alpha = 0.05$: $0.0164 < 0.05$, so reject H_0. There is convincing evidence that the proportion of purple-flowered plants from this cross is different from 0.75 (the sample suggests it is lower).
8. At $\alpha = 0.01$: $0.0164 \geq 0.01$, so we would fail to reject H_0. The stricter standard demands stronger evidence than these data give.

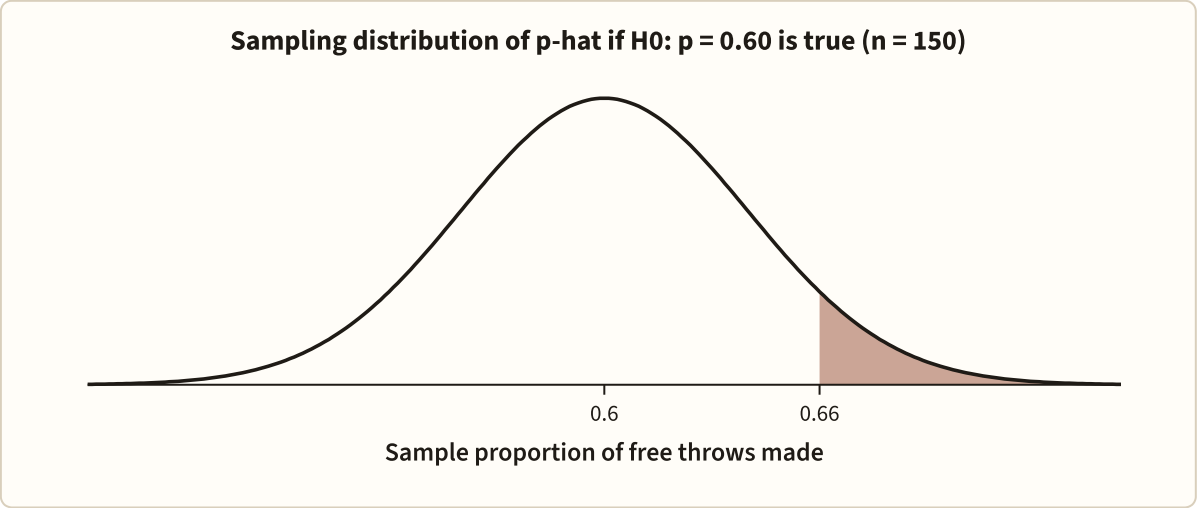
Answer: $z = -2.40$, P-value $= 0.0164$. At $\alpha = 0.05$ reject H_0 (convincing evidence that the true proportion is not 0.75); at $\alpha = 0.01$ fail to reject H_0.

PRACTICE

1. EASY For a test of $H_0 : p = 0.5$ with a sample of $n = 100$, find the standard deviation of $\hat{p}$ assuming H_0 is true.
2. EASY In past years, 20% of the Miller family's pumpkins weighed more than 20 pounds. This year, in a random sample of 400 pumpkins, 92 weighed more than 20 pounds. For $H_0 : p = 0.20$, find (a) the standard deviation of $\hat{p}$ assuming H_0 is true and (b) the test statistic z.
3. EASY A hatchery claims that only 4% of its chicken eggs fail to hatch. A farmer plans to test this claim with a random sample of 150 eggs. Is the Large Counts condition met?
4. EASY A test of $H_0 : p = 0.30$ against $H_a : p > 0.30$ gives $z = 1.75$. Find the P-value.



5. MEDIUM A test of $H_0 : p = 0.65$ against $H_a : p \neq 0.65$ gives $z = -1.20$. Find the P-value. Round to 4 decimal places.
6. MEDIUM Henry has made 60% of his free throws in past seasons. After a summer of practice, he makes 99 of 150 shots. Treat these shots as a random sample of his shooting. Test $H_0 : p = 0.60$ against $H_a : p > 0.60$ at the $\alpha = 0.05$ level. Give z, the P-value and your decision.



7. MEDIUM In a confidence interval for p we use $\sqrt{\hat{p}(1-\hat{p})/n}$, but in a significance test we use $\sqrt{p_0(1-p_0)/n}$. Explain why.
8. MEDIUM A parish librarian has records showing that 30% of the books checked out are children's books. This year, in a random sample of 525 checkouts, 168 were children's books. Test whether the proportion has changed, at the $\alpha = 0.05$ level. Give z, the P-value and your decision.
9. MEDIUM Interpret the P-value of 0.0668 from Problem 6 in context.
10. HARD With her old feed, 25% of a poultry farmer's eggs were graded extra-large. After switching to a new feed, she grades a random sample of 300 eggs and finds 93 extra-large. Is there convincing evidence at the $\alpha = 0.05$ level that the proportion of extra-large eggs has increased? Use the four-step process.
11. HARD A board-game spinner has four equal sections, one of them red. Grace suspects the spinner is not fair. She spins it 300 times and it lands on red 60 times. For $H_0 : p = 0.25$ against $H_a : p \neq 0.25$: (a) find z; (b) find the P-value (4 decimal places); (c) give the decision at $\alpha = 0.05$; (d) give the decision at $\alpha = 0.01$.
12. HARD An orchard owner wants to test the claim that 10% of her apples have blemishes, using a random sample of 40 apples. (a) Is the Large Counts condition met? (b) Explain why a P-value from the normal table would not be trustworthy here.
13. CHALLENGE A garden club wonders whether more than half of the plants of a new sunflower variety grow taller than 6 feet ($H_0 : p = 0.5$, $H_a : p > 0.5$). Club A grows a random sample of 100 plants and Club B grows a random sample of 400 plants; in both, 55% grow taller than 6 feet. (a) Find z for Club A. (b) Find Club A's P-value. (c) Find z for Club B. (d) Find Club B's P-value. (e) Explain why the same sample proportion gives such different evidence.

COMMON MISTAKES

- Using $\hat{p}$ in the standard deviation or the Large Counts check. In a test, assume H_0 is true and use p_0: $\sqrt{p_0(1-p_0)/n}$, np_0 and $n(1-p_0)$.
- Forgetting to double the tail area for a two-sided alternative. If H_a has $\neq$, the P-value is $2P(Z \geq |z|)$.
- Shading the wrong tail. Let H_a decide: "greater than" means the area to the right of z, "less than" the area to the left; a sketch prevents this mistake.
- Stopping at "reject H_0." Finish with a sentence in context about the parameter: "there is convincing evidence that the true germination rate is less than 90%."

ANSWERS TO PRACTICE

1. $\sqrt{\dfrac{0.5(0.5)}{100}} = \sqrt{0.0025} = 0.05$
2. (a) $\sqrt{\dfrac{0.20(0.80)}{400}} = \sqrt{0.0004} = 0.02$. (b) $\hat{p} = 92/400 = 0.23$, so $z = \dfrac{0.23 - 0.20}{0.02} = 1.50$.
3. No. $np_0 = 150(0.04) = 6$, which is less than 10 (even though $n(1-p_0) = 144 \geq 10$).

4. P-value $= P(Z \geq 1.75) = 1 - 0.9599 = 0.0401$
5. P-value $= 2P(Z \leq -1.20) = 2(0.1151) = 0.2301$ (technology; table: 0.2302)
6. Large Counts: $150(0.60) = 90$ and $150(0.40) = 60$, both at least 10. $\hat{p} = 99/150 = 0.66$; standard deviation $= \sqrt{0.60(0.40)/150} = \sqrt{0.0016} = 0.04$; $z = \dfrac{0.66 - 0.60}{0.04} = 1.50$; P-value $= P(Z \geq 1.50) = 0.0668$. Since $0.0668 \geq 0.05$, fail to reject H_0: there is not convincing evidence that Henry's long-run free-throw percentage is now above 60%.
7. The P-value is computed assuming H_0 is true, and if H_0 is true the population proportion is exactly p_0, so the standard deviation of $\hat{p}$ is exactly $\sqrt{p_0(1-p_0)/n}$. In a confidence interval there is no assumed value of p, so we must estimate the standard deviation with $\hat{p}$.
8. $H_0 : p = 0.30$, $H_a : p \neq 0.30$. Large Counts: $525(0.30) = 157.5$ and $525(0.70) = 367.5$. $\hat{p} = 168/525 = 0.32$; standard deviation $= \sqrt{0.30(0.70)/525} = \sqrt{0.0004} = 0.02$; $z = \dfrac{0.32 - 0.30}{0.02} = 1.00$; P-value $= 2P(Z \geq 1.00) = 2(0.1587) = 0.3173$ (table: 0.3174). Fail to reject H_0: there is not convincing evidence that the proportion of checkouts that are children's books has changed from 30%.
9. If Henry's long-run free-throw percentage were still 60%, there would be about a 6.68% chance that he would make 66% or more of 150 shots just by chance.
10. State: $H_0 : p = 0.25$, $H_a : p > 0.25$, where p is the proportion of all eggs laid on the new feed that are extra-large; $\alpha = 0.05$. Plan: one-sample z test for p; random sample; 300 is less than 10% of all eggs laid on the new feed; $300(0.25) = 75 \geq 10$ and $300(0.75) = 225 \geq 10$. Do: $\hat{p} = 93/300 = 0.31$; standard deviation $= \sqrt{0.25(0.75)/300} = \sqrt{0.000625} = 0.025$; $z = \dfrac{0.31 - 0.25}{0.025} = 2.40$; P-value $= P(Z \geq 2.40) = 0.0082$. Conclude: $0.0082 < 0.05$, reject H_0; there is convincing evidence that the proportion of extra-large eggs is greater than 25% on the new feed.
11. (a) $\hat{p} = 60/300 = 0.20$; standard deviation $= \sqrt{0.25(0.75)/300} = 0.025$; $z = \dfrac{0.20 - 0.25}{0.025} = -2.00$. (b) P-value $= 2P(Z \leq -2.00) = 0.0455$ (table: $2(0.0228) = 0.0456$). (c) $0.0455 < 0.05$: reject H_0; convincing evidence the spinner lands on red in a proportion different from 0.25. (d) $0.0455 \geq 0.01$: fail to reject H_0.
12. (a) No: $np_0 = 40(0.10) = 4 < 10$. (b) With an expected count of only 4 blemished apples, the sampling distribution of $\hat{p}$ is strongly skewed right, not normal, so normal-curve areas could be far from the true P-value. She should use a larger sample (at least 100 apples) or a simulation or binomial calculation instead.
13. (a) Standard deviation $= \sqrt{0.25/100} = 0.05$; $z = \dfrac{0.55 - 0.50}{0.05} = 1.00$. (b) $P(Z \geq 1.00) = 0.1587$. (c) Standard deviation $= \sqrt{0.25/400} = 0.025$; $z = \dfrac{0.05}{0.025} = 2.00$. (d) $P(Z \geq 2.00) = 0.0228$. (e) A larger sample makes $\hat{p}$ vary less (its standard deviation shrinks like $1/\sqrt{n}$; four times the sample gives half the spread), so the same 5-point excess is 1 standard deviation for Club A but 2 standard deviations for Club B. A difference chance could easily produce in 100 plants is much harder to blame on chance in 400 plants.

LESSON 11.3 · 2 SESSIONS OF 45 MINUTES

Testing a Claim about a Mean

OBJECTIVES

- Compute the one-sample t statistic $t = \dfrac{\bar{x} - \mu_0}{s/\sqrt{n}}$ and its degrees of freedom $n - 1$, and explain why t is used instead of z.
- Find a P-value with technology and give bounds for it from the t table.
- Check the Random, 10% and Normal/Large Sample conditions for a test about a mean.
- Carry out a one-sample t test with the four-step process and state the conclusion in context.

VIDEO

WATCH

t Tests for One Mean: Introduction

jbstatistics · 13:46

When to watch: Before Example 1: the one-sample t test.

https://www.youtube.com/watch?v=T9nI6vhTU1Y

BACKUP

t Tests for One Mean: An Example

jbstatistics · 9:43

When to watch: Extra practice: a one-sample t test.

https://www.youtube.com/watch?v=kQ4xcx6N0o4

KEY TERMS

one-sample t test	A significance test of $H_0 : \mu = \mu_0$ that compares a sample mean with the null value and finds the P-value from a t distribution.
standard error of the mean	The estimate $s/\sqrt{n}$ of the standard deviation $\sigma/\sqrt{n}$ of $\bar{x}$, used because σ is unknown.
t statistic	The number of standard errors the sample mean lies from the null value: $t = (\bar{x} - \mu_0)/(s/\sqrt{n})$.
degrees of freedom (df)	The number that picks out which t distribution to use; for a one-sample t test, $\text{df} = n - 1$.

robust	Describes a procedure whose results stay nearly correct even when a condition (such as a normal population) is not exactly met. The t test is robust unless the data are strongly skewed or have outliers, especially in small samples.

CORE IDEAS

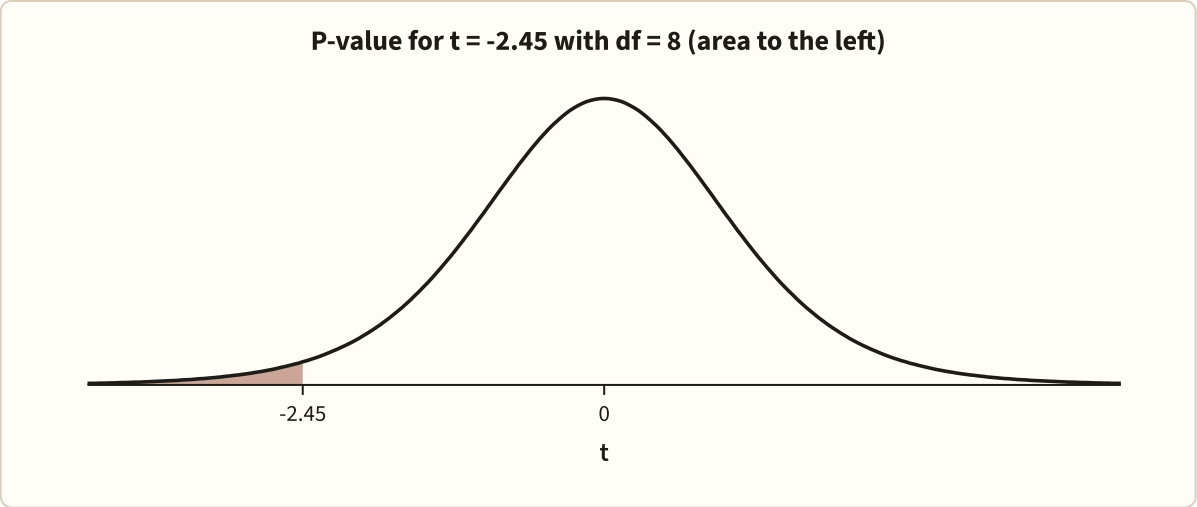
A test about a mean follows the same logic as a test about a proportion: assume $H_0 : \mu = \mu_0$ is true and ask how unusual the observed sample mean $\bar{x}$ would be. From Unit 9, if H_0 is true, $\bar{x}$ is centered at μ_0 with standard deviation $\sigma/\sqrt{n}$. But we almost never know σ, so we estimate it with the sample standard deviation s. The estimate $s/\sqrt{n}$ is the **standard error** of $\bar{x}$.

The test statistic counts standard errors:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \qquad \text{df} = n - 1$$

Why a t distribution and not the standard normal? As you saw in Lesson 10.3, s itself changes from sample to sample. When a sample happens to have a small s, the statistic comes out farther from 0 than z would, so the values of t are more spread out than z values: the t curve has heavier tails. Using the normal table would make P-values too small and claim evidence that isn't really there. The t distribution with $n - 1$ degrees of freedom accounts for the extra uncertainty, and as n grows (and s becomes a steadier estimate of σ) it approaches the normal curve.

Example: honey jars are labeled 16 ounces. A random sample of 9 jars has $\bar{x} = 15.90$ ounces and $s = 0.1225$ ounce, so the standard error is $0.1225/\sqrt{9} = 0.0408$ and $t = (15.90 - 16)/0.0408 = -2.45$ with df $= 8$. For $H_a : \mu < 16$, the P-value is the area to the left of -2.45 under the t curve with 8 degrees of freedom. Technology gives 0.0200: if the jars really averaged 16 ounces, only about 2% of samples of 9 jars would have a mean this far below 16.



Without technology, use the t table. Find the row for the degrees of freedom and locate $|t|$ between two entries; the P-value is between the two tail probabilities at the tops of those columns. For df $= 8$, the value 2.45 is essentially the entry 2.449 in the 0.02 column, so the P-value is about 0.02. If $|t|$ had been

2.60 (between 2.449 and 2.896), the one-sided P-value would be between 0.01 and 0.02. For a two-sided alternative, double both bounds.

Part of the t table: critical values t* with upper-tail probability p

df	p = 0.10	0.05	0.025	0.02	0.01	0.005
8	1.397	1.860	2.306	2.449	2.896	3.355
9	1.383	1.833	2.262	2.398	2.821	3.250
12	1.356	1.782	2.179	2.303	2.681	3.055
15	1.341	1.753	2.131	2.249	2.602	2.947
24	1.318	1.711	2.064	2.172	2.492	2.797

Conditions. **Random:** a random sample or randomized experiment. **10%:** n is at most 10% of the population when sampling without replacement. **Normal/Large Sample:** the population is normal, OR $n \geq 30$ (the central limit theorem makes $\bar{x}$ approximately normal), OR, for a small sample, a dotplot, stemplot or boxplot of the data shows no strong skewness and no outliers. Because t procedures are robust, "roughly symmetric with no outliers" is enough for small samples. Outliers are dangerous because they pull both $\bar{x}$ and s.

Use the same four steps as for proportions: State (hypotheses about μ in words, α), Plan (one-sample t test; check conditions, with a graph when $n < 30$), Do ($\bar{x}$, s, t, df, P-value), Conclude (decision and a sentence about the population mean, in context). Keep full precision in s and the standard error until the final value of t.

KEY FORMULAS

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} \qquad \text{df} = n - 1$$

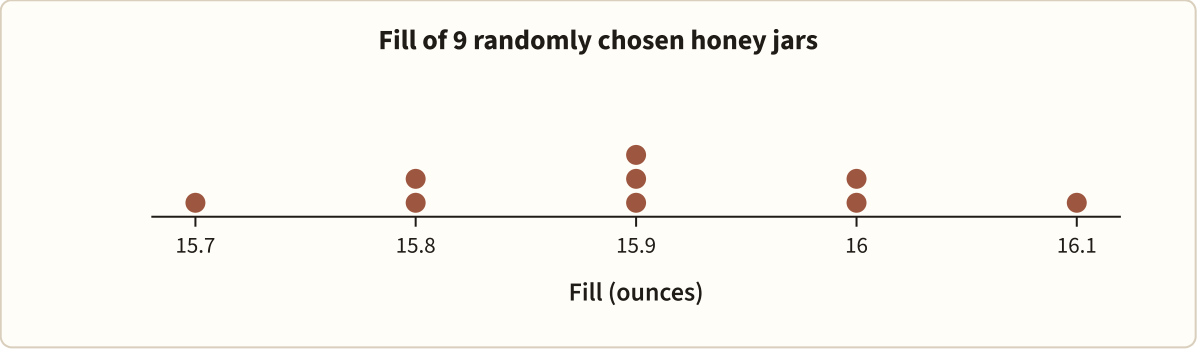
$$\text{standard error of } \bar{x} = \frac{s}{\sqrt{n}}$$

$$H_a : \mu > \mu_0 : P(T \geq t) \qquad H_a : \mu < \mu_0 : P(T \leq t) \qquad H_a : \mu \neq \mu_0 : 2P(T \geq |t|)$$

Practice habit. Keep full precision until the last step: compute s and $s/\sqrt{n}$ without rounding, and round only the final t (2 decimal places) and P-value (4 decimal places).

WORKED EXAMPLES

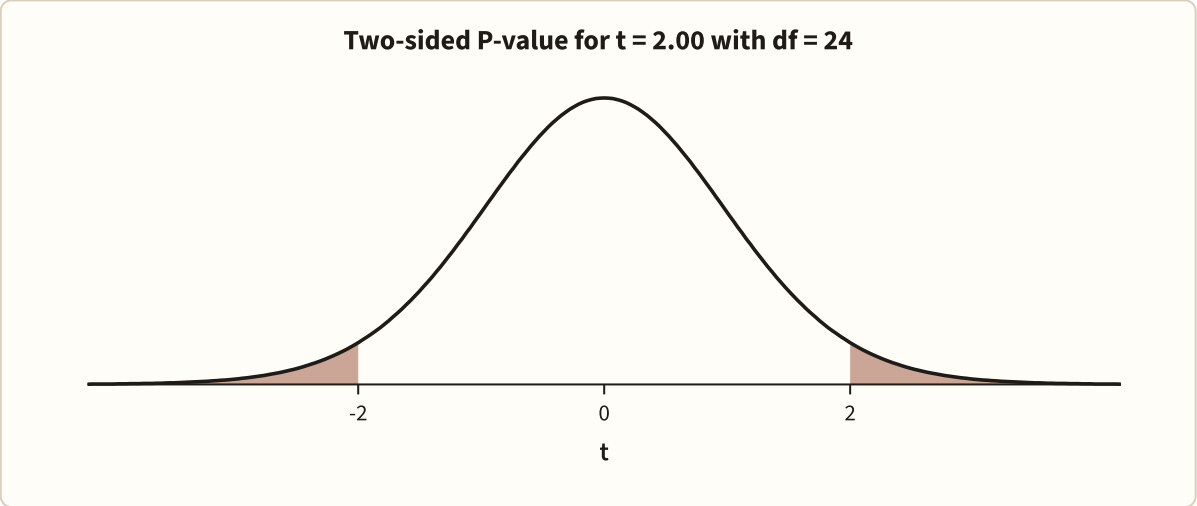
Example 1. A beekeeper's filling machine is set to put 16 ounces of honey in each jar. She suspects it is underfilling. A random sample of 9 jars from one day's run contains these amounts, in ounces: 15.9, 16.0, 15.8, 15.9, 16.1, 15.7, 15.9, 15.8, 16.0. Is there convincing evidence at the $\alpha = 0.05$ level that the mean fill of all jars is less than 16 ounces?



1. State: $H_0 : \mu = 16$ and $H_a : \mu < 16$, where μ is the mean fill (ounces) of all jars filled that day. $\alpha = 0.05$.
2. Plan: one-sample t test for μ. Random: random sample of jars. 10%: 9 jars is less than 10% of the day's jars. Normal/Large Sample: $n = 9 < 30$, but the dotplot shows no strong skewness and no outliers.
3. Do: $\bar{x} = \dfrac{143.1}{9} = 15.90$ ounces.
4. Deviations from 15.9: 0, 0.1, −0.1, 0, 0.2, −0.2, 0, −0.1, 0.1; their squares add to 0.12.
5. $s = \sqrt{\dfrac{0.12}{9-1}} = \sqrt{0.015} = 0.122474$ ounce.
6. Standard error: $\dfrac{0.122474}{\sqrt{9}} = 0.040825$.
7. $t = \dfrac{15.90 - 16}{0.040825} = -2.45$ with df $= 9 - 1 = 8$.
8. P-value $= P(T \le -2.45) = 0.0200$ (technology). Table: 2.449 is in the 0.02 column of the df = 8 row, so P-value ≈ 0.02.
9. Conclude: $0.0200 < 0.05$, so reject H_0. There is convincing evidence that the machine's mean fill is less than 16 ounces; the beekeeper should adjust it.

Answer: $t = -2.45$, df $= 8$, P-value $= 0.0200$. Reject H_0: there is convincing evidence that the mean fill of all jars is less than 16 ounces.

Example 2. On the standard ration, a rancher's calves gain a mean of 2.10 pounds per day. She tries a new feed on a random sample of 25 calves, which gain a mean of $\bar{x} = 2.22$ pounds per day with $s = 0.30$ pound per day. A dotplot of the 25 gains is roughly symmetric with no outliers. Does the new feed change the mean daily gain? Test at the $\alpha = 0.05$ level.



1. State: $H_0 : \mu = 2.10$ and $H_a : \mu \neq 2.10$, where μ is the mean daily gain (pounds) of calves like these on the new feed. The question asks whether the feed changes the gain, so H_a is two-sided. $\alpha = 0.05$.
2. Plan: one-sample t test for μ. Random: random sample of calves. 10%: 25 is less than 10% of all such calves. Normal/Large Sample: $n = 25 < 30$, but the dotplot is roughly symmetric with no outliers.
3. Do: standard error $= \dfrac{0.30}{\sqrt{25}} = 0.06$.
4. $t = \dfrac{2.22 - 2.10}{0.06} = 2.00$ with df $= 25 - 1 = 24$.
5. P-value $= 2P(T \geq 2.00) = 0.0569$ (technology). Table: in the df = 24 row, 2.00 is between 1.711 (0.05) and 2.064 (0.025), so the two-sided P-value is between 0.05 and 0.10.
6. Conclude: $0.0569 \geq 0.05$, so fail to reject H_0. There is not convincing evidence that the new feed changes the mean daily gain of calves.
7. Note: the one-sided P-value would have been about 0.0285, but H_a must come from the question asked before seeing the data, not from the direction the data happened to go. A real gain of 0.12 pound per day may exist; this sample is not enough to show it.

Answer: $t = 2.00$, df $= 24$, P-value $= 0.0569$. Fail to reject H_0: there is not convincing evidence that the new feed changes the mean daily gain.

PRACTICE

1. EASY A beekeeper's hives have averaged 50 pounds of honey per hive in past years. This year a random sample of 16 hives gives $\bar{x} = 52.5$ pounds and $s = 6$ pounds. For $H_0 : \mu = 50$, find (a) the standard error of $\bar{x}$, (b) the t statistic (2 decimal places) and (c) the degrees of freedom.

2. EASY A one-sample t test of $H_0 : \mu = 20$ against $H_a : \mu > 20$ gives $t = 2.20$ with df $= 12$. Use the t table to give bounds for the P-value.

Part of the t table: critical values t* with upper-tail probability p

df	p = 0.10	0.05	0.025	0.02	0.01	0.005
8	1.397	1.860	2.306	2.449	2.896	3.355
9	1.383	1.833	2.262	2.398	2.821	3.250
12	1.356	1.782	2.179	2.303	2.681	3.055
15	1.341	1.753	2.131	2.249	2.602	2.947
24	1.318	1.711	2.064	2.172	2.492	2.797

3. EASY A one-sample t test of $H_0 : \mu = 8$ against $H_a : \mu \neq 8$ gives $t = -1.90$ with df $= 9$. Use the t table to give bounds for the P-value.

Part of the t table: critical values t* with upper-tail probability p

df	p = 0.10	0.05	0.025	0.02	0.01	0.005
8	1.397	1.860	2.306	2.449	2.896	3.355
9	1.383	1.833	2.262	2.398	2.821	3.250
12	1.356	1.782	2.179	2.303	2.681	3.055
15	1.341	1.753	2.131	2.249	2.602	2.947
24	1.318	1.711	2.064	2.172	2.492	2.797

4. EASY Samuel wants to test a claim about the mean daily milk yield of his family's cows using a random sample of 12 days. A stemplot of the 12 yields is strongly skewed right with one high outlier. Is the Normal/Large Sample condition met?

5. MEDIUM A tomato variety is known to produce fruit with a mean weight of 6.0 ounces. Grace grows the variety in compost-rich soil and weighs a random sample of 9 tomatoes: $\bar{x} = 6.5$ ounces and $s = 0.6$ ounce, with no strong skewness or outliers in the data. Test $H_0 : \mu = 6.0$ against $H_a : \mu > 6.0$ at the $\alpha = 0.05$ level. Give t, df, the P-value and your decision.

6. MEDIUM Interpret the P-value of 0.0185 from Problem 5 in context.

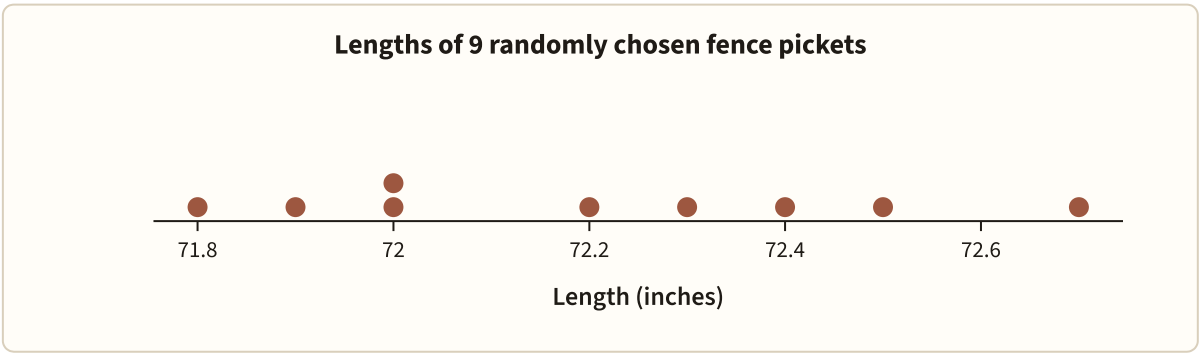
7. MEDIUM A sawmill's boards should average 96 inches long. A random sample of 16 boards has $\bar{x} = 96.3$ inches and $s = 0.6$ inch, with no outliers or strong skewness. For $H_0 : \mu = 96$ against $H_a : \mu \neq 96$: (a) find t; (b) use the t table to give bounds for the P-value; (c) give the decision at $\alpha = 0.05$.

Part of the t table: critical values t* with upper-tail probability p

df	p = 0.10	0.05	0.025	0.02	0.01	0.005
8	1.397	1.860	2.306	2.449	2.896	3.355
9	1.383	1.833	2.262	2.398	2.821	3.250
12	1.356	1.782	2.179	2.303	2.681	3.055
15	1.341	1.753	2.131	2.249	2.602	2.947
24	1.318	1.711	2.064	2.172	2.492	2.797

8. MEDIUM Explain why we use a t distribution rather than the standard normal distribution to find the P-value when we standardize $\bar{x}$ with $s/\sqrt{n}$.

9. HARD Fence pickets are supposed to be 72.0 inches long. Henry measures a random sample of 9 pickets from a new saw setting: 72.7, 71.8, 72.5, 71.9, 72.4, 72.0, 72.0, 72.3, 72.2 inches. Is there convincing evidence at the $\alpha = 0.05$ level that the mean length of all pickets cut at this setting differs from 72.0 inches? Use the four-step process.



10. HARD The cows of a county dairy cooperative have averaged 70 pounds of milk per day. After the cooperative's farms switch to a new grazing plan, a random sample of 25 of the cooperative's cows gives $\bar{x} = 72.4$ pounds and $s = 6.0$ pounds; the data are roughly symmetric with no outliers. Is there convincing evidence at the $\alpha = 0.05$ level that the mean daily yield has increased? Use the four-step process.

11. HARD A county extension agent records the honey harvest from a random sample of 40 hives. A histogram of the 40 harvests is moderately skewed right with no outliers. Is it reasonable to use a one-sample t test for the mean harvest of all hives in the county? Explain.

12. CHALLENGE Suppose Henry in Problem 9 had measured 36 pickets and found the same $\bar{x} = 72.2$ inches and $s = 0.3$ inch. (a) Find the new t statistic. (b) Explain why the evidence against H_0 is now much stronger.

13. **CHALLENGE** In a test of $H_0 : \mu = 16$ against $H_a : \mu < 16$ for honey jars, suppose $n = 9$ and $s = 0.12$ ounce. At $\alpha = 0.05$, what sample mean would give exactly $t = -1.860$, the boundary for rejecting H_0 (the 0.05 entry for df = 8)? Round to 3 decimal places.

COMMON MISTAKES

- Using n instead of $n - 1$ for the degrees of freedom, or using the normal table. For a one-sample t test, df $= n - 1$ and the P-value comes from the t distribution.
- Rounding s or the standard error early. For example, rounding $s = 0.122474$ to 0.12 in the honey example changes t from -2.45 to -2.50. Keep full precision until the final answer.
- Skipping the Normal/Large Sample check for a small sample. When $n < 30$, look at a graph of the data and say what you see ("no strong skewness or outliers").
- Forgetting to double the table bounds for a two-sided alternative. If the one-tail area is between 0.025 and 0.05, the two-sided P-value is between 0.05 and 0.10.

ANSWERS TO PRACTICE

1. (a) $\frac{6}{\sqrt{16}} = 1.5$ pounds. (b) $t = \frac{52.5 - 50}{1.5} = 1.67$. (c) df $= 16 - 1 = 15$.
2. In the df = 12 row, 2.20 is between 2.179 (0.025) and 2.303 (0.02), so the P-value is between 0.02 and 0.025.
3. In the df = 9 row, $|t| = 1.90$ is between 1.833 (0.05) and 2.262 (0.025), so one tail is between 0.025 and 0.05. Doubling for the two-sided test, the P-value is between 0.05 and 0.10.
4. No. The sample is small ($n = 12 < 30$) and the data show strong skewness and an outlier, so a t test would not be trustworthy.
5. Standard error $= 0.6/\sqrt{9} = 0.2$; $t = \frac{6.5 - 6.0}{0.2} = 2.50$; df $= 8$; P-value $= P(T \geq 2.50) = 0.0185$ (table: between 0.01 and 0.02). Since $0.0185 < 0.05$, reject H_0: there is convincing evidence that the mean weight of tomatoes grown in compost-rich soil is greater than 6.0 ounces.
6. If tomatoes grown in compost-rich soil really had a mean weight of 6.0 ounces, there would be only about a 1.85% chance that a random sample of 9 of them would have a mean weight of 6.5 ounces or more.
7. (a) Standard error $= 0.6/\sqrt{16} = 0.15$; $t = \frac{96.3 - 96}{0.15} = 2.00$, df $= 15$. (b) 2.00 is between 1.753 (0.05) and 2.131 (0.025); doubling, the P-value is between 0.05 and 0.10. (c) The P-value is greater than 0.05, so fail to reject H_0: there is not convincing evidence that the mean board length differs from 96 inches.
8. Because σ is unknown we replace it with s, which varies from sample to sample. That extra variability makes the statistic more spread out than a z statistic, with heavier tails. The t distribution with $n - 1$ degrees of freedom has exactly this extra spread; using the normal curve would give P-values that are too small.

9. State: $H_0 : \mu = 72.0$, $H_a : \mu \neq 72.0$, where μ is the mean length (inches) of all pickets cut at this setting; $\alpha = 0.05$. Plan: one-sample t test; random sample; 9 is less than 10% of all pickets; $n < 30$ but the dotplot shows no strong skewness or outliers. Do: $\bar{x} = 649.8/9 = 72.2$; squared deviations add to 0.72, so $s = \sqrt{0.72/8} = 0.3$; standard error $= 0.3/3 = 0.1$; $t = \dfrac{72.2 - 72.0}{0.1} = 2.00$, df $= 8$; P-value $= 2P(T \geq 2.00) = 0.0805$. Conclude: $0.0805 \geq 0.05$, fail to reject H_0; there is not convincing evidence that the mean picket length at this setting differs from 72.0 inches.
10. State: $H_0 : \mu = 70$, $H_a : \mu > 70$, where μ is the mean daily milk yield (pounds) of all the cooperative's cows on the new grazing plan; $\alpha = 0.05$. Plan: one-sample t test; random sample; 25 cows is less than 10% of the cooperative's cows; $n < 30$ but the data have no strong skewness or outliers. Do: standard error $= 6.0/\sqrt{25} = 1.2$; $t = \dfrac{72.4 - 70}{1.2} = 2.00$, df $=$ 24; P-value $= P(T \geq 2.00) = 0.0285$ (table: between 0.025 and 0.05). Conclude: $0.0285 < 0.05$, reject H_0; there is convincing evidence that the mean daily milk yield on the new pasture is greater than 70 pounds.
11. Yes. The sample is large ($n = 40 \geq 30$), so by the central limit theorem $\bar{x}$ is approximately normal even though the harvests are moderately skewed, and there are no outliers to distort $\bar{x}$ and s.
12. (a) Standard error $= 0.3/\sqrt{36} = 0.05$; $t = \dfrac{72.2 - 72.0}{0.05} = 4.00$ (df = 35, P-value well below 0.01). (b) The same 0.2-inch difference is now 4 standard errors instead of 2, because averages of 36 pickets vary much less than averages of 9. A difference that chance could produce with 9 pickets is very hard to blame on chance with 36.
13. Standard error $= 0.12/\sqrt{9} = 0.04$. Solve $\dfrac{\bar{x} - 16}{0.04} = -1.860$: $\bar{x} = 16 - 1.860(0.04) = 16 - 0.0744 = 15.926$ ounces. Any sample mean below about 15.926 ounces would lead to rejecting H_0.

LESSON 11.4 · 2 SESSIONS OF 45 MINUTES

Type I and Type II Errors and Power

OBJECTIVES

- Describe a Type I error and a Type II error in context, with a consequence of each.
- Explain why the probability of a Type I error equals α, and define β and power.
- Compute α, power and β for a test with a stated rejection rule in a simple normal setting.
- Explain how sample size, significance level, the size of the true effect and variability change the power of a test.

VIDEO

WATCH

Type I Errors, Type II Errors, and the Power of the Test

jbstatistics · 8:10

When to watch: Before Example 1: Type I and Type II errors and power.

https://www.youtube.com/watch?v=7mE-K_w1v90

BACKUP

What Factors Affect the Power of a Z Test?

jbstatistics · 12:25

When to watch: After Example 2: what makes power larger.

https://www.youtube.com/watch?v=K6tado8Xcug

KEY TERMS

Type I error	Rejecting H_0 when H_0 is actually true: a false alarm. Its probability is α.
Type II error	Failing to reject H_0 when H_0 is actually false: a missed effect. Its probability is called β.
power	The probability that a test rejects H_0 when a particular alternative value of the parameter is true; power $= 1 - \beta$. It is the chance of detecting an effect that really exists.
rejection rule	The set of values of the statistic (for example, "$\bar{x} \leq 15.92$") for which the test rejects H_0.

CORE IDEAS

A significance test ends in a decision, and the decision is based on a sample, so it can be wrong. There are two ways to be right and two ways to be wrong, depending on the truth (which we never see) and

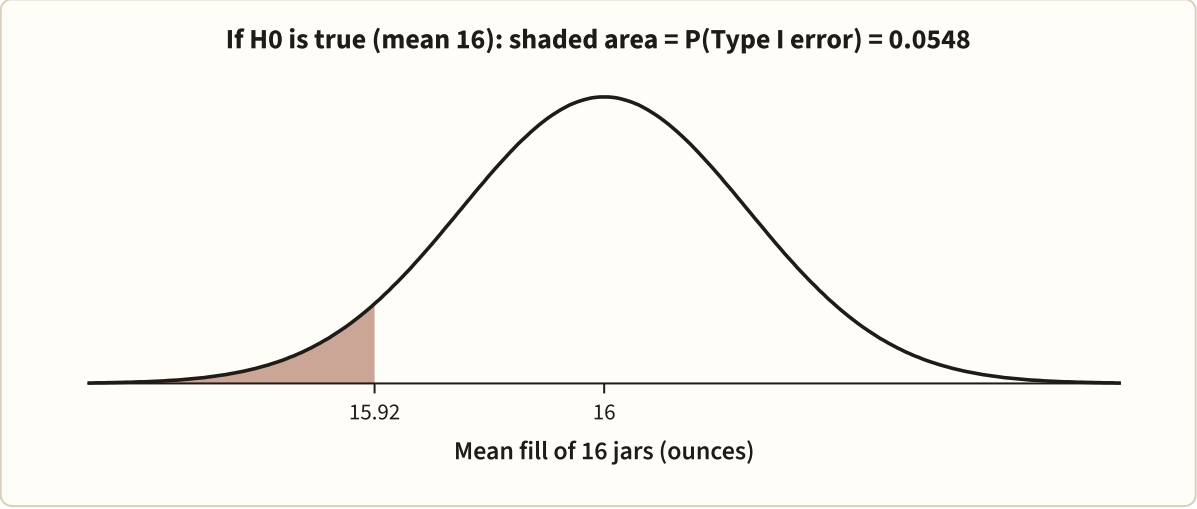
our decision.

The four possible outcomes of a significance test

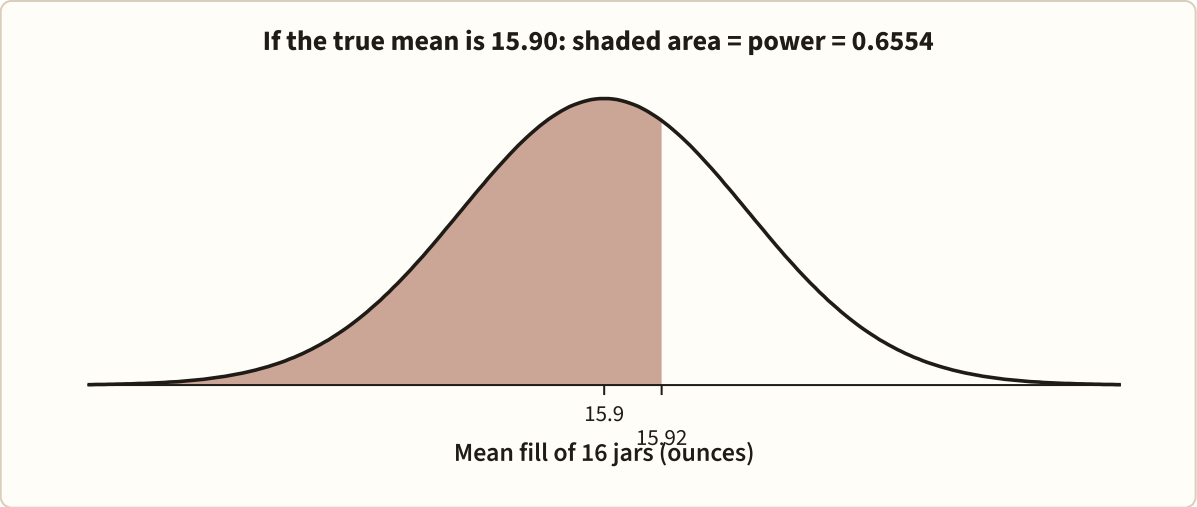
Decision	H0 is true	H0 is false
Reject H0	Type I error (probability alpha)	Correct decision (probability = power)
Fail to reject H0	Correct decision	Type II error (probability beta)

A **Type I error** is rejecting a true H_0: a false alarm. A **Type II error** is failing to reject a false H_0: a real effect goes unnoticed. Always describe them in context. A beekeeper tests $H_0 : \mu = 16$ against $H_a : \mu < 16$ for the mean fill of her honey jars. A Type I error would mean concluding the machine is underfilling when it is actually fine; she would stop and adjust a machine that needed nothing. A Type II error would mean failing to find underfilling that is really happening; her customers would keep getting less honey than the label promises. Which error is worse depends on these consequences.

Why is $P(\text{Type I error}) = \alpha$? If H_0 is true, the P-value is less than α only when the result lands in the most extreme part of the sampling distribution, a part that by design holds probability α. So a test at $\alpha = 0.05$ will reject a true H_0 in 5% of samples in the long run. More generally, for any rejection rule, $P(\text{Type I error})$ is the area of the rejection region under the sampling distribution centered at the null value. For the honey machine, suppose experience shows the fills have $\sigma = 0.20$ ounce, and she weighs 16 jars, so $\bar{x}$ has standard deviation $0.20/\sqrt{16} = 0.05$. Her rule is to adjust the machine if $\bar{x} \leq 15.92$. If the machine is fine ($\mu = 16$), $P(\bar{x} \leq 15.92) = P(Z \leq \dfrac{15.92 - 16}{0.05}) = P(Z \leq -1.60) = 0.0548$.



Power answers the other question: if the machine really is underfilling, how likely is the test to catch it? Power depends on HOW wrong H_0 is, so we must pick a particular alternative value. If the true mean is $\mu = 15.90$, the sampling distribution of $\bar{x}$ is centered at 15.90 instead, and power $= P(\bar{x} \leq 15.92) = P(Z \leq \dfrac{15.92 - 15.90}{0.05}) = P(Z \leq 0.40) = 0.6554$. So $\beta = 1 - 0.6554 = 0.3446$: about one-third of the time, a machine underfilling by 0.1 ounce would slip past this test. A worse machine, with $\mu = 15.85$, would be caught with power $P(Z \leq 1.40) = 0.9192$.



Four things increase power. **A larger sample** makes $\bar{x}$ less variable (standard deviation $\sigma/\sqrt{n}$), so the null and alternative curves overlap less. **A larger** α widens the rejection region, but at the price of more Type I errors. **A larger true effect** puts the alternative curve farther from the null value. **Less variability** (more careful measurement, more uniform conditions, blocking) narrows both curves. Because lowering α lowers power, the usual way to get both a small α and high power is a larger sample. Good studies choose n in advance to give high power (often 0.80 or more) against an effect large enough to matter, and a "fail to reject" from a low-power study tells us very little.

KEY FORMULAS

$$P(\text{Type I error}) = \alpha$$

$$P(\text{Type II error}) = \beta \qquad \text{power} = 1 - \beta$$

$$\text{power} = P(\text{reject } H_0 \mid \text{a particular alternative value is true})$$

From history. In papers written in the late 1920s and early 1930s, Jerzy Neyman and Egon Pearson described a test as a rule for deciding between two hypotheses and named the two kinds of error it can make. Their idea of choosing a test to keep the Type I error rate at α while making power as large as possible still guides how studies are planned.

Practice habit. Describe each error in context with its consequence: say what the truth is, what you decided, and what happens as a result ("The machine is fine, but she concludes it is underfilling, so she wastes time adjusting it").

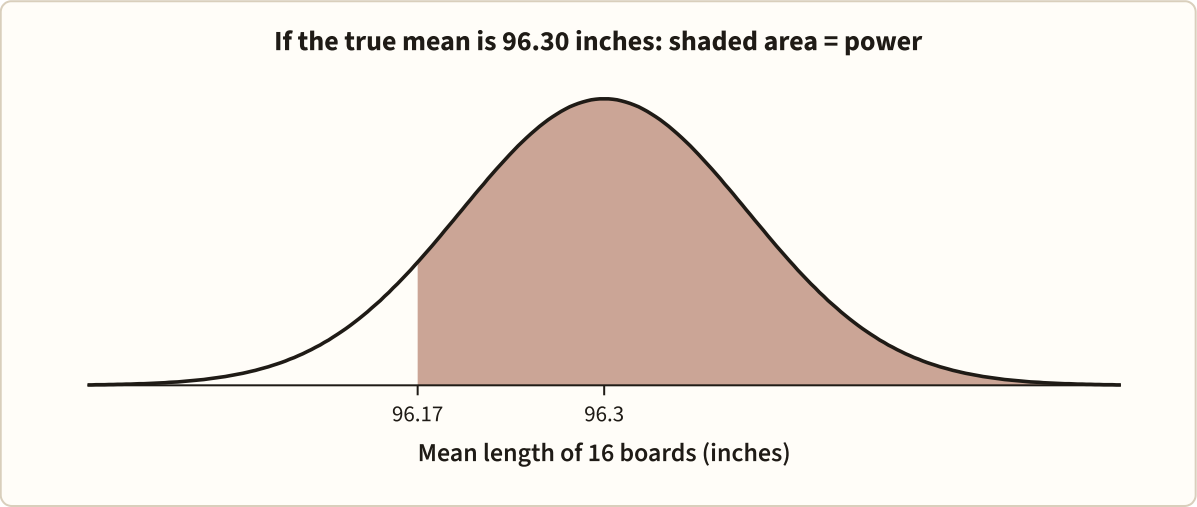
WORKED EXAMPLES

Example 1. An orchard owner tries a new spray and tests $H_0 : p = 0.10$ against $H_a : p < 0.10$, where p is the proportion of all her apples this season with worm damage (10% is the usual rate). (a) Describe a Type I error and its consequence. (b) Describe a Type II error and its consequence.

1. A Type I error rejects a true H_0. Here H_0 true means the spray does not lower the damage rate below 10%.
2. (a) Type I: she concludes the spray reduces worm damage below 10% when in fact it does not. Consequence: she keeps buying and applying a spray that doesn't help.
3. A Type II error fails to reject a false H_0. Here H_0 false means the spray really does lower the damage rate.
4. (b) Type II: she finds no convincing evidence that the spray helps when in fact it does reduce damage below 10%. Consequence: she may give up a spray that would have saved part of her crop.

Answer: (a) Type I: concluding the spray lowers worm damage below 10% when it does not, so she wastes money on it. (b) Type II: failing to find that the spray lowers worm damage when it really does, so she abandons a helpful spray.

Example 2. Boards that run long waste lumber, so a sawmill foreman tests $H_0 : \mu = 96$ against $H_a : \mu > 96$, where μ is the mean length (inches) of boards the saw is cutting. Board lengths have $\sigma = 0.4$ inch, and he measures 16 boards. His rule: adjust the saw if $\bar{x} \geq 96.17$ inches. (a) Find the probability of a Type I error. (b) Find the power of the test if the true mean is 96.30 inches. (c) Find β for that alternative and interpret the power.



1. Standard deviation of $\bar{x}$: $\dfrac{0.4}{\sqrt{16}} = 0.1$ inch.
2. (a) A Type I error happens if H_0 is true ($\mu = 96$) but $\bar{x} \geq 96.17$. $z = \dfrac{96.17 - 96}{0.1} = 1.70$.
3. $P(\text{Type I}) = P(Z \geq 1.70) = 1 - 0.9554 = 0.0446$.
4. (b) Now suppose $\mu = 96.30$. The test rejects when $\bar{x} \geq 96.17$: $z = \dfrac{96.17 - 96.30}{0.1} = -1.30$.
5. Power $= P(Z \geq -1.30) = 1 - 0.0968 = 0.9032$.
6. (c) $\beta = 1 - 0.9032 = 0.0968$.
7. Interpret: if the saw is really cutting boards that average 96.30 inches, the foreman's rule will detect it (and he will adjust the saw) about 90% of the time; about 10% of the time he will miss it.

Answer: (a) $P(\text{Type I}) = P(Z \geq 1.70) = 0.0446$. (b) Power $= P(Z \geq -1.30) = 0.9032$. (c) $\beta = 0.0968$; if the true mean is 96.30 inches, the rule catches the problem about 90% of the time.

PRACTICE

1. EASY A test is carried out at the $\alpha = 0.01$ significance level. If H_0 is actually true, what is the probability of making a Type I error?
2. EASY A test has power 0.80 against a particular alternative. What is the probability of a Type II error for that alternative?

3. EASY A baker tests $H_0 : \mu = 24$ against $H_a : \mu < 24$, where μ is the mean weight (ounces) of the loaves from his new oven. Describe a Type I error in context and give one consequence.

4. EASY For the baker in Problem 3, describe a Type II error in context and give one consequence.

5. MEDIUM Before selling a batch of seed, a seed company tests $H_0 : p = 0.85$ against $H_a : p < 0.85$, where p is the batch's germination rate, and throws the batch away if it rejects H_0. Which error would the company's customers consider more serious, and should the company choose $\alpha = 0.01$ or $\alpha = 0.10$ to protect them? Explain.

6. MEDIUM For each change, say whether the power of a test increases or decreases (all else staying the same), and why: (a) the sample size is increased from 16 to 64; (b) α is lowered from 0.05 to 0.01; (c) the true mean is farther from the null value; (d) more careful measuring reduces σ.

7. MEDIUM Use the sawmill test from the second worked example ($\sigma = 0.4$ inch, $n = 16$, reject $H_0 : \mu = 96$ if $\bar{x} \geq 96.17$). (a) Find the power if the true mean is 96.20 inches. (b) Explain why the power is lower than it was for a true mean of 96.30 inches.

8. MEDIUM In the second worked example, the power against $\mu = 96.30$ inches is 0.9032. Interpret this value in context.

9. MEDIUM For the honey machine in the core ideas ($\sigma = 0.20$ ounce, $n = 16$, adjust if $\bar{x} \leq 15.92$), find (a) the power if the true mean fill is 15.85 ounces and (b) β for that alternative.

10. HARD The sawmill foreman decides to measure 64 boards instead of 16 ($\sigma = 0.4$ inch). His new rule is to adjust the saw if $\bar{x} \geq 96.09$. (a) Find the probability of a Type I error. (b) Find the power if the true mean is 96.20 inches. (c) Compare with the power of 0.6179 for $n = 16$ (Problem 7) and explain the difference.

11. HARD Suppose the foreman, measuring 16 boards ($\sigma = 0.4$ inch), wants fewer false alarms and changes his rule to adjust the saw only if $\bar{x} \geq 96.21$. (a) Find the new probability of a Type I error. (b) Find the new power against a true mean of 96.30 inches. (c) Compare with the worked example (0.0446 and 0.9032) and describe the trade-off.

12. HARD A rancher compares a new feed with the old one using only 10 calves and gets P-value $=$ 0.40. A statistician tells her the study had power of only 0.25 to detect a difference of 0.2 pound per day in mean daily gain. What should the rancher conclude, and what would you advise?

13. CHALLENGE A seed company tests $H_0 : p = 0.90$ against $H_a : p < 0.90$ for its germination rate with a random sample of 400 seeds, and rejects H_0 if $\hat{p} \leq 0.87$. (a) Find the probability of a Type I error. (b) Find the power if the true germination rate is $p = 0.80$. (c) Explain why the standard deviation of $\hat{p}$ is different in parts (a) and (b).

COMMON MISTAKES

- Describing an error without context or consequence ("rejecting H_0 when it is true"). Say what the truth is and what you concluded in the setting, then what would happen.

- Thinking power is the probability that H_a is true. Power is the probability that the test rejects H_0, calculated assuming a particular alternative value is the truth.
- Using the null value as the center when computing power. The rejection rule comes from H_0, but power is computed with the sampling distribution centered at the alternative value.
- Treating a "fail to reject" from a small, low-power study as evidence that there is no effect. Low power means real effects are often missed.

ANSWERS TO PRACTICE

1. $P(\text{Type I error}) = \alpha = 0.01$
2. $\beta = 1 - 0.80 = 0.20$
3. A Type I error would be concluding that the loaves average less than 24 ounces when they really average 24 ounces. Consequence: he might adjust his recipe or oven, or throw out good loaves, for no reason.
4. A Type II error would be failing to find convincing evidence that the loaves are light when they really average less than 24 ounces. Consequence: he keeps selling underweight loaves, and customers get less bread than they pay for.
5. Customers are hurt by a Type II error: a batch with a germination rate below 85% is not caught and is sold. To make Type II errors less likely, the company should use the larger $\alpha = 0.10$, which makes the test reject more readily (higher power), at the cost of throwing away some good batches (more Type I errors).
6. (a) Increases: a larger sample makes $\bar{x}$ less variable, so a real difference is easier to see. (b) Decreases: a smaller α shrinks the rejection region, so it is harder to reject even a false H_0. (c) Increases: a bigger effect is easier to detect. (d) Increases: less variability makes $\bar{x}$ vary less, just as a larger sample does.
7. (a) $z = \dfrac{96.17 - 96.20}{0.1} = -0.30$; power $= P(Z \geq -0.30) = 0.6179$. (b) A true mean of 96.20 is closer to 96, so the sampling distribution of $\bar{x}$ overlaps more with the region where we fail to reject. A smaller effect is harder to detect.
8. If the saw is really cutting boards with a mean length of 96.30 inches, there is about a 90% chance that the mean of 16 measured boards will be 96.17 inches or more, so the foreman will correctly conclude the boards are too long and adjust the saw.
9. (a) Standard deviation of $\bar{x} = 0.05$; $z = \dfrac{15.92 - 15.85}{0.05} = 1.40$; power $= P(Z \leq 1.40) = 0.9192$. (b) $\beta = 1 - 0.9192 = 0.0808$.
10. (a) Standard deviation of $\bar{x} = 0.4/\sqrt{64} = 0.05$; $z = \dfrac{96.09 - 96}{0.05} = 1.80$; $P(\text{Type I}) = P(Z \geq 1.80) = 0.0359$. (b) $z = \dfrac{96.09 - 96.20}{0.05} = -2.20$; power $= P(Z \geq -2.20) = 0.9861$. (c) Power rose from 0.6179 to 0.9861 (with an even smaller α) because four times as many boards cut the standard deviation of $\bar{x}$ in half, so the null and alternative distributions barely overlap.

11. (a) $z = \dfrac{96.21 - 96}{0.1} = 2.10$; $P(\text{Type I}) = P(Z \geq 2.10) = 0.0179$. (b) $z = \dfrac{96.21 - 96.30}{0.1} = -0.90$; power $= P(Z \geq -0.90) = 0.8159$. (c) The stricter rule lowers the chance of a false alarm from 0.0446 to 0.0179, but the power drops from 0.9032 to 0.8159, so long boards are missed more often. With a fixed sample size, reducing one kind of error increases the other.

12. She should not conclude that the new feed makes no difference. With power 0.25, even a real 0.2-pound-per-day improvement would be missed 75% of the time, so failing to reject H_0 tells her little. She should repeat the study with many more calves so the test has a good chance (such as 0.80 or more) of detecting a difference that matters.

13. (a) If $p = 0.90$: standard deviation $= \sqrt{0.90(0.10)/400} = 0.015$; $z = \dfrac{0.87 - 0.90}{0.015} = -2.00$; $P(\text{Type I}) = P(Z \leq -2.00) = 0.0228$. (b) If $p = 0.80$: standard deviation $= \sqrt{0.80(0.20)/400} = 0.02$; $z = \dfrac{0.87 - 0.80}{0.02} = 3.50$; power $= P(Z \leq 3.50) = 0.9998$. (c) The standard deviation of $\hat{p}$ is $\sqrt{p(1-p)/n}$, which depends on the true value of p. In (a) we assume $p = 0.90$; in (b) we assume $p = 0.80$, so the spread changes along with the center.

LESSON 11.5 · 2 SESSIONS OF 45 MINUTES

Tests and Intervals Together

OBJECTIVES

- Use a $100(1-\alpha)\%$ confidence interval to decide a two-sided test at level α, and explain why the two agree.
- Explain the difference between statistical significance and practical importance, using a confidence interval to judge the size of an effect.
- Compute the probability of at least one false alarm among several tests and explain why reporting only significant results misleads.
- Explain why no test or interval can rescue data from a badly designed sample or experiment.

VIDEO

WATCH

The Relationship Between Confidence Intervals and Hypothesis Tests

jbstatistics · 5:35

When to watch: Before Example 1: how a confidence interval and a two-sided test agree.

https://www.youtube.com/watch?v=k1at8VukIbw

BACKUP

Statistical Significance versus Practical Significance

jbstatistics · 4:47

When to watch: Before Example 2: statistical versus practical significance.

https://www.youtube.com/watch?v=_k1MQTUCXmU

KEY TERMS

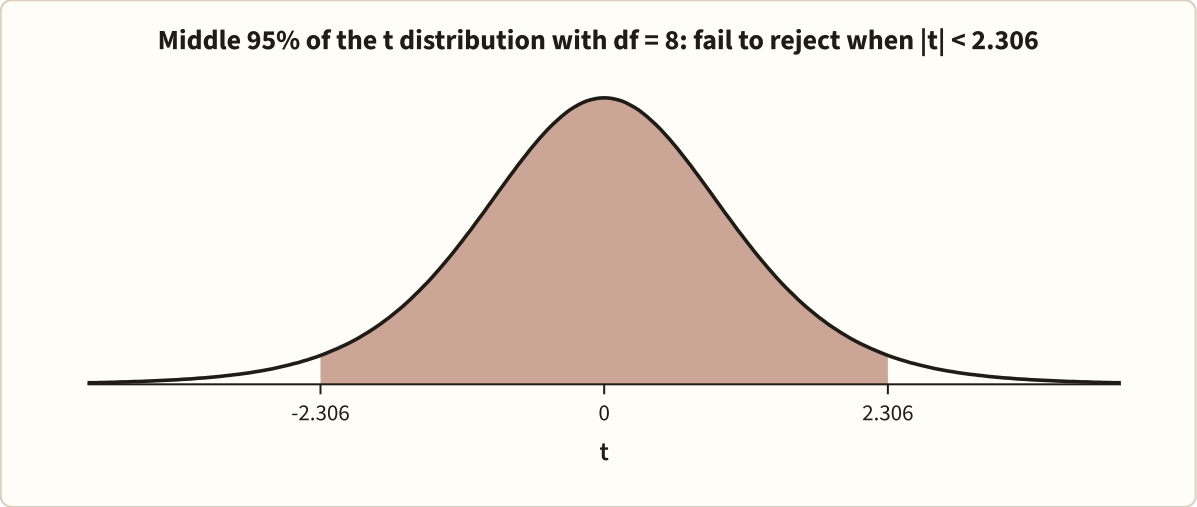
duality of tests and intervals	The fact that a two-sided test at level α rejects $H_0 : \mu = \mu_0$ exactly when μ_0 falls outside the $100(1-\alpha)\%$ confidence interval for μ.
practical importance	Whether an effect is large enough to matter in the real situation. A statistically significant effect can be too small to matter.
multiple testing	Running many significance tests on the same study or data. Even when every null hypothesis is true, some tests will be "significant" by chance alone.

effect size	How large the difference between the true parameter and the null value is, in the units of the problem; a confidence interval estimates it.

CORE IDEAS

A confidence interval and a two-sided test are built from the same ingredients: the statistic, its standard error and a critical value. The interval gives the set of parameter values that are plausible in light of the data. A two-sided test asks whether one particular value, μ_0, is plausible. So it should not surprise you that they give the same answer.

Here is why, for a mean. A two-sided t test at level α rejects H_0 when $|t| > t^*$, where t^* is the critical value for $100(1-\alpha)\%$ confidence. Multiplying both sides by the standard error, that is the same as $|\bar{x} - \mu_0| > t^* \cdot \dfrac{s}{\sqrt{n}}$: the null value is farther from $\bar{x}$ than the margin of error. And that says exactly that μ_0 lies outside the interval $\bar{x} \pm t^* \dfrac{s}{\sqrt{n}}$.



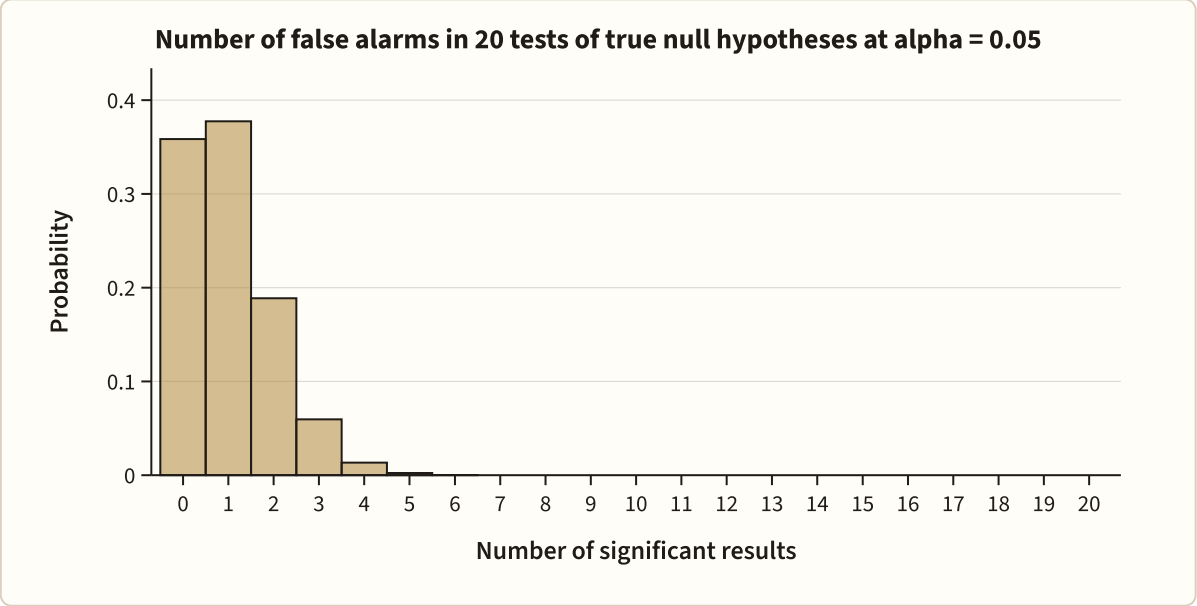
Example: Henry's 9 fence pickets had $\bar{x} = 72.2$ inches and $s = 0.3$ inch, so the standard error is 0.1 and the 95% interval is $72.2 \pm 2.306(0.1) = (71.969, 72.431)$. Every null value inside the interval gives $|t| < 2.306$ and is not rejected at $\alpha = 0.05$; every value outside is rejected. In Lesson 11.3 the test of $\mu_0 = 72.0$ gave P-value $= 0.0805$, and indeed 72.0 is inside the interval. For proportions the test uses p_0 in the standard deviation and the interval uses $\hat{p}$, so in borderline cases the two can disagree slightly; nearly always they agree.

Fence pickets (n = 9, mean 72.2, standard error 0.1): 95% interval (71.969, 72.431)

Null value (inches)	t statistic	Inside the 95% interval?	Decision at alpha = 0.05
71.9	3	No	Reject H0
72	2	Yes	Fail to reject H0
72.4	-2	Yes	Fail to reject H0
72.5	-3	No	Reject H0

Report the interval along with the test. The test answers "is there an effect?" and the interval answers "how big is it?" The second question matters because **statistical significance is not the same as practical importance**. With a very large sample, even a tiny difference is many standard errors from the null value. A seed company that tests 10,000 seeds and finds a 79% germination rate can show convincingly that the rate is below its claimed 80%, but the interval (78.2%, 79.8%) shows the shortfall is about one percentage point, which no gardener would notice. On the other hand, a small study can miss an effect that is large and important (low power).

Beware of many tests. If every null hypothesis is true, each test at $\alpha = 0.05$ still has a 5% chance of a false alarm. Run 20 independent tests and the number of false alarms is binomial with $n = 20$ and $p = 0.05$: the expected number is $20(0.05) = 1$, and $P(\text{at least one}) = 1 - 0.95^{20} = 0.6415$. A study that tries many things and reports only the one "significant" result is reporting what chance alone would usually produce. Honest reports say how many tests were run.



Finally, inference cannot rescue badly collected data. P-values and confidence intervals measure only the random variation that comes from random sampling or random assignment. They say nothing about bias from a voluntary response or convenience sample, poorly worded questions, nonresponse,

or confounding in an observational study. A precise-looking P-value computed from bad data is still bad information. Always ask first how the data were produced.

KEY FORMULAS

$$\text{two-sided test at level } \alpha \text{ rejects } H_0 : \mu = \mu_0 \iff \mu_0 \text{ is outside the } 100(1-\alpha)\% \text{ confidence interval}$$

$$|t| > t^* \iff |\bar{x} - \mu_0| > t^* \frac{s}{\sqrt{n}}$$

$$P(\text{at least one false alarm in } k \text{ independent tests}) = 1 - (1-\alpha)^k$$

Practice habit. Ask "compared to what?" and "how big?": report a confidence interval with every test, and say in context whether the size of the effect matters in practice.

WORKED EXAMPLES

Example 1. A 4-H member weighs a random sample of lambs at weaning and computes a 95% confidence interval for the mean weaning weight of all lambs on the farm: (82.4, 88.6) pounds. Without doing any more calculations, decide each two-sided test at the $\alpha = 0.05$ level: (a) $H_0 : \mu = 80$; (b) $H_0 : \mu = 85$; (c) $H_0 : \mu = 90$.

1. A two-sided test at $\alpha = 0.05$ matches a $100(1 - 0.05)\% = 95\%$ confidence interval.
2. (a) 80 is below 82.4, outside the interval, so 80 is not a plausible value of μ: reject H_0.
3. (b) 85 is inside (82.4, 88.6), so it is plausible: fail to reject H_0.
4. (c) 90 is above 88.6, outside the interval: reject H_0.
5. In context: there is convincing evidence that the mean weaning weight is not 80 pounds and not 90 pounds, but not convincing evidence that it differs from 85 pounds.

Answer: (a) Reject H_0 (80 is outside the interval). (b) Fail to reject H_0 (85 is inside). (c) Reject H_0 (90 is outside).

Example 2. A seed company claims that 80% of its carrot seeds germinate. A testing laboratory plants a random sample of 10,000 seeds, and 7,900 germinate. (a) For $H_0 : p = 0.80$ against $H_a : p \neq 0.80$, find the z statistic. (b) Find the P-value and the decision at $\alpha = 0.05$. (c) Construct a 95% confidence interval for p (3 decimal places). (d) Is the difference from 80% practically important to a home gardener?

1. Conditions: random sample; 10,000 is less than 10% of the company's seeds; for the test, $10{,}000(0.80) = 8{,}000$ and $10{,}000(0.20) = 2{,}000$ are both at least 10; for the interval, the observed counts of 7,900 successes and 2,100 failures are both at least 10.
2. (a) $\hat{p} = \dfrac{7{,}900}{10{,}000} = 0.79$. Standard deviation if H_0 is true: $\sqrt{\dfrac{0.80(0.20)}{10{,}000}} = 0.004$.
3. $z = \dfrac{0.79 - 0.80}{0.004} = -2.50$.
4. (b) P-value $= 2P(Z \leq -2.50) = 2(0.0062) = 0.0124$. Since $0.0124 < 0.05$, reject H_0: there is convincing evidence that the germination rate is not 80%.
5. (c) Standard error for the interval uses $\hat{p}$: $\sqrt{\dfrac{0.79(0.21)}{10{,}000}} = 0.0040731$.
6. $0.79 \pm 1.96(0.0040731) = 0.79 \pm 0.0080 = (0.782, 0.798)$. The interval does not contain 0.80, agreeing with the test.
7. (d) We are 95% confident that the true germination rate is between 78.2% and 79.8%, only about 0.2 to 1.8 percentage points below the claim. Out of a packet of 100 seeds that is about one fewer carrot: statistically significant, but not practically important for a gardener.

Answer: (a) $z = -2.50$. (b) P-value $= 0.0124$; reject H_0. (c) (0.782, 0.798). (d) No: the true rate is within about 2 percentage points of 80%, a difference too small to matter in a garden, even though it is statistically significant.

PRACTICE

1. EASY A 95% confidence interval for the mean tomato yield per plant in a garden is (3.1, 3.9) pounds. Decide each two-sided test at the $\alpha = 0.05$ level: (a) $H_0 : \mu = 4$; (b) $H_0 : \mu = 3.5$.
2. EASY A two-sided test is carried out at the $\alpha = 0.01$ level. What confidence level, as a percent, gives an interval that always agrees with this test (for a mean)?
3. EASY A researcher runs 40 independent significance tests at the $\alpha = 0.05$ level, and every null hypothesis is actually true. How many "significant" results should she expect by chance alone?
4. EASY Explain in plain words what it means to say that a result is "statistically significant at the $\alpha = 0.05$ level."

5. **MEDIUM** An egg farm says its large eggs have a mean weight of 58 grams. A random sample of 9 eggs has $\bar{x} = 57.0$ grams and $s = 1.8$ grams, with no strong skewness or outliers. (a) Construct a 95% confidence interval for the mean weight (2 decimal places). (b) Use your interval to decide the test of $H_0 : \mu = 58$ against $H_a : \mu \neq 58$ at $\alpha = 0.05$.

6. **MEDIUM** Samuel computes a 90% confidence interval for the mean weight of the cabbages in his garden: (5.2, 6.0) pounds. (a) For what significance level does this interval decide a two-sided test? (b) What is the decision for $H_0 : \mu = 6.2$?

7. **MEDIUM** A golf-ball maker tests a new ball with 5,000 drives by a driving machine. The new ball's drives average 1.2 yards longer than the old ball's (P-value $= 0.001$; 95% confidence interval for the increase: 0.5 to 1.9 yards). An advertisement says the new ball is "proven to go significantly farther." Comment.

8. **MEDIUM** Grace's coach says she makes 80% of her free throws. A 95% confidence interval for her true free-throw proportion, from a random sample of her shots, is (0.62, 0.74). (a) What is the decision for a two-sided test of $H_0 : p = 0.80$ at $\alpha = 0.05$? (b) What does the interval tell you that the test alone does not?

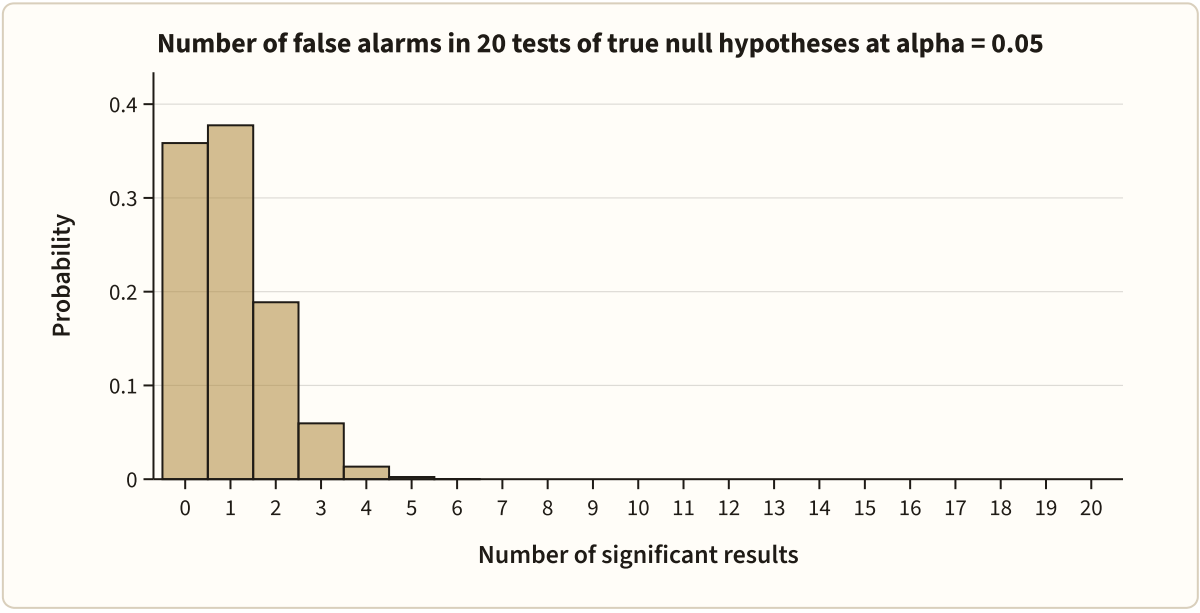
9. **MEDIUM** During the county fair, a radio station invites listeners to call in and vote for their favorite pie; 2,000 people call, and 41% choose apple. A student computes a 95% confidence interval of (0.388, 0.432) for the proportion of all county residents who prefer apple pie. Explain why the interval cannot be trusted.

10. **HARD** In a random sample of 400 apples from a large orchard, 52 have blemishes. The owner's usual rate is 10%. (a) For $H_0 : p = 0.10$ against $H_a : p \neq 0.10$, find z. (b) Find the P-value. (c) Construct a 95% confidence interval for p (3 decimal places). (d) At $\alpha = 0.05$ the test rejects H_0, yet 0.10 is inside the interval. Explain how this can happen.

11. **HARD** A 99% confidence interval for the mean height of corn plants 60 days after planting is (41.2, 46.8) inches. (a) Decide the two-sided test of $H_0 : \mu = 40$ at $\alpha = 0.01$. (b) Decide the two-sided test of $H_0 : \mu = 46$ at $\alpha = 0.01$. (c) What was the sample mean?

12. **HARD** A large honey packer checks a random sample of 2,500 jars labeled 16 ounces and finds $\bar{x} = 16.01$ ounces and $s = 0.25$ ounce. (a) For $H_0 : \mu = 16$ against $H_a : \mu \neq 16$, find t. (b) Find the P-value (df = 2,499; with so many degrees of freedom, the t distribution is almost exactly normal). (c) Is the result practically important? Explain.

13. **CHALLENGE** A garden club tests 20 different folk remedies for tomato plants (crushed eggshells, coffee grounds, and so on), each with its own independent experiment at $\alpha = 0.05$. Suppose none of the remedies actually works. (a) What is the expected number of "significant" results? (b) Find the probability that at least one result is significant (4 decimal places). (c) The club's newsletter reports only the one remedy that came out significant. Why is this misleading?



COMMON MISTAKES

- Matching a test with the wrong confidence level. A two-sided test at level α matches a $100(1-\alpha)\%$ interval: $\alpha = 0.05$ with 95%, $\alpha = 0.01$ with 99%, $\alpha = 0.10$ with 90%.
- Treating "statistically significant" as "large" or "important." Look at the confidence interval and ask whether an effect of that size matters in context.
- Reporting only the tests that came out significant. Say how many tests were run; with many tests, some false alarms are expected.
- Trusting a P-value or interval computed from a voluntary response sample, a convenience sample or a confounded comparison. Inference measures only chance variation, not bias.

ANSWERS TO PRACTICE

1. (a) 4 is outside the interval: reject H_0. (b) 3.5 is inside the interval: fail to reject H_0.
2. $100(1 - 0.01)\% = 99\%$
3. $40(0.05) = 2$
4. It means the P-value is less than 0.05: if the null hypothesis were true, a result as extreme as the one observed would happen less than 5% of the time, so the result is too unusual to be explained comfortably by chance alone. It does not mean the effect is large or important.

5. (a) Standard error $= 1.8/\sqrt{9} = 0.6$; $t^* = 2.306$ (df = 8); $57.0 \pm 2.306(0.6) = 57.0 \pm 1.3836 = (55.62, 58.38)$ grams. (b) 58 is inside the interval, so fail to reject H_0: there is not convincing evidence that the mean weight differs from 58 grams. (Check: $t = (57.0 - 58)/0.6 = -1.67$ and $|t| < 2.306$.)
6. (a) $\alpha = 1 - 0.90 = 0.10$. (b) 6.2 is outside the interval, so reject H_0 at $\alpha = 0.10$: there is convincing evidence that the mean cabbage weight is not 6.2 pounds.
7. The difference is statistically significant: 5,000 drives make even a small difference stand out from chance. But the interval shows the true increase is only about 0.5 to 1.9 yards on drives of a couple hundred yards, too small for an ordinary golfer to notice. "Significantly" in the statistical sense does not mean "by a lot," and "proven" overstates what a test shows.
8. (a) 0.80 is outside the interval, so reject H_0 (convincing evidence her true rate is not 80%). (b) The interval gives the size of the shortfall: her true rate is plausibly between 62% and 74%, about 6 to 18 percentage points below the coach's claim.
9. This is a voluntary response sample: the callers chose themselves, and people with strong opinions (or who were listening that day) are over-represented. The interval accounts only for random sampling variation, not for this bias, so a large sample and a narrow interval do not make the estimate trustworthy.
10. (a) $\hat{p} = 52/400 = 0.13$; standard deviation under $H_0 = \sqrt{0.10(0.90)/400} = 0.015$; $z = \dfrac{0.13 - 0.10}{0.015} = 2.00$. (b) P-value $= 2P(Z \geq 2.00) = 0.0455$. (c) Standard error $= \sqrt{0.13(0.87)/400} = 0.016815$; $0.13 \pm 1.96(0.016815) = 0.13 \pm 0.0330 = (0.097, 0.163)$. (d) The test uses the standard deviation computed from $p_0 = 0.10$ (0.015), while the interval uses the standard error computed from $\hat{p} = 0.13$ (0.0168). With different spreads, a borderline case like this one (P-value just under 0.05, 0.10 just inside the interval) can come out differently. For proportions the two methods agree nearly always, not always.
11. (a) 40 is outside the interval: reject H_0. (b) 46 is inside the interval: fail to reject H_0. (c) The interval is centered at the sample mean: $\bar{x} = \dfrac{41.2 + 46.8}{2} = 44.0$ inches.
12. (a) Standard error $= 0.25/\sqrt{2{,}500} = 0.005$; $t = \dfrac{16.01 - 16}{0.005} = 2.00$. (b) P-value $= 2P(T \geq 2.00) = 0.0456$ (normal table: $2(0.0228) = 0.0456$). The result is significant at $\alpha = 0.05$. (c) No: the jars average about one-hundredth of an ounce MORE than the label, which no customer could notice and which costs the packer almost nothing. The huge sample makes a trivial difference statistically significant.
13. (a) $20(0.05) = 1$. (b) $1 - 0.95^{20} = 1 - 0.3585 = 0.6415$. (c) Even if no remedy works, there is about a 64% chance that at least one of 20 tests will be significant by chance alone. Reporting only the "winner" hides the 19 failures, so readers cannot tell that the result is just what chance would usually produce. An honest report says how many remedies were tested, and a significant result should be checked with a new experiment.

Unit 11 test

Time: 60 minutes. You may use a scientific or graphing calculator and the normal and t tables at the back of the course. Show your work; for a complete significance test, show all four steps (State, Plan, Do, Conclude) and end with a conclusion in context. Unless a problem says otherwise, round z and t statistics to 2 decimal places and probabilities and P-values to 4 decimal places, using technology values when you can (answers from the tables may differ in the last digit). The test is worth 100 points.

1. **5 PTS** A seed packet says that its bush-bean seeds sprout in a mean of 8 days. Ruth suspects the beans take longer than that in her cool spring garden. Define the parameter and state H_0 and H_a in symbols and words.

2. **4 PTS** A significance test gives a P-value of 0.032. What is the decision (a) at the $\alpha = 0.05$ level and (b) at the $\alpha = 0.01$ level?

3. **6 PTS** A beekeeper tests whether a new hive design increases the mean honey harvest above 60 pounds per hive ($H_0 : \mu = 60$, $H_a : \mu > 60$) and gets a P-value of 0.12. (a) Interpret this P-value in context. (b) She writes in her notebook, "The new design does not increase the honey harvest." Explain what is wrong with her conclusion.

4. **6 PTS** Samuel claims he can tell well water from city water by taste. In 16 trials, each with one cup of each in random order, he picks the well water correctly 13 times. The table shows 100 simulated sets of 16 pure guesses. (a) Estimate the P-value for $H_0 : p = 0.5$ against $H_a : p > 0.5$, where p is the proportion of all such trials Samuel gets right. (b) What do you conclude at $\alpha = 0.05$?

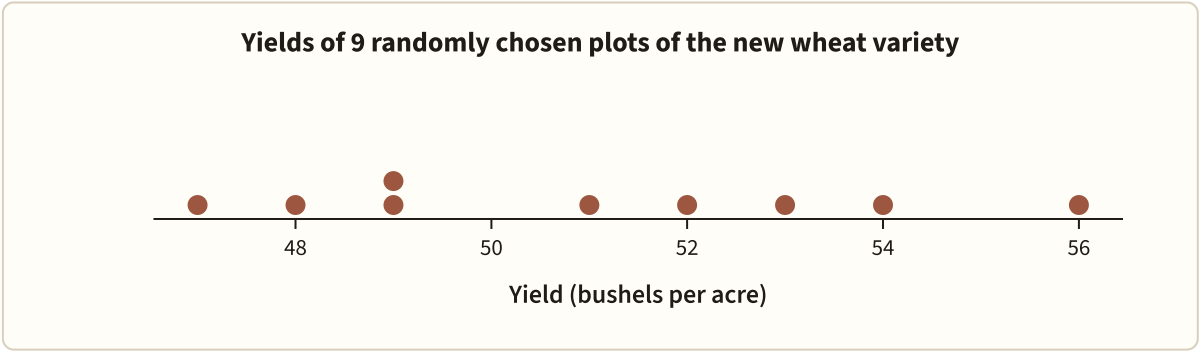
100 simulated sets of 16 pure guesses (each guess right with probability 0.5)

Number correct (out of 16)	Number of simulated sets
4 or fewer	3
5	7
6	12
7	17
8	20
9	17
10	12
11	7
12	3
13	1
14 or more	1
Total	100

5. **10 PTS** A tree nursery says that 60% of its oak seedlings survive their first winter. A forester plants a random sample of 600 of the nursery's seedlings with a new mulch, and 381 survive the winter. Is there convincing evidence at the $\alpha = 0.05$ level that more than 60% of mulched seedlings survive? Use the four-step process.

6. **5 PTS** A poultry farmer wants to test the claim that 5% of her eggs have double yolks, using a random sample of 80 eggs. Is the Large Counts condition for a one-sample z test met? Explain.

7. **10 PTS** The mean wheat yield in a county is 48 bushels per acre. A farmer plants a new variety on a random sample of 9 plots, with these yields in bushels per acre: 56, 47, 54, 48, 53, 49, 49, 52, 51. Is there convincing evidence at the $\alpha = 0.01$ level that the new variety's mean yield is greater than 48 bushels per acre? Use the four-step process.



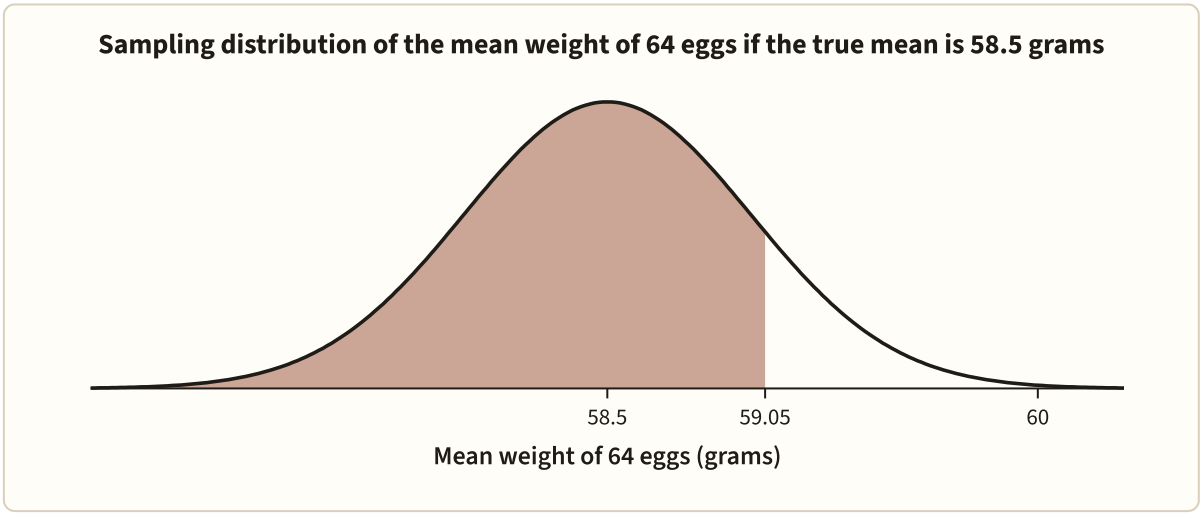
8. **6 PTS** A carpenter tests whether the mean thickness of a shipment of boards differs from the stated 0.75 inch, using a random sample of 9 boards. The test of $H_0 : \mu = 0.75$ against $H_a : \mu \neq 0.75$ gives $t = -1.50$. (a) Use the t table to give bounds for the P-value. (b) The technology P-value is one of 0.0860 or 0.1720. Which is it, and why?

Part of the t table: critical values t* with upper-tail probability p

df	p = 0.20	0.10	0.05	0.025	0.01	0.005
7	0.896	1.415	1.895	2.365	2.998	3.499
8	0.889	1.397	1.860	2.306	2.896	3.355
9	0.883	1.383	1.833	2.262	2.821	3.250

9. **6 PTS** A lumberyard tests each kiln load of boards with $H_0 : \mu = 19$ against $H_a : \mu > 19$, where μ is the mean moisture content (percent) of the boards; boards above 19% are too wet to sell for building. (a) Describe a Type I error and its consequence. (b) Describe a Type II error and its consequence.

10. **9 PTS** An egg farm tests $H_0 : \mu = 60$ against $H_a : \mu < 60$, where μ is the mean weight (grams) of its jumbo eggs. Egg weights have $\sigma = 4$ grams, and each day 64 eggs are weighed. The rule is to adjust the grading machine if $\bar{x} \le 59.05$ grams. (a) Find the probability of a Type I error. (b) Find the power of the test if the true mean is 58.5 grams. (c) Find β for that alternative.



11. **5 PTS** For the egg farm in Problem 10, describe two different changes that would increase the power of the test against $\mu = 58.5$ grams, and explain why each one works.

12. **8 PTS** A 95% confidence interval for the mean weight of the pumpkins at a community pumpkin patch is (14.2, 17.8) pounds. (a) Decide the two-sided test of $H_0 : \mu = 15$ at $\alpha = 0.05$. (b) Decide the two-sided test of $H_0 : \mu = 18$ at $\alpha = 0.05$. (c) Explain what the interval tells you that the test decisions alone do not.

13. **9 PTS** A board-game spinner is supposed to land on red half the time. A game company spins a randomly chosen spinner 10,000 times, and it lands on red 5,120 times. (a) For $H_0 : p = 0.5$ against $H_a : p \ne 0.5$, find z. (b) Find the P-value. (c) Construct a 95% confidence interval for p (3 decimal places). (d) Is the departure from 0.5 important for players of the game? Explain.

14. **6 PTS** A 4-H club tests 15 different gardening tips, each with its own independent experiment at $\alpha = 0.05$. Suppose none of the tips actually works. (a) What is the expected number of significant results? (b) Find the probability that at least one result is significant. (c) Why would it be misleading to publish only the tips that came out significant?

15. **5 PTS** A church newsletter asks readers to mail in a card if they would like the summer potluck moved to Saturday. Of the 60 cards returned, 52 favor Saturday, and a member computes a P-value of less than 0.0001 for $H_0 : p = 0.5$. Does this show that most members of the congregation prefer Saturday? Explain.

UNIT 11 TEST: ANSWER KEY

1. Let μ = the mean number of days for bush-bean seeds of this kind to sprout in Ruth's garden. $H_0 : \mu = 8$ (the beans sprout in a mean of 8 days); $H_a : \mu > 8$ (they take longer than 8 days on average).
2. (a) $0.032 < 0.05$: reject H_0. (b) $0.032 \geq 0.01$: fail to reject H_0.
3. (a) If the new design really produced a mean harvest of 60 pounds per hive, there would be about a 12% chance of getting a sample mean as large as or larger than hers just by chance. (b) A large P-value means only that her data are not convincing evidence of an increase; it does not show there is no increase. The design might help, and a larger study might detect it. She should write "there is not convincing evidence that the new design increases the mean harvest."
4. (a) 13 or more correct: $1 + 1 = 2$ of 100, so P-value ≈ 0.02. (b) $0.02 < 0.05$, so reject H_0: there is convincing evidence that Samuel does better than guessing.
5. State: $H_0 : p = 0.60$, $H_a : p > 0.60$, where p is the proportion of all the nursery's seedlings that would survive the first winter with the new mulch; $\alpha = 0.05$. Plan: one-sample z test for p; random sample; 600 is less than 10% of the nursery's seedlings; $600(0.60) = 360 \geq 10$ and $600(0.40) = 240 \geq 10$. Do: $\hat{p} = 381/600 = 0.635$; standard deviation $= \sqrt{0.60(0.40)/600} = \sqrt{0.0004} = 0.02$; $z = \dfrac{0.635 - 0.60}{0.02} = 1.75$; P-value $= P(Z \geq 1.75) = 0.0401$. Conclude: $0.0401 < 0.05$, reject H_0; there is convincing evidence that more than 60% of mulched seedlings survive their first winter.
6. No. $np_0 = 80(0.05) = 4$, which is less than 10, so the sampling distribution of $\hat{p}$ would be skewed and a normal P-value would not be trustworthy.
7. State: $H_0 : \mu = 48$, $H_a : \mu > 48$, where μ is the mean yield (bushels per acre) of the new variety on plots like these; $\alpha = 0.01$. Plan: one-sample t test; random sample of plots; 9 is less than 10% of all such plots; $n < 30$ but the dotplot shows no strong skewness or outliers. Do: $\bar{x} = 459/9 = 51.0$; squared deviations add to 72, so $s = \sqrt{72/8} = 3.0$; standard error $= 3.0/\sqrt{9} = 1.0$; $t = \dfrac{51.0 - 48}{1.0} = 3.00$, df $= 8$; P-value $= P(T \geq 3.00) = 0.0085$ (table: between 0.005 and 0.01). Conclude: $0.0085 < 0.01$, reject H_0; there is convincing evidence that the new variety's mean yield is greater than 48 bushels per acre.
8. (a) df $= 8$; $|t| = 1.50$ is between 1.397 (0.10) and 1.860 (0.05), so one tail is between 0.05 and 0.10, and the two-sided P-value is between 0.10 and 0.20. (b) 0.1720: the alternative is two-sided, so the P-value is twice the one-tail area of 0.0860.
9. (a) Type I: concluding that the load's mean moisture is above 19% when it really is not; the yard would dry good boards again (or hold them back), wasting time and fuel. (b) Type II: failing to find that the mean moisture is above 19% when it really is; the yard would sell boards that are too wet, which may warp or shrink after builders use them.

10. Standard deviation of $\bar{x} = 4/\sqrt{64} = 0.5$. (a) $z = \dfrac{59.05 - 60}{0.5} = -1.90$; $P(\text{Type I}) = P(Z \le -1.90) = 0.0287$. (b) $z = \dfrac{59.05 - 58.5}{0.5} = 1.10$; power $= P(Z \le 1.10) = 0.8643$. (c) $\beta = 1 - 0.8643 = 0.1357$.
11. Any two of: (1) Weigh more eggs each day; a larger n shrinks the standard deviation $\sigma/\sqrt{n}$ of $\bar{x}$, so the distributions under $\mu = 60$ and $\mu = 58.5$ overlap less. (2) Use a larger α (a cutoff closer to 60); the rejection region grows, though Type I errors become more likely. (3) Reduce the variability of the measurements (a more precise scale, more uniform eggs), which also shrinks the spread of $\bar{x}$.
12. (a) 15 is inside the interval: fail to reject H_0. (b) 18 is outside the interval: reject H_0. (c) The interval gives the whole range of plausible values for the mean (14.2 to 17.8 pounds), showing how large or small the mean could reasonably be, while a test only says whether one particular value is plausible.
13. (a) $\hat{p} = 5{,}120/10{,}000 = 0.512$; standard deviation $= \sqrt{0.5(0.5)/10{,}000} = 0.005$; $z = \dfrac{0.512 - 0.5}{0.005} = 2.40$. (b) P-value $= 2P(Z \ge 2.40) = 0.0164$, significant at $\alpha = 0.05$. (c) $0.512 \pm 1.96\sqrt{0.512(0.488)/10{,}000} = 0.512 \pm 1.96(0.0049986) = 0.512 \pm 0.0098 = (0.502, 0.522)$. (d) No: the spinner lands on red between about 50.2% and 52.2% of the time, so players would not notice any difference in a game. The huge number of spins makes a tiny departure statistically significant.
14. (a) $15(0.05) = 0.75$. (b) $1 - 0.95^{15} = 1 - 0.4633 = 0.5367$. (c) There is better than a 50% chance that at least one useless tip will look significant by chance alone. Publishing only the "winners" hides how many tests were run, so readers would mistake chance results for real effects.
15. No. The cards are a voluntary response sample: members who care strongly (most likely those who want the change) are the ones who mail cards. The P-value measures only chance variation in a random sample; it cannot correct this bias. A random sample of members would be needed.

UNIT 12

Comparing Two Groups

Most of the questions people really care about are comparisons. Does this seed lot germinate better than that one? Do lambs gain more on clover than on grass? Did the new ration raise each cow's milk yield? In this unit you extend the confidence intervals of Unit 10 and the significance tests of Unit 11 from one group to two: two proportions (Lesson 12.1), two independent means (Lesson 12.2), and paired data, where each measurement has a partner (Lesson 12.3). Lesson 12.4 pulls everything together into a decision guide for choosing the right procedure.

The reasoning is the same as before: a statistic varies from sample to sample, and its standard error measures by how much. What is new is WHY the standard errors look the way they do: for two independent groups the variances add (Lesson 7.3), for a test of two proportions we pool because the null hypothesis says the groups are alike, and for paired data we subtract within pairs to remove the variation between subjects. Throughout, the course uses the conservative df (the smaller of $n_1 - 1$ and $n_2 - 1$) for two-sample t procedures, so every answer can be found with a scientific calculator and the printed tables.

For parents: this unit is the payoff of the inference units, because it is how real experiments on farms, in kitchens and on playing fields are judged. Encourage your student to say out loud which procedure fits and why before calculating, and to finish every problem with a sentence in context. Unit 13 then adds chi-square tests for categorical variables with more than two categories.

PREREQUISITES

- Mean and standard deviation of a sum or difference of independent random variables: variances add (Lesson 7.3).
- Sampling distributions of $\hat{p}$ and $\bar{x}$ (Lessons 9.2 and 9.3).
- One-sample z interval for a proportion and one-sample t interval for a mean, with the four-step process (Lessons 10.2 and 10.4).
- Reading the t table for t^* and for bounds on a P-value (Lessons 10.3 and 11.3).
- Significance tests: hypotheses, P-values, decisions at level α, and conclusions in context (Lessons 11.1 to 11.3).

MASTERY CHECKLIST

☐ I can describe the sampling distribution of $\hat{p}_1 - \hat{p}_2$ and explain why its variance is the sum of two variances.

☐ I can construct and interpret a two-proportion z interval using the unpooled standard error.

☐ I can carry out a two-proportion z test using the pooled proportion $\hat{p}_c$, and explain why we pool.

☐ I can compute the standard error $\sqrt{s_1^2/n_1 + s_2^2/n_2}$ and the conservative df for two-sample t procedures.

☐ I can construct and interpret a two-sample t interval and carry out a two-sample t test in context.

- Random sampling versus random assignment and the scope of conclusions (Lesson 5.6).

☐ I can recognize paired data and analyze the differences with a paired t interval or test.
☐ I can explain why pairing removes variation between subjects and why treating paired data as independent samples is a mistake.
☐ I can check the Random, 10% and Large Counts or Normal/Large Sample conditions for each two-group procedure.
☐ I can choose the right procedure by asking: proportion or mean? one sample, two independent samples, or paired? interval or test?
☐ I can state whether a conclusion shows cause and effect or only generalizes to a population, based on how the data were collected.

LESSON 12.1 · 2 SESSIONS OF 45 MINUTES

Comparing Two Proportions

OBJECTIVES

- Describe the sampling distribution of $\hat{p}_1 - \hat{p}_2$ and explain why its variance is the sum of two variances.
- Construct and interpret a two-proportion z interval for $p_1 - p_2$ in context.
- Carry out a two-proportion z test using the pooled proportion $\hat{p}_c$, and explain why we pool when we assume H_0.
- Check the Random, 10% and Large Counts conditions for two-proportion procedures and state the scope of the conclusion.

VIDEO

WATCH

An Introduction to Inference for Two Proportions

jbstatistics · 15:10

When to watch: Before Example 1: the two-proportion z interval and test.

https://www.youtube.com/watch?v=g0at6LpYvHc

BACKUP

Inference for Two Proportions: An Example of a Confidence Interval and a Hypothesis Test

jbstatistics · 13:23

When to watch: After Example 2: an interval and a test worked through (homing pigeons).

https://www.youtube.com/watch?v=OIYkOiQX3fk

> **Note:** Like the course, both videos pool the two samples for the test but not for the interval.

KEY TERMS

Two-sample problem	A study that compares two populations or two treatments using two separate, independent groups (two random samples, or two groups formed by random assignment).
Difference of proportions	The parameter $p_1 - p_2$, the true proportion of successes in group 1 minus the true proportion in group 2. Its estimate is $\hat{p}_1 - \hat{p}_2$.

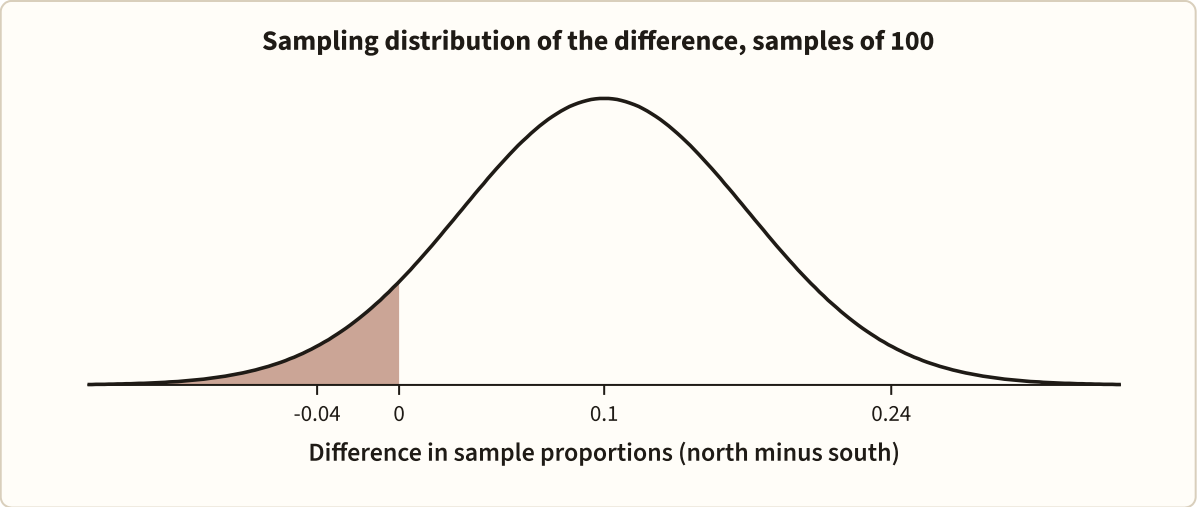
Pooled (combined) proportion	$\hat{p}_c = \dfrac{X_1 + X_2}{n_1 + n_2}$: all the successes in both samples divided by all the individuals in both samples. It estimates the common proportion when $H_0 : p_1 = p_2$ is true.
Two-proportion z interval	A confidence interval for $p_1 - p_2$: the observed difference plus or minus z^* times the unpooled standard error.
Two-proportion z test	A significance test of $H_0 : p_1 = p_2$ that measures how many standard errors the observed difference $\hat{p}_1 - \hat{p}_2$ is from 0, using the pooled standard error.

CORE IDEAS

Much of the most useful statistics is comparison. Does lot A of seed germinate better than lot B? Does a new spray reduce blight on tomatoes? Does one free-throw routine work better than another? Each question asks about two proportions, p_1 and p_2, and the natural parameter is their difference $p_1 - p_2$. Our estimate is the difference of the two sample proportions, $\hat{p}_1 - \hat{p}_2$. Always say in words which group is 1 and which is 2, and keep the same order of subtraction from start to finish.

How much does $\hat{p}_1 - \hat{p}_2$ vary from sample to sample? Its center is $p_1 - p_2$, because each $\hat{p}$ is centered on its own parameter, so the difference is an unbiased estimator. For the spread, recall from Lesson 7.3 that variances ADD for independent random variables, even when we subtract. Each sample brings its own chance error, and subtracting does not cancel them; the errors can pile up in either direction. So the variance of the difference is $\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}$, and the standard deviation is its square root. When the samples are large, the shape is approximately normal.

Here is a picture. Suppose that in truth 50% of the apples in the north orchard and 40% in the south orchard are over 3 inches across, and we pick 100 apples at random from each. Then $\hat{p}_N - \hat{p}_S$ has mean $0.50 - 0.40 = 0.10$ and standard deviation $\sqrt{\dfrac{0.5(0.5)}{100} + \dfrac{0.4(0.6)}{100}} = \sqrt{0.0025 + 0.0024} = \sqrt{0.0049} = 0.07$. The shaded area is the chance that the samples point the wrong way and show the south orchard ahead: $z = \dfrac{0 - 0.10}{0.07} = -1.43$, an area of about 0.077. A real difference of 10 percentage points can hide in samples of 100 almost 8% of the time.



To ESTIMATE $p_1 - p_2$ we use the familiar form statistic $\pm$ (critical value)(standard error). We do not know p_1 and p_2, so the standard error uses $\hat{p}_1$ and $\hat{p}_2$ in their places. The interval tells us how big the difference might be. If the whole interval lies above 0, the data give convincing evidence that p_1 is larger; if the interval contains 0, a difference of zero is plausible.

To TEST we begin with $H_0 : p_1 = p_2$ (no difference). A P-value is computed assuming H_0 is true, just as Lesson 11.2 used p_0 in the standard deviation. If H_0 is true, the two groups share ONE common proportion, so the best estimate of it uses all the data at once: the pooled proportion $\hat{p}_c = \dfrac{X_1 + X_2}{n_1 + n_2}$. Pooling is not a trick; it is simply being consistent with the assumption that both groups are alike. The test statistic z then counts how many standard errors the observed difference lies from 0.

Conditions. Random: the data come from two independent random samples or from two groups formed by random assignment in an experiment. 10%: when sampling without replacement, each sample is at most 10% of its population (not needed for random assignment). Large Counts: for an interval, the four counts $n_1\hat{p}_1$, $n_1(1-\hat{p}_1)$, $n_2\hat{p}_2$, $n_2(1-\hat{p}_2)$ (the observed successes and failures) are all at least 10; for a test, use $n_1\hat{p}_c$, $n_1(1-\hat{p}_c)$, $n_2\hat{p}_c$, $n_2(1-\hat{p}_c)$.

Scope. Random assignment lets a small P-value show that the treatment CAUSED the difference. Two random samples from two populations let us generalize to those populations but, with no assignment, they do not show cause: the groups may differ in other ways.

KEY FORMULAS

$$\mu_{\hat{p}_1-\hat{p}_2} = p_1 - p_2 \qquad \sigma_{\hat{p}_1-\hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

$$\text{Interval: } (\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

$$\hat{p}_c = \frac{X_1 + X_2}{n_1 + n_2}$$

$$\text{Test of } H_0 : p_1 = p_2 : \quad z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}$$

Practice habit. Define p_1 and p_2 in words before you calculate, and keep the same order of subtraction in the hypotheses, the statistic and the conclusion. Remember: pool for the test, do not pool for the interval.

WORKED EXAMPLES

Example 1. Ruth buys carrot seed from two lots. She plants a random sample of 200 seeds from lot A and 172 germinate; she plants a random sample of 250 seeds from lot B under the same conditions and 195 germinate. Construct and interpret a 95% confidence interval for $p_A - p_B$, the difference in the true germination proportions. Round the endpoints to 3 decimal places.

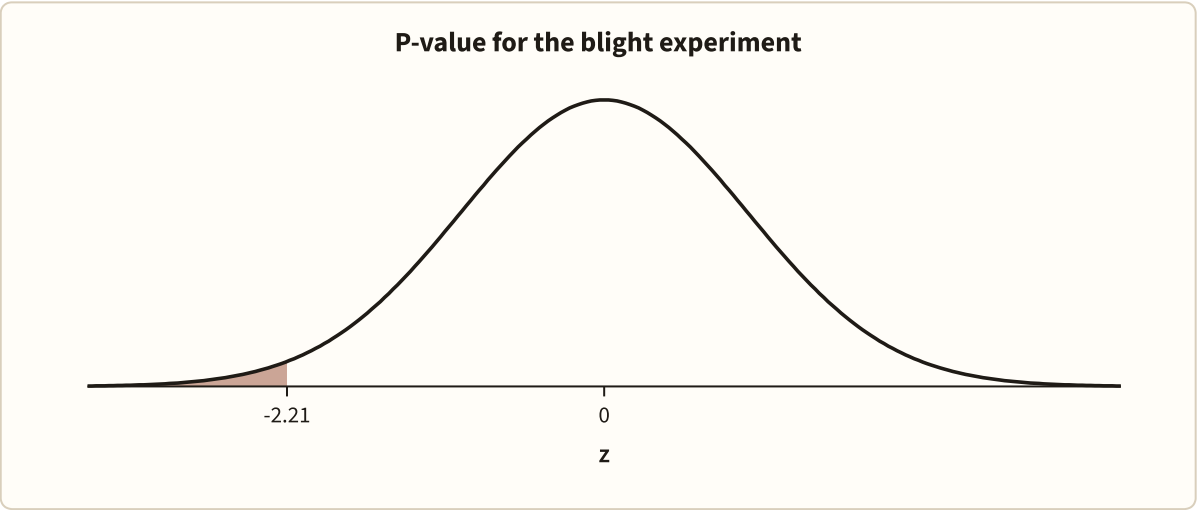
Carrot seed germination

Seed lot	Germinated	Did not germinate	Total
Lot A	172	28	200
Lot B	195	55	250
Total	367	83	450

1. State: estimate $p_A - p_B$, where p_A and p_B are the proportions of ALL seeds in lot A and lot B that germinate, with 95% confidence.
2. Plan: two-proportion z interval. Random: independent random samples from each lot. 10%: each lot holds many thousands of seeds, far more than 10 times 200 or 250. Large Counts: 172, 28, 195 and 55 are all at least 10.
3. Do: $\hat{p}_A = \dfrac{172}{200} = 0.86$ and $\hat{p}_B = \dfrac{195}{250} = 0.78$, so $\hat{p}_A - \hat{p}_B = 0.08$.
4. Standard error: $SE = \sqrt{\dfrac{0.86(0.14)}{200} + \dfrac{0.78(0.22)}{250}} = \sqrt{0.000602 + 0.0006864} = \sqrt{0.0012884} \approx 0.035894$.
5. Margin of error: $1.960 \times 0.035894 \approx 0.070352$.
6. Interval: 0.08 ± 0.070352, which is $(0.009648, 0.150352) \approx (0.010, 0.150)$.
7. Conclude: We are 95% confident that the germination rate of all lot A seeds is between 1.0 and 15.0 percentage points higher than that of lot B. Because 0 is not in the interval, the data give convincing evidence that lot A germinates better.

Answer: $(0.010, 0.150)$. We are 95% confident that the true germination rate for lot A is between 1.0 and 15.0 percentage points higher than for lot B.

Example 2. A garden club randomly assigned 300 tomato seedlings to two groups: 150 were sprayed weekly with compost tea and 150 were not sprayed. By midsummer, 21 sprayed plants and 36 unsprayed plants showed early blight. Do the data give convincing evidence at the $\alpha = 0.05$ level that the spray reduces the proportion of plants that get blight?



1. State: $H_0 : p_1 = p_2$ and $H_a : p_1 < p_2$, where p_1 and p_2 are the true proportions of plants like these that would get blight when sprayed and when not sprayed. $\alpha = 0.05$.
2. Plan: two-proportion z test. Random: the seedlings were randomly assigned to the two treatments. (The 10% condition is not needed for random assignment.)
3. Pooled proportion: $\hat{p}_c = \dfrac{21 + 36}{150 + 150} = \dfrac{57}{300} = 0.19$.
4. Large Counts: $150(0.19) = 28.5$ and $150(0.81) = 121.5$ are both at least 10, for each group.
5. Do: $\hat{p}_1 = \dfrac{21}{150} = 0.14$, $\hat{p}_2 = \dfrac{36}{150} = 0.24$, difference -0.10.
6. Standard error: $\sqrt{0.19(0.81)\left(\dfrac{1}{150} + \dfrac{1}{150}\right)} = \sqrt{0.1539 \times 0.013333} = \sqrt{0.002052} \approx 0.045299$.
7. $z = \dfrac{-0.10 - 0}{0.045299} \approx -2.21$.
8. P-value $= P(Z < -2.21) \approx 0.0136$ (table: 0.0136).
9. Conclude: Because $0.0136 < 0.05$, we reject H_0. There is convincing evidence that compost-tea spray reduces the proportion of tomato plants like these that get early blight. Because the treatments were randomly assigned, we may conclude the spray caused the reduction.

Answer: $z \approx -2.21$, P-value ≈ 0.0136; reject H_0. There is convincing evidence that the spray reduces the blight rate.

PRACTICE

1. EASY In a random sample of 120 apples from the north orchard, 18 had blemishes. In a random sample of 150 apples from the south orchard, 30 had blemishes. Find (a) $\hat{p}_N$, (b) $\hat{p}_S$, and (c) $\hat{p}_N - \hat{p}_S$.

2. **EASY** Using the apple data in Problem 1 (18 of 120 and 30 of 150 blemished), find the pooled proportion $\hat{p}_c$ that a test of $H_0 : p_N = p_S$ would use. Round to 4 decimal places.

3. **EASY** In Field A, 30% of the pumpkins weigh more than 20 pounds; in Field B, 20% do. A farmer picks a random sample of 100 pumpkins from each field. Find the mean and standard deviation of the sampling distribution of $\hat{p}_A - \hat{p}_B$. Round the standard deviation to 4 decimal places.

4. **EASY** Of 40 duck eggs incubated at a higher temperature, 6 failed to hatch; of 50 eggs at the usual temperature, 9 failed to hatch. A student wants to run a two-proportion z test on the failure rates. Is the Large Counts condition met? Answer yes or no and show the check.

5. **MEDIUM** From independent random samples of households in two towns, a 95% confidence interval for $p_A - p_B$, the difference in the proportions of households that keep backyard chickens, is $(-0.04, 0.09)$. (a) Interpret the interval in context. (b) Does the interval give convincing evidence that the two proportions differ? Answer yes or no.

6. **MEDIUM** In a random sample of 150 boards from Sawmill A, 27 had serious knots; in a random sample of 200 boards from Sawmill B, 24 had serious knots. Construct a 90% confidence interval for $p_A - p_B$. Round the endpoints to 3 decimal places.

7. **MEDIUM** In a random sample of 200 families in County A, 124 keep a vegetable garden; in a random sample of 200 families in County B, 106 do. Test $H_0 : p_A = p_B$ against $H_a : p_A \neq p_B$ at the $\alpha = 0.05$ level. Give the z statistic (2 decimal places), the P-value (4 decimal places) and your decision.

8. **MEDIUM** Explain why the two-proportion z test uses the pooled proportion $\hat{p}_c$ in its standard error, while the two-proportion z interval uses $\hat{p}_1$ and $\hat{p}_2$ separately.

9. **MEDIUM** An extension agent takes a random sample of farms in a county and finds that the proportion of farms with high tomato yields is significantly larger (P-value = 0.003) among farms that use drip irrigation than among farms that do not. Can the agent conclude that drip irrigation causes higher yields? Answer yes or no and explain.

10. HARD Coach Miller randomly assigned 160 players at a summer camp to learn one of two free-throw routines, 80 to each. After a week, each player shot 10 free throws, and the coach recorded whether the player made at least 7 of the 10. The results are in the table. Do the data give convincing evidence at the $\alpha = 0.05$ level that routine A leads to a higher proportion of players making at least 7? Use the four-step process; give z to 2 decimal places and the P-value to 4 decimal places.

Free-throw routine experiment

Routine	Made at least 7 of 10	Made fewer than 7	Total
Routine A	52	28	80
Routine B	40	40	80
Total	92	68	160

11. HARD A poultry farmer randomly assigned 400 fertile eggs to two incubator settings, 200 to each. The results are in the table. (a) Construct a 99% confidence interval for $p_A - p_B$, the difference in true hatching proportions. Round the endpoints to 3 decimal places. (b) Does the interval give convincing evidence that the settings have different hatching rates? Answer yes or no.

Incubator experiment

Setting	Hatched	Did not hatch	Total
Setting A	176	24	200
Setting B	158	42	200
Total	334	66	400

12. HARD A random sample of 150 days from Weather Station A's records shows that its rain forecast was correct on 123 days; a random sample of 250 days from Station B's records shows 190 correct. Test at the $\alpha = 0.05$ level whether the two stations' true proportions of correct forecasts differ. Give z (2 decimal places), the P-value (4 decimal places) and your decision.

13. CHALLENGE Return to Problem 7 (62% of 200 families garden in County A and 53% of 200 in County B, $z \approx 1.82$). Suppose instead that the samples had 800 families each, with the same sample proportions 0.62 and 0.53. (a) Find the new z statistic (2 decimal places). (b) Find the new two-sided P-value (4 decimal places). (c) Explain why the same difference now gives a different conclusion.

14. CHALLENGE Grace says, "When we subtract $\hat{p}_2$ from $\hat{p}_1$, the chance errors should partly cancel, so the standard deviation of $\hat{p}_1 - \hat{p}_2$ ought to be SMALLER than the standard deviation of either one." Explain what is wrong with her reasoning.

COMMON MISTAKES

- Using the pooled proportion in a confidence interval, or the unpooled standard error in a test. Pool only for the test, because only the test assumes $p_1 = p_2$.
- Switching the order of subtraction partway through (for example, $H_a : p_1 > p_2$ but computing $\hat{p}_2 - \hat{p}_1$). Write down which group is 1 and keep that order everywhere.
- Adding the two standard deviations instead of the two variances. Add the variances under one square root.
- Claiming cause and effect from two random samples of existing groups. Only random assignment of the treatments supports a cause-and-effect conclusion.

ANSWERS TO PRACTICE

1. (a) $\frac{18}{120} = 0.15$ (b) $\frac{30}{150} = 0.20$ (c) $0.15 - 0.20 = -0.05$
2. $\hat{p}_c = \frac{18 + 30}{120 + 150} = \frac{48}{270} \approx 0.1778$
3. Mean $= 0.30 - 0.20 = 0.10$; SD $= \sqrt{\frac{0.3(0.7)}{100} + \frac{0.2(0.8)}{100}} = \sqrt{0.0037} \approx 0.0608$
4. No. $\hat{p}_c = \frac{15}{90} \approx 0.1667$, and $40(0.1667) \approx 6.7$ and $50(0.1667) \approx 8.3$ are less than 10.
5. (a) We are 95% confident that the proportion of households in Town A that keep backyard chickens is somewhere between 4 percentage points lower and 9 percentage points higher than in Town B. (b) No: 0 is inside the interval, so no difference is a plausible value.
6. $\hat{p}_A = 0.18$, $\hat{p}_B = 0.12$; $SE = \sqrt{\frac{0.18(0.82)}{150} + \frac{0.12(0.88)}{200}} = \sqrt{0.001512} \approx 0.038884$; $0.06 \pm 1.645(0.038884) = 0.06 \pm 0.063964$, so $(-0.004, 0.124)$.
7. $\hat{p}_A = 0.62$, $\hat{p}_B = 0.53$, $\hat{p}_c = \frac{230}{400} = 0.575$; $SE = \sqrt{0.575(0.425)\left(\frac{2}{200}\right)} \approx 0.049434$; $z = \frac{0.09}{0.049434} \approx 1.82$; P-value ≈ 0.0687 (table with $z = 1.82$: 0.0688). Since $0.0687 > 0.05$, fail to reject H_0: there is not convincing evidence that the proportions of families who garden differ in the two counties.
8. A P-value is computed assuming $H_0 : p_1 = p_2$ is true. If the two groups share one common proportion, the best estimate of it combines all the data, so the test pools. An interval makes no such assumption; it is trying to estimate how far apart p_1 and p_2 are, so it must let them differ and uses each sample's own proportion.
9. No. This is an observational study: the farmers chose whether to use drip irrigation. Farms that use it may differ in other ways (more experience, better soil, more money for fertilizer) that could explain the higher yields. Only random assignment of the treatment would let us conclude cause.

10. State: $H_0 : p_A = p_B$, $H_a : p_A > p_B$, where p is the true proportion of players like these who would make at least 7 of 10 with each routine; $\alpha = 0.05$. Plan: two-proportion z test; random assignment; $\hat{p}_c = \dfrac{92}{160} = 0.575$, and $80(0.575) = 46$ and $80(0.425) = 34$ are at least 10. Do: $\hat{p}_A = 0.65$, $\hat{p}_B = 0.50$; $SE = \sqrt{0.575(0.425)\left(\dfrac{2}{80}\right)} \approx 0.078163$; $z = \dfrac{0.15}{0.078163} \approx 1.92$; P-value ≈ 0.0275 (table: 0.0274). Conclude: $0.0275 < 0.05$, so reject H_0. There is convincing evidence that routine A causes a higher proportion of players like these to make at least 7 of 10.

11. (a) $\hat{p}_A = 0.88$, $\hat{p}_B = 0.79$; $SE = \sqrt{\dfrac{0.88(0.12)}{200} + \dfrac{0.79(0.21)}{200}} = \sqrt{0.0013575} \approx 0.036844$; $0.09 \pm 2.576(0.036844) = 0.09 \pm 0.094911$, so $(-0.005, 0.185)$. (b) No: at 99% confidence, 0 is (just) inside the interval, so a difference of zero is still plausible.

12. $\hat{p}_A = 0.82$, $\hat{p}_B = 0.76$, $\hat{p}_c = \dfrac{313}{400} = 0.7825$; $SE = \sqrt{0.7825(0.2175)\left(\dfrac{1}{150} + \dfrac{1}{250}\right)} \approx 0.042607$; $z = \dfrac{0.06}{0.042607} \approx 1.41$; two-sided P-value ≈ 0.1591 (table with $z = 1.41$: 0.1586). Since $0.1591 > 0.05$, fail to reject H_0: there is not convincing evidence that the stations' forecast accuracy differs.

13. (a) $\hat{p}_c$ is still 0.575; $SE = \sqrt{0.575(0.425)\left(\dfrac{2}{800}\right)} \approx 0.024717$; $z = \dfrac{0.09}{0.024717} \approx 3.64$. (b) P-value ≈ 0.0003. (c) Four times the sample size cuts the standard error in half (it shrinks like $1/\sqrt{n}$), so the same 9-point difference is now twice as many standard errors from 0. With large samples, chance alone rarely produces a 9-point gap, so the difference becomes convincing.

14. The two samples are independent, so their chance errors are unrelated: $\hat{p}_1$ may come out too high while $\hat{p}_2$ comes out too low, which makes the difference even farther off. Errors do not reliably cancel; they can pile up. That is why the variances add, $\sigma^2 = \dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}$, and the difference is MORE variable than either proportion alone.

LESSON 12.2 · 2 SESSIONS OF 45 MINUTES

Comparing Two Means

OBJECTIVES

- Describe the sampling distribution of $\bar{x}_1 - \bar{x}_2$ and explain why its variance is the sum of the two variances.
- Compute the standard error $\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$ and the conservative degrees of freedom, and explain why the conservative choice is safe.
- Construct and interpret a two-sample t interval for $\mu_1 - \mu_2$ in context.
- Carry out a two-sample t test with the four-step process and state the scope of the conclusion.

VIDEO

WATCH

Welch (Unpooled Variance) t Tests and Confidence Intervals: Introduction

jbstatistics · 9:42

When to watch: Before Example 1: the two-sample t procedures.

https://www.youtube.com/watch?v=2-ecXltt2vI

BACKUP

Welch (Unpooled Variance) t Tests and Confidence Intervals: An Example

jbstatistics · 10:13

When to watch: After Example 1: a two-sample t interval and test worked through.

https://www.youtube.com/watch?v=gzrmHpA54Sc

> **Note:** The videos use the Welch–Satterthwaite degrees of freedom from software; the course uses the conservative df, the smaller of $n_1 - 1$ and $n_2 - 1$, which gives a slightly wider interval and a slightly larger P-value.

KEY TERMS

Difference of means — The parameter $\mu_1 - \mu_2$, the true mean of population (or treatment) 1 minus the true mean of population (or treatment) 2. Its estimate is $\bar{x}_1 - \bar{x}_2$.

Standard error of $\bar{x}_1 - \bar{x}_2$ — $SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$, our estimate from the data of how far $\bar{x}_1 - \bar{x}_2$ typically falls from $\mu_1 - \mu_2$.

Two-sample t interval — A confidence interval for $\mu_1 - \mu_2$: $(\bar{x}_1 - \bar{x}_2) \pm t^* \cdot SE$.

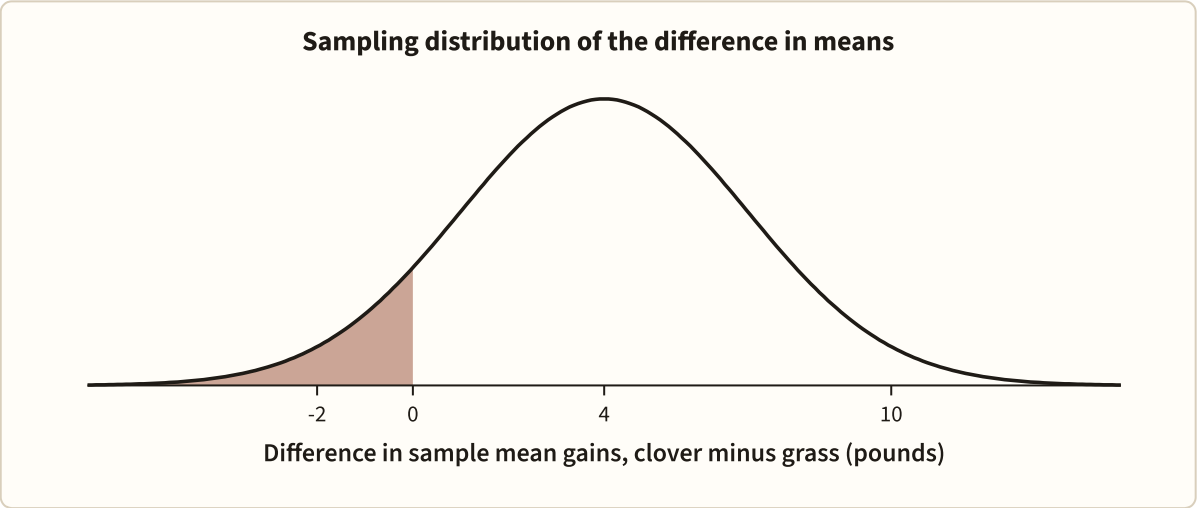
Two-sample t test	A significance test of $H_0 : \mu_1 = \mu_2$ using $t = \dfrac{\bar{x}_1 - \bar{x}_2}{SE}$.
Conservative degrees of freedom	In this course, df = the smaller of $n_1 - 1$ and $n_2 - 1$. It gives a slightly wider interval and a slightly larger P-value than technology, so it errs on the safe side.

CORE IDEAS

Lesson 12.1 compared two proportions; now we compare two MEANS. Does wheat variety A yield more bushels per acre than variety B? Do lambs gain more weight on a clover pasture than on a grass pasture? The parameter is $\mu_1 - \mu_2$ and the statistic is $\bar{x}_1 - \bar{x}_2$. As before, name group 1 and group 2 in words and keep the order of subtraction the same throughout.

The sampling distribution of $\bar{x}_1 - \bar{x}_2$ is centered at $\mu_1 - \mu_2$, since each sample mean is unbiased. Because the two samples are independent, their variances ADD (Lesson 7.3): each $\bar{x}$ has variance σ^2/n, so the difference has variance $\dfrac{\sigma_1^2}{n_1} + \dfrac{\sigma_2^2}{n_2}$. If both populations are normal, the difference is exactly normal; with large samples it is approximately normal by the central limit theorem.

For example, suppose lambs on a clover pasture gain on average 4 pounds more in a month than lambs on a grass pasture, with $\sigma_1 = 6$ pounds and $\sigma_2 = 10$ pounds. With 9 lambs on clover and 20 on grass, the standard deviation of $\bar{x}_1 - \bar{x}_2$ is $\sqrt{\dfrac{36}{9} + \dfrac{100}{20}} = \sqrt{4 + 5} = 3$ pounds. The shaded area, about 0.09, is the chance that the grass lambs come out ahead in the samples even though clover is truly better: $z = \dfrac{0 - 4}{3} \approx -1.33$.



In practice we do not know σ_1 and σ_2, so we put s_1 and s_2 in their places to get the standard error $SE = \sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}$. Just as in Lesson 10.3, estimating the spread adds extra variability, so we use a t distribution instead of z. The standardized difference does not follow a t distribution exactly, but it is close to one whose df lies somewhere between the smaller of $n_1 - 1$ and $n_2 - 1$ and $n_1 + n_2 - 2$. This course uses the SMALLEST possible value, df = the smaller of $n_1 - 1$ and $n_2 - 1$. Fewer df means a

larger t^*, so our intervals are a little wider and our P-values a little larger than the ones technology gives with its more exact formula. If anything we claim less than the data show, never more; that is why it is called conservative.

The interval is $(\bar{x}_1 - \bar{x}_2) \pm t^* \cdot SE$. The test of $H_0 : \mu_1 = \mu_2$ uses $t = \dfrac{(\bar{x}_1 - \bar{x}_2) - 0}{SE}$, the number of standard errors between the observed difference and the difference of 0 claimed by H_0. Notice that we do NOT pool here, unlike the two-proportion test. For proportions, $p_1 = p_2$ forces the spreads to be equal too, because the spread of a proportion depends only on p. For means, H_0 says the centers are equal but says nothing about the spreads, so each sample keeps its own s.

Conditions. Random: two independent random samples, or two groups formed by random assignment. 10%: each sample is at most 10% of its population when sampling without replacement. Normal/Large Sample: for EACH group, either the population is normal, or $n \geq 30$, or (for a small sample) a graph of the data shows no strong skewness and no outliers. The t procedures are robust: they work well even when the populations are not exactly normal, as long as there are no outliers or strong skewness in small samples.

Always finish in context with units: "We are 95% confident that the mean yield of variety A is between 1.4 and 13.0 bushels per acre higher than variety B." If the interval contains 0, a difference of 0 (no difference) is plausible. With random assignment, a significant difference can be blamed on the treatment; with random samples from two populations, we can generalize but not claim cause.

KEY FORMULAS

$$\mu_{\bar{x}_1 - \bar{x}_2} = \mu_1 - \mu_2 \qquad \sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

$$SE = \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}} \qquad \text{df} = \text{smaller of } n_1 - 1 \text{ and } n_2 - 1$$

$$\text{Interval: } (\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}$$

$$\text{Test of } H_0 : \mu_1 = \mu_2 : \quad t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{\dfrac{s_1^2}{n_1} + \dfrac{s_2^2}{n_2}}}$$

Practice habit. Square the standard deviations before you add: $SE = \sqrt{s_1^2/n_1 + s_2^2/n_2}$, never $s_1 + s_2$. Then write "df = smaller of $n_1 - 1$ and $n_2 - 1$" next to every two-sample t calculation.

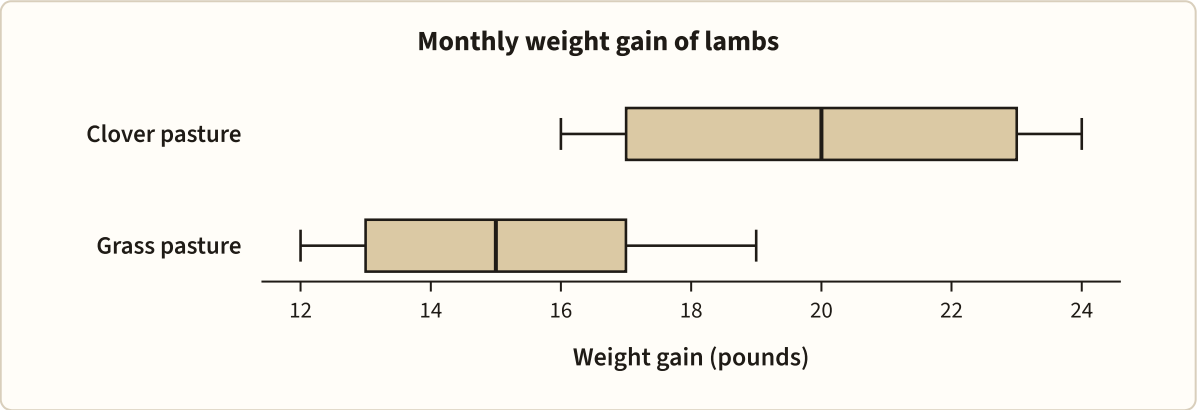
WORKED EXAMPLES

Example 1. An agricultural college randomly assigned 25 test plots to two varieties of wheat: 9 plots to variety A and 16 plots to variety B. Variety A yielded $\bar{x}_A = 63.2$ bushels per acre with $s_A = 4.5$; variety B yielded $\bar{x}_B = 56.0$ bushels per acre with $s_B = 8.0$. Dotplots of both samples show no strong skewness and no outliers. Construct and interpret a 95% confidence interval for $\mu_A - \mu_B$. Round the endpoints to 1 decimal place.

1. State: estimate $\mu_A - \mu_B$, the difference in true mean yield (bushels per acre) for plots like these planted with variety A and with variety B, with 95% confidence.
2. Plan: two-sample t interval. Random: plots were randomly assigned to varieties. Normal/Large Sample: both samples are small, but the dotplots show no strong skewness or outliers.
3. Do: difference $\bar{x}_A - \bar{x}_B = 63.2 - 56.0 = 7.2$ bushels per acre.
4. $SE = \sqrt{\dfrac{4.5^2}{9} + \dfrac{8.0^2}{16}} = \sqrt{\dfrac{20.25}{9} + \dfrac{64}{16}} = \sqrt{2.25 + 4} = \sqrt{6.25} = 2.5$.
5. df = smaller of $9 - 1 = 8$ and $16 - 1 = 15$, so df = 8 and $t^* = 2.306$.
6. Margin of error: $2.306 \times 2.5 = 5.765$.
7. Interval: $7.2 \pm 5.765 = (1.435, 12.965) \approx (1.4, 13.0)$.
8. Conclude: We are 95% confident that the true mean yield of variety A is between 1.4 and 13.0 bushels per acre higher than that of variety B on plots like these. Since 0 is not in the interval, the data give convincing evidence that variety A has a higher mean yield than variety B on plots like these. Because the plots were randomly assigned to varieties, we may conclude that the variety caused the difference.

Answer: $(1.4, 13.0)$ bushels per acre. We are 95% confident that variety A's true mean yield is between 1.4 and 13.0 bushels per acre higher than variety B's.

Example 2. Henry randomly assigned 12 lambs to two pastures: 5 to a clover pasture and 7 to a grass pasture. Their weight gains in one month, in pounds, were: clover 16, 20, 22, 24, 18; grass 12, 15, 17, 14, 13, 19, 15. Do the data give convincing evidence at the $\alpha = 0.05$ level that lambs gain more weight on average on the clover pasture? Give t to 2 decimal places and the P-value to 4 decimal places.



1. State: $H_0 : \mu_C = \mu_G$ and $H_a : \mu_C > \mu_G$, where μ_C and μ_G are the true mean monthly gains of lambs like these on clover and on grass. $\alpha = 0.05$.
2. Plan: two-sample t test. Random: lambs were randomly assigned to pastures. Normal/Large Sample: the samples are small, but the boxplots show no outliers and no strong skewness.
3. Clover: $\bar{x}_C = \dfrac{100}{5} = 20$; deviations $-4, 0, 2, 4, -2$; squares add to 40; $s_C^2 = \dfrac{40}{4} = 10$.
4. Grass: $\bar{x}_G = \dfrac{105}{7} = 15$; deviations $-3, 0, 2, -1, -2, 4, 0$; squares add to 34; $s_G^2 = \dfrac{34}{6} \approx 5.6667$.
5. $SE = \sqrt{\dfrac{10}{5} + \dfrac{5.6667}{7}} = \sqrt{2 + 0.80952} = \sqrt{2.80952} \approx 1.67616$ pounds.
6. $t = \dfrac{20 - 15}{1.67616} \approx 2.98$, with df = smaller of 4 and 6 = 4.
7. P-value $= P(t > 2.98) \approx 0.0203$ with df = 4. (Table: 2.98 lies between 2.776 and 2.999, so the P-value is between 0.02 and 0.025.)
8. Conclude: Because $0.0203 < 0.05$, we reject H_0. There is convincing evidence that lambs like these gain more weight on average on the clover pasture. Since the lambs were randomly assigned, the pasture caused the difference.

Answer: $t \approx 2.98$, df = 4, P-value ≈ 0.0203; reject H_0. There is convincing evidence that lambs gain more on average on clover.

PRACTICE

1. EASY Samuel records golf drives with two drivers. With driver 1, 25 drives have $s_1 = 15$ yards; with driver 2, 16 drives have $s_2 = 12$ yards. Find the standard error of $\bar{x}_1 - \bar{x}_2$. Round to 4 decimal places.

2. EASY An orchard keeper measures random samples of 13 trees of one apple variety and 21 trees of another and wants a 95% two-sample t interval. Find (a) the conservative degrees of freedom and (b) t^*.

3. EASY Birth weights of Angus calves on a large ranch have mean 75 pounds and standard deviation 8 pounds; Hereford calves have mean 80 pounds and standard deviation 10 pounds. A rancher takes random samples of 16 Angus and 25 Hereford calves. Find the mean and standard deviation of the sampling distribution of $\bar{x}_H - \bar{x}_A$. Round the standard deviation to 4 decimal places.

4. EASY A beekeeping club recorded the honey harvest from random samples of hives: 9 Italian hives averaged 62.0 pounds with $s = 6.0$ pounds, and 25 Carniolan hives averaged 56.0 pounds with $s = 7.5$ pounds. Find the two-sample t statistic for $H_0 : \mu_I = \mu_C$ and the conservative df.

5. MEDIUM A 95% confidence interval for $\mu_A - \mu_B$, the difference in mean daily weight gain (pounds) of calves on feed A and feed B, is $(-0.12, 0.31)$. (a) Interpret the interval in context. (b) Does the interval give convincing evidence that the mean gains differ? Answer yes or no.

6. MEDIUM Random samples of 2-by-4 studs were measured from two sawmills. Mill A: $n = 11$, $\bar{x} = 96.20$ inches, $s = 0.30$ inch. Mill B: $n = 15$, $\bar{x} = 95.90$ inches, $s = 0.40$ inch. Construct a 90% confidence interval for $\mu_A - \mu_B$. Round the endpoints to 3 decimal places.

7. MEDIUM For the honey data in Problem 4 ($t = 2.40$, df = 8), test $H_0 : \mu_I = \mu_C$ against $H_a : \mu_I \neq \mu_C$ at the $\alpha = 0.05$ level. Give the P-value to 4 decimal places and state your conclusion in context.

8. MEDIUM Explain why this course uses df = the smaller of $n_1 - 1$ and $n_2 - 1$ for two-sample t procedures, and what effect this choice has on intervals and P-values compared with technology.

9. MEDIUM Grace wants to compare the weekly practice time of choir sopranos and altos. Her random sample of 8 sopranos includes times of 30 to 60 minutes and one of 300 minutes; her random sample of 40 altos has a slightly skewed distribution. Is the Normal/Large Sample condition met for a two-sample t test? Answer yes or no and explain.

10. HARD A rancher randomly assigned 28 calves to two feeds. The summary statistics are in the table, and histograms of both groups show no strong skewness or outliers. Do the data give convincing evidence at the $\alpha = 0.05$ level that feed A produces a higher mean daily gain? Use the four-step process; give t to 2 decimal places and the P-value to 4 decimal places.

Daily weight gain of calves

Feed	Number of calves	Mean gain (lb per day)	SD (lb per day)
Feed A	13	2.45	0.30
Feed B	15	2.20	0.36

11. HARD Samuel compared two brands of golf ball by hitting 11 drives with each, in a random order, on the same morning. The summary statistics are in the table, and dotplots of both sets of distances show no strong skewness or outliers. (a) Construct a 95% confidence interval for $\mu_A - \mu_B$. Round the endpoints to 1 decimal place. (b) Does the interval give convincing evidence that the mean drive distances differ? Answer yes or no.

Golf drive distances

Ball	Number of drives	Mean (yards)	SD (yards)
Ball A	11	231.4	8.2
Ball B	11	224.6	9.5

12. HARD Ruth randomly assigned 10 bean plants to two groups of 5. One group got fertilizer and one did not. After four weeks the heights, in centimeters, were: fertilized 24, 28, 25, 30, 23; unfertilized 20, 22, 19, 23, 21. Test at the $\alpha = 0.05$ level whether fertilizer increases the mean height. Give t to 2 decimal places, the df, and the P-value to 4 decimal places.

13. CHALLENGE On a large farm, brown eggs have weights that are normally distributed with mean 62 grams and standard deviation 4 grams; white eggs are normal with mean 60 grams and standard deviation 3 grams. A worker picks a random sample of 16 eggs of each color. Find the probability that the mean weight of the brown eggs is more than 4 grams greater than the mean weight of the white eggs. Round to 4 decimal places.

14. CHALLENGE For the calf data in Problem 10, a computer program reports df of about 26 and a one-sided P-value of about 0.028, smaller than the 0.0341 found with df = 12. Explain why the computer's P-value is smaller, and why reporting 0.0341 is still honest.

COMMON MISTAKES

- Adding the standard deviations, $s_1 + s_2$, or forgetting to divide each variance by its own sample size. The standard error is $\sqrt{s_1^2/n_1 + s_2^2/n_2}$.
- Using df = $n_1 + n_2 - 2$ or $n - 1$ for the larger sample. In this course df is the SMALLER of $n_1 - 1$ and $n_2 - 1$.
- Checking the Normal/Large Sample condition for only one group. Check each sample separately: a large sample in one group does not rescue a small, skewed sample in the other.
- Pooling the standard deviations as in the two-proportion test. $H_0 : \mu_1 = \mu_2$ says nothing about the spreads, so each group keeps its own s.

ANSWERS TO PRACTICE

1. $SE = \sqrt{\dfrac{225}{25} + \dfrac{144}{16}} = \sqrt{9 + 9} = \sqrt{18} \approx 4.2426$ yards

2. (a) df = smaller of 12 and 20 = 12. (b) $t^* = 2.179$.

3. Mean $= 80 - 75 = 5$ pounds; SD $= \sqrt{\dfrac{100}{25} + \dfrac{64}{16}} = \sqrt{4 + 4} = \sqrt{8} \approx 2.8284$ pounds.

4. $SE = \sqrt{\dfrac{36}{9} + \dfrac{56.25}{25}} = \sqrt{4 + 2.25} = 2.5$; $t = \dfrac{62.0 - 56.0}{2.5} = 2.40$; df = smaller of 8 and 24 = 8.

5. (a) We are 95% confident that the true mean daily gain on feed A is somewhere between 0.12 pound less and 0.31 pound more than on feed B. (b) No: 0 is in the interval, so no difference is a plausible value.

6. $SE = \sqrt{\dfrac{0.09}{11} + \dfrac{0.16}{15}} = \sqrt{0.0188485} \approx 0.137290$; df = 10, $t^* = 1.812$; $0.30 \pm 1.812(0.137290) = 0.30 \pm 0.248769$, so $(0.051, 0.549)$ inch.

7. Two-sided P-value $= 2P(t > 2.40) \approx 0.0432$ with df = 8 (table: between 0.04 and 0.05). Since $0.0432 < 0.05$, reject H_0: there is convincing evidence that the mean honey harvest differs for Italian and Carniolan hives in this club.

8. The true df of the two-sample t statistic lies somewhere between the smaller $n - 1$ and $n_1 + n_2 - 2$. Using the smallest value gives the largest t^* and the fattest tails, so our intervals are a little wider and our P-values a little larger than technology's. The method errs on the safe side: we never claim more confidence or stronger evidence than the data support.

9. No. The alto sample is large ($n = 40 \geq 30$), but the soprano sample is small ($n = 8$) and has a clear outlier at 300 minutes, so t procedures are not safe for it.

10. State: $H_0 : \mu_A = \mu_B$, $H_a : \mu_A > \mu_B$, where μ is the true mean daily gain of calves like these on each feed; $\alpha = 0.05$. Plan: two-sample t test; random assignment; histograms show no strong skewness or outliers. Do: $SE = \sqrt{\dfrac{0.30^2}{13} + \dfrac{0.36^2}{15}} = \sqrt{0.0155631} \approx 0.124752$; $t = \dfrac{2.45 - 2.20}{0.124752} \approx 2.00$; df = smaller of 12 and 14 = 12; P-value ≈ 0.0341 (table: between 0.025 and 0.05). Conclude: $0.0341 < 0.05$, so reject H_0. There is convincing evidence that feed A causes a higher mean daily gain for calves like these.

11. (a) $SE = \sqrt{\dfrac{8.2^2}{11} + \dfrac{9.5^2}{11}} = \sqrt{\dfrac{67.24 + 90.25}{11}} = \sqrt{14.31727} \approx 3.78382$; df = 10, $t^* = 2.228$; $6.8 \pm 2.228(3.78382) = 6.8 \pm 8.4303$, so $(-1.6, 15.2)$ yards. (b) No: 0 is in the interval, so the data do not show a real difference in mean distance.

12. $\bar{x}_F = 26$, $s_F^2 = \dfrac{34}{4} = 8.5$; $\bar{x}_U = 21$, $s_U^2 = \dfrac{10}{4} = 2.5$; $SE = \sqrt{\dfrac{8.5}{5} + \dfrac{2.5}{5}} = \sqrt{2.2} \approx 1.48324$; $t = \dfrac{5}{1.48324} \approx 3.37$, df = 4; P-value ≈ 0.0140 (table: between 0.01 and 0.02). Since $0.0140 < 0.05$, reject H_0: there is convincing evidence that fertilizer increases the mean height of bean plants like these.

13. $\bar{x}_B - \bar{x}_W$ is normal with mean $62 - 60 = 2$ and SD $\sqrt{\dfrac{16}{16} + \dfrac{9}{16}} = \sqrt{\dfrac{25}{16}} = 1.25$ grams. $z = \dfrac{4 - 2}{1.25} = 1.60$; $P(Z > 1.60) = 1 - 0.9452 = 0.0548$.

14. The computer uses a more exact formula for df, which here gives about 26 instead of the smallest possible value, 12. A t distribution with more df has thinner tails, so the same t = 2.00 cuts off a smaller tail area. Our conservative df makes the P-value a little too large, never too small, so we never overstate the evidence; here both methods lead to the same decision.

LESSON 12.3 · 2 SESSIONS OF 45 MINUTES

Paired Data

OBJECTIVES

- Recognize paired data (before and after on the same subject, matched pairs, a field split in half) and tell it apart from two independent samples.
- Construct and interpret a paired t interval for the mean difference μ_d.
- Carry out a paired t test on the differences with the four-step process.
- Explain why pairing removes variation between subjects and gives a more powerful test, and why treating paired data as two independent samples is a mistake.

VIDEO

WATCH

The Paired-Difference t Procedure

jbstatistics · 8:20

When to watch: Before Example 1: paired data analyzed as one sample of differences, with an interval and a test.

https://www.youtube.com/watch?v=TMRBGFIj0RM

BACKUP

Matched Pairs t Confidence Interval

ProfRobBob · 12:18

When to watch: After Example 2: a matched-pairs t interval for pretest and posttest scores.

https://www.youtube.com/watch?v=l-L9nSjfdQo

KEY TERMS

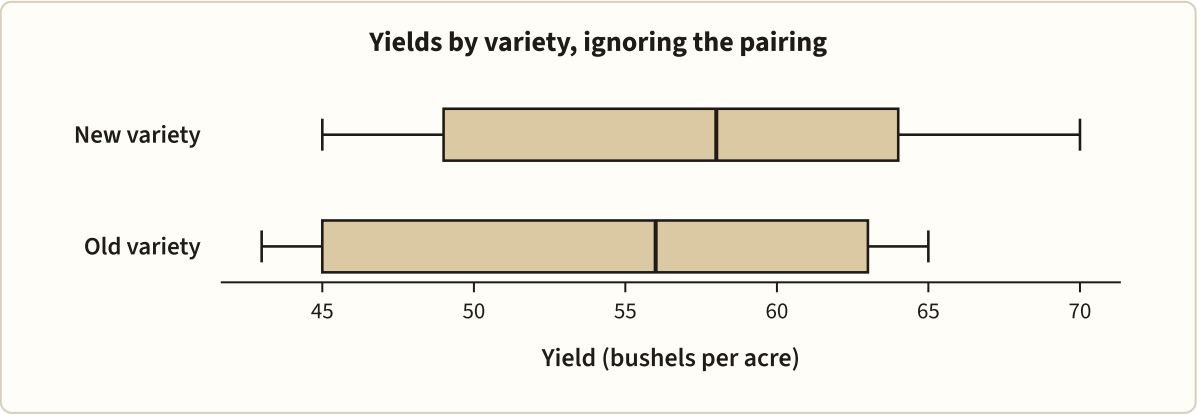
Paired data	Two measurements that belong together because they come from the same individual (before and after), from two halves of the same unit (a field split in two), or from two closely matched individuals (twin lambs).
Difference (d)	For each pair, the first measurement minus the second, $d = x_1 - x_2$, always subtracted in the same order.
Mean difference μ_d	The true mean of the differences for all pairs in the population (or for all pairs that could receive the treatments). It is the parameter of a paired analysis.
Paired t procedures	One-sample t procedures (interval and test) applied to the list of differences, with df = n − 1 where n is the number of pairs.

CORE IDEAS

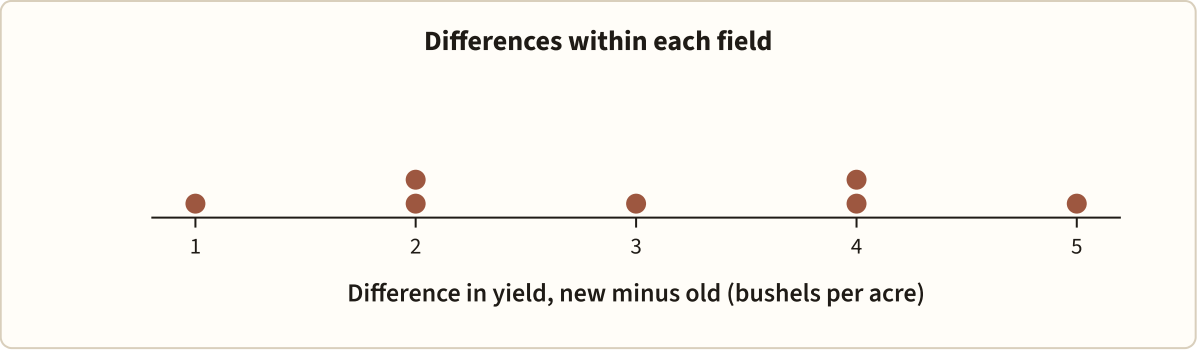
Sometimes the two sets of measurements are not separate groups at all. A farmer splits each of 7 fields in half and plants a new wheat variety on one half and the old variety on the other. A woodworker measures the same boards before and after drying. A shepherd gives one lamb of each pair of twins a supplement. In each case every value in one list has a PARTNER in the other list. Such data are paired, and the question to ask is simple: "Is each value in one group linked to one particular value in the other?"

The right way to handle paired data is to subtract within each pair, always in the same order, and work with the ONE list of differences d. The parameter is μ_d, the true mean difference, and $H_0 : \mu_d = 0$ says there is no difference on average. From here we use the one-sample t procedures of Lessons 10.4 and 11.3, applied to the differences: the interval $\bar{x}_d \pm t^* \dfrac{s_d}{\sqrt{n}}$ and the test statistic $t = \dfrac{\bar{x}_d - 0}{s_d/\sqrt{n}}$, with df = n − 1, where n is the number of PAIRS.

Why pair at all? Look at the wheat yields from the 7 split fields below. Fields differ a great deal: rich bottomland yields about 70 bushels per acre and a stony hillside about 45. The boxplots for the two varieties overlap almost completely, because the field-to-field variation is so large.



Now look at the differences, new minus old, field by field: 3, 4, 2, 5, 2, 4, 1. Every one is positive! A rich field is rich for both halves, so subtracting within a field cancels the field's richness and leaves only the effect of the variety plus a little leftover variation. The standard deviation of the yields is about 9 bushels, but the standard deviation of the differences is only about 1.4 bushels. A smaller standard deviation means a smaller standard error, a larger t, and a better chance of detecting a real difference: pairing gives a more POWERFUL test.



Treating paired data as two independent samples is a serious mistake for two reasons. First, the two samples are not independent (the halves of a rich field are both high), so the formula for the two-sample standard error does not apply. Second, that standard error includes all the variation between fields, which the pairing was designed to remove, so a real effect gets buried. For these fields the wrong two-sample analysis gives $t \approx 0.64$, while the correct paired analysis gives $t \approx 5.61$.

Conditions for paired t procedures: Random: the pairs are a random sample, or the two treatments are randomly assigned within each pair (for example, a coin flip decides which half of a field gets the new variety). 10%: when sampling pairs without replacement, n is at most 10% of the population of pairs. Normal/Large Sample: check the DIFFERENCES: the population of differences is normal, or $n \geq 30$, or a graph of the differences shows no strong skewness and no outliers.

KEY FORMULAS

$$d = x_1 - x_2 \text{ for each pair (same order every time)}$$

$$\text{Paired t interval: } \bar{x}_d \pm t^* \frac{s_d}{\sqrt{n}}, \quad \text{df} = n - 1$$

$$\text{Paired t test of } H_0 : \mu_d = 0 : \quad t = \frac{\bar{x}_d - 0}{s_d / \sqrt{n}}, \quad \text{df} = n - 1$$

Practice habit. Before choosing a method, ask: "Is each value in one group linked to one particular value in the other?" If yes, subtract within pairs first, write down the order of subtraction, and do a one-sample analysis of the differences.

WORKED EXAMPLES

Example 1. Henry chose 6 oak boards at random from a large stack at the sawmill and measured the width of each, in millimeters, before and after kiln drying. The data are in the table. Construct and interpret a 95% confidence interval for the true mean shrinkage. Round the endpoints to 1 decimal place.

Oak board widths before and after kiln drying (mm)

Board	Before	After	Shrinkage (before minus after)
1	140	136	4
2	146	140	6
3	138	135	3
4	151	146	5
5	143	136	7
6	135	130	5

1. These are paired data: each board is measured twice. Work with the shrinkage d = before minus after: 4, 6, 3, 5, 7, 5.
2. State: estimate μ_d, the true mean shrinkage in width (mm) for all boards in the stack, with 95% confidence.
3. Plan: paired t interval. Random: a random sample of boards. 10%: 6 boards are far fewer than 10% of a large stack. Normal/Large Sample: a dotplot of the 6 differences (3 to 7) shows no strong skewness or outliers.
4. Do: $\bar{x}_d = \dfrac{30}{6} = 5$ mm. Deviations $-1, 1, -2, 0, 2, 0$; squares add to 10; $s_d^2 = \dfrac{10}{5} = 2$, $s_d = \sqrt{2} \approx 1.41421$ mm.
5. $SE = \dfrac{1.41421}{\sqrt{6}} \approx 0.57735$ mm; df = 5, $t^* = 2.571$.
6. Margin of error: $2.571 \times 0.57735 \approx 1.48437$; interval $5 \pm 1.48437 = (3.51563, 6.48437) \approx (3.5, 6.5)$ mm.
7. Conclude: We are 95% confident that kiln drying shrinks the width of boards from this stack by between 3.5 and 6.5 millimeters on average.

Answer: $(3.5, 6.5)$ mm. We are 95% confident that the true mean shrinkage of boards from this stack is between 3.5 and 6.5 millimeters.

Example 2. A farmer split each of 7 fields in half and flipped a coin to decide which half got a new wheat variety; the other half got the old variety. The yields are in the table. Do the data give convincing evidence at the $\alpha = 0.05$ level that the new variety has a higher mean yield? Then compare with the (wrong) two-sample analysis, which uses $s_{new} \approx 8.832$ and $s_{old} \approx 8.583$.

Wheat yields on split fields (bushels per acre)

Field	New variety	Old variety	Difference (new minus old)
1	52	49	3
2	61	57	4
3	45	43	2
4	70	65	5
5	58	56	2
6	49	45	4
7	64	63	1

1. Paired data: both varieties grow in each field. Let d = new minus old: 3, 4, 2, 5, 2, 4, 1.
2. State: $H_0 : \mu_d = 0$ and $H_a : \mu_d > 0$, where μ_d is the true mean difference in yield (new minus old, bushels per acre) for fields like these. $\alpha = 0.05$.
3. Plan: paired t test. Random: a coin flip assigned the varieties within each field. Normal/Large Sample: the dotplot of the 7 differences shows no strong skewness or outliers.
4. Do: $\bar{x}_d = \dfrac{21}{7} = 3$. Deviations $0, 1, -1, 2, -1, 1, -2$; squares add to 12; $s_d^2 = \dfrac{12}{6} = 2$, $s_d \approx 1.41421$.
5. $SE = \dfrac{1.41421}{\sqrt{7}} \approx 0.53452$; $t = \dfrac{3 - 0}{0.53452} \approx 5.61$ with df = 6.
6. P-value $= P(t > 5.61) \approx 0.0007$ (table: between 0.0005 and 0.001).
7. Conclude: $0.0007 < 0.05$, so reject H_0. There is convincing evidence that the new variety produces a higher mean yield than the old one on fields like these, about 3 bushels per acre more.
8. Wrong analysis for comparison: $SE = \sqrt{\dfrac{8.832^2}{7} + \dfrac{8.583^2}{7}} \approx 4.655$ and $t = \dfrac{57 - 54}{4.655} \approx 0.64$, with a P-value of about 0.27. Ignoring the pairing lets the large field-to-field variation hide a real and consistent difference.

Answer: $t \approx 5.61$, df = 6, P-value ≈ 0.0007; reject H_0. There is convincing evidence that the new variety has a higher mean yield. The two-sample analysis would give only $t \approx 0.64$.

PRACTICE

1. **EASY** Are the data paired? Answer yes or no for each. (a) The weights of 10 calves at birth and again at one month old. (b) The yields of 8 plots planted with fertilizer A and 8 different plots planted with fertilizer B.

2. **EASY** Five runners ran a mile before and after a summer of training. Their times in minutes were: before 6.2, 5.8, 7.0, 6.5, 6.0; after 5.9, 5.7, 6.6, 6.1, 6.0. Using d = before minus after, find $\bar{x}_d$ and s_d. Round s_d to 3 decimal places.

3. **EASY** Sixteen students counted their push-ups before and after a four-week program. The gains (after minus before) had $\bar{x}_d = 2.4$ and $s_d = 3.0$. Find (a) the standard error, (b) the t statistic for $H_0 : \mu_d = 0$, and (c) the df.

4. **EASY** A paired t interval at 90% confidence is based on 12 pairs. Find (a) the df and (b) t^*.

5. **MEDIUM** A company wants to compare two brands of work-boot soles. Plan 1: give 10 farmhands brand A and 10 other farmhands brand B. Plan 2: give each of 10 farmhands one sole of each brand, with a coin flip deciding which foot gets brand A. Explain why Plan 2 is the better design.

6. **MEDIUM** Nine dairy cows were each fed an old ration for two weeks and a new ration for two weeks, in a random order. The differences in daily milk yield (new minus old) had $\bar{x}_d = 4.2$ pounds and $s_d = 3.6$ pounds, and their dotplot showed no outliers. Construct and interpret a 95% confidence interval for μ_d. Round the endpoints to 2 decimal places.

7. **MEDIUM** On each of 11 apple trees, one branch (chosen by a coin flip) was sprayed with a thinning spray and a similar branch was not. The differences in mean apple diameter (sprayed minus unsprayed) had $\bar{x}_d = 1.5$ mm and $s_d = 2.0$ mm, with no outliers. Test $H_0 : \mu_d = 0$ against $H_a : \mu_d > 0$ at the $\alpha = 0.05$ level. Give t (2 decimal places), the P-value (4 decimal places) and your decision.

8. **MEDIUM** Samuel measured the time to bake 8 batches of bread in an old oven and the same 8 recipes in a new oven. He ran a two-sample t test and found a P-value of 0.41. Explain what is wrong with his analysis and what he should do instead.

9. **MEDIUM** Forty members of a church walking club, chosen at random from the club's roll, recorded their weekly walking minutes before and after a new Saturday walk was added. The 40 differences are moderately skewed to the right with no extreme outliers. Is the Normal/Large Sample condition met for a paired t test? Answer yes or no and explain.

10. HARD Henry and Ruth each estimated the brightness (magnitude) of the same 5 variable stars, chosen at random from their astronomy club's observing list, on one clear night. Their estimates are in the table. Test at the $\alpha = 0.05$ level whether their estimates differ on average. Give t (2 decimal places), df, the P-value (4 decimal places) and a conclusion in context.

Estimated magnitudes of five variable stars

Star	Henry	Ruth	Difference (Henry minus Ruth)
A	4.6	4.4	0.2
B	5.1	5	0.1
C	3.9	3.6	0.3
D	4.4	4.5	-0.1
E	5.5	5.5	0

11. HARD A shepherd took 8 pairs of twin lambs and, by a coin flip, gave one lamb in each pair a mineral supplement. The differences in monthly weight gain, in pounds (supplement minus none), were 3, 1, 4, 2, 5, 0, 3, 2. Construct and interpret a 95% confidence interval for μ_d. Round the endpoints to 1 decimal place.

12. HARD Thirteen cross-country runners each ran the same course once in old shoes and once in new shoes, in a random order. The differences in time (new minus old) had $\bar{x}_d = -8$ seconds and $s_d = 12$ seconds, with no outliers. Test $H_0 : \mu_d = 0$ against $H_a : \mu_d \neq 0$ at the $\alpha = 0.01$ level. Give t (2 decimal places), the P-value (4 decimal places) and your decision.

13. CHALLENGE Two thermometers, an old one and a new one, were read at the same moment at 5 weather stations. Old: 61, 72, 55, 80, 67; new: 59, 71, 52, 78, 65 (degrees Fahrenheit). (a) Find the paired t statistic for d = old minus new. (b) Find the two-sample t statistic you would get by wrongly treating the readings as independent samples. Round both to 2 decimal places. (c) Explain why the two answers are so different.

COMMON MISTAKES

- Running a two-sample t test on paired data. If each value has a partner, subtract within pairs and do a one-sample analysis of the differences.
- Using the number of measurements (2n) instead of the number of pairs (n) for the sample size and df. In a paired analysis, n is the number of differences.
- Mixing the order of subtraction (some pairs new minus old, others old minus new), or stating a conclusion with the sign reversed. Write the order in the definition of d and use it every time.
- Checking the Normal condition on each list separately instead of on the differences. The condition is about the distribution of the differences.

ANSWERS TO PRACTICE

1. (a) Yes: each calf is weighed twice. (b) No: the plots are separate, unlinked groups.
2. d: 0.3, 0.1, 0.4, 0.4, 0.0; $\bar{x}_d = \dfrac{1.2}{5} = 0.24$ minute; squared deviations add to 0.132, so $s_d = \sqrt{\dfrac{0.132}{4}} = \sqrt{0.033} \approx 0.182$ minute.
3. (a) $\dfrac{3.0}{\sqrt{16}} = 0.75$ (b) $t = \dfrac{2.4}{0.75} = 3.20$ (c) df = 15
4. (a) df = 12 − 1 = 11 (b) $t^* = 1.796$
5. Farmhands differ a lot in how hard they are on boots (weight, walking, type of work). In Plan 2 both brands get exactly the same wear from the same person, so subtracting within each farmhand removes that person-to-person variation. The differences vary much less, the standard error is smaller, and a real difference between the brands is easier to detect.
6. $SE = \dfrac{3.6}{\sqrt{9}} = 1.2$; df = 8, $t^* = 2.306$; $4.2 \pm 2.306(1.2) = 4.2 \pm 2.7672$, so $(1.43, 6.97)$ pounds. We are 95% confident that the new ration increases mean daily milk yield for cows like these by between 1.43 and 6.97 pounds.
7. $SE = \dfrac{2.0}{\sqrt{11}} \approx 0.603023$; $t = \dfrac{1.5}{0.603023} \approx 2.49$, df = 10; P-value ≈ 0.0161 (table: between 0.01 and 0.02). Since $0.0161 < 0.05$, reject H_0: there is convincing evidence that the spray increases mean apple diameter on trees like these.
8. The data are paired: each recipe was baked in both ovens. A two-sample test treats the ovens' lists as independent and lets the large differences between recipes (a dense rye takes far longer than rolls) inflate the standard error. He should find the difference (old minus new) for each recipe and do a paired t test on those 8 differences.
9. Yes. There are $n = 40$ differences, and $40 \geq 30$, so the sampling distribution of $\bar{x}_d$ is approximately normal by the central limit theorem even though the differences are somewhat skewed.
10. d = Henry minus Ruth: 0.2, 0.1, 0.3, −0.1, 0.0. $H_0 : \mu_d = 0$, $H_a : \mu_d \neq 0$. $\bar{x}_d = \dfrac{0.5}{5} = 0.1$; squared deviations add to 0.10, $s_d = \sqrt{0.025} \approx 0.158114$; $SE = \dfrac{0.158114}{\sqrt{5}} \approx 0.070711$; $t = \dfrac{0.1}{0.070711} \approx 1.41$, df = 4; two-sided P-value ≈ 0.2302. Since $0.2302 > 0.05$, fail to reject H_0: there is not convincing evidence that Henry's and Ruth's estimates differ on average.
11. $\bar{x}_d = \dfrac{20}{8} = 2.5$; squared deviations add to 18; $s_d = \sqrt{\dfrac{18}{7}} \approx 1.603567$; $SE = \dfrac{1.603567}{\sqrt{8}} \approx 0.566947$; df = 7, $t^* = 2.365$; 2.5 ± 1.34083, so $(1.2, 3.8)$ pounds. We are 95% confident that the supplement increases mean monthly gain for twin lambs like these by between 1.2 and 3.8 pounds.

12. $SE = \dfrac{12}{\sqrt{13}} \approx 3.32820$; $t = \dfrac{-8}{3.32820} \approx -2.40$, df = 12; two-sided P-value ≈ 0.0333 (table: between 0.02 and 0.04). Since $0.0333 > 0.01$, fail to reject H_0: at the 0.01 level there is not convincing evidence that the new shoes change mean race time. (At $\alpha = 0.05$ the result would have been significant.)

13. (a) d: 2, 1, 3, 2, 2; $\bar{x}_d = 2$, $s_d = \sqrt{\dfrac{2}{4}} \approx 0.70711$, $SE \approx 0.31623$, $t \approx 6.32$. (b) Means 67 and 65; $s^2_{old} = \dfrac{374}{4} = 93.5$, $s^2_{new} = \dfrac{410}{4} = 102.5$; $SE = \sqrt{\dfrac{93.5}{5} + \dfrac{102.5}{5}} = \sqrt{39.2} \approx 6.26099$; $t = \dfrac{2}{6.26099} \approx 0.32$. (c) The temperatures differ greatly from station to station (55 to 80 degrees), but at each station the two thermometers differ by only 1 to 3 degrees. Pairing subtracts out the station-to-station variation; the two-sample analysis leaves it in the standard error, which hides the consistent 2-degree difference.

LESSON 12.4 · 2 SESSIONS OF 45 MINUTES

Choosing the Right Procedure

OBJECTIVES

- Identify the parameter in a problem by deciding whether the variable is categorical (a proportion) or quantitative (a mean).
- Choose the correct inference procedure (one or two samples, paired or independent, interval or test) and justify the choice in a sentence.
- Check the conditions for the chosen procedure and carry out a complete four-step solution.
- Explain why every procedure in the course has the same structure: an estimate, a standard error and a critical value or test statistic.

KEY TERMS

Inference procedure	A method for drawing a conclusion about a parameter from sample data: a confidence interval (to estimate) or a significance test (to weigh evidence for a claim).
Parameter of interest	The single number about the population or treatments that the question is really asking about, such as p, $p_1 - p_2$, μ, μ_d or $\mu_1 - \mu_2$. Always define it in words.
Independent samples	Two groups in which the individuals in one group have no link to particular individuals in the other group.
Standard error	An estimate, computed from the data, of the standard deviation of a statistic's sampling distribution: how far the statistic typically falls from the parameter.

CORE IDEAS

You now know five families of inference procedures, each with an interval and a test. In a textbook chapter it is obvious which one to use; in real life (and on the unit test) the hardest step is choosing. Three questions, asked in order, lead you to the right procedure almost every time.

Question 1: Is the variable categorical or quantitative? If each individual gives a category (germinated or not, made the free throw or not), the parameter is a PROPORTION. If each individual gives a measured number (weight, yield, time), the parameter is a MEAN. Watch for the word "percent": the percent of eggs that are Grade AA is a proportion, but the moisture percent of each hay bale is a measurement, so the average moisture is a mean.

Question 2: How many groups, and are they paired? One group compared with a claimed value calls for a one-sample procedure. Two groups call for a two-sample procedure, unless each value in one group is linked to a particular value in the other (the same subject twice, two halves of one field,

twins). Linked values are paired: subtract within pairs and use a one-sample t procedure on the differences.

Question 3: Estimate or test? If the question asks "how much" or "how large is the difference", build a confidence interval. If it asks "is there evidence that", carry out a significance test. Often the best report gives both: the test says whether there is an effect, and the interval says how big it is.

Decision guide for inference

Variable	Design	Parameter	Procedure (interval and test)	Lesson
Categorical	One sample	p	One-proportion z	10.2, 11.2
Categorical	Two independent groups	p1 − p2	Two-proportion z (pool only for the test)	12.1
Quantitative	One sample	μ	One-sample t, df = n − 1	10.4, 11.3
Quantitative	Paired data	μd	Paired t on the differences, df = number of pairs − 1	12.3
Quantitative	Two independent groups	μ1 − μ2	Two-sample t, df = smaller of n1 − 1 and n2 − 1	12.2

Why do all these procedures look so much alike? Every one rests on the same idea: a statistic varies from sample to sample, and its standard error tells us by how much. An interval is always statistic $\pm$ (critical value)(standard error), because about 95% (say) of samples produce a statistic within that distance of the parameter. A test statistic is always $\dfrac{\text{statistic} - \text{null value}}{\text{standard error}}$, the number of standard errors between what we saw and what H_0 claims. Only two things change: the formula for the standard error, and whether we use z (proportions, where H_0 or $\hat{p}$ fixes the spread) or t (means, where s estimates σ and adds extra uncertainty).

Once you have chosen, check the conditions that go with the procedure, then follow the four steps: State the parameter, hypotheses or confidence level; Plan by naming the procedure and checking conditions; Do the calculations; Conclude in context. Two cautions apply to every procedure: inference only makes sense when the data come from random sampling or random assignment, and if you already have data on the WHOLE population (a census), you know the parameter exactly and no inference is needed.

Conditions to check

Procedure	Random	10% (sampling without replacement)	Shape
One proportion	Random sample or random assignment	n at most 10% of population	Interval: successes and failures at least 10; test: np0 and n(1 − p0) at least 10
Two proportions	Two independent random samples or random assignment	Each n at most 10% of its population	Interval: all four observed counts at least 10; test: counts with pooled p-hat at least 10
One mean or paired	Random sample or random assignment	n at most 10% of population	Population normal, or n at least 30, or no strong skew or outliers (in the differences, if paired)
Two means	Two independent random samples or random assignment	Each n at most 10% of its population	For each group: normal, or n at least 30, or no strong skew or outliers

KEY FORMULAS

$$\text{Confidence interval: statistic} \pm (\text{critical value})(\text{standard error})$$

$$\text{Test statistic} = \frac{\text{statistic} - \text{null value}}{\text{standard error}}$$

Practice habit. Before any calculation, answer three questions in writing: Proportion or mean? One sample, two independent samples, or paired? Interval or test? Then name the procedure in the Plan step.

WORKED EXAMPLES

Example 1. An orchard manager chooses 10 trees at random from the more than 200 trees in the east block. From each tree she picks one apple on the sunny south side and one on the shaded north side and weighs both. The differences (south minus north) have mean 12 grams and standard deviation 8 grams, and a dotplot of the differences shows no strong skewness or outliers. She wants to know how much heavier south-side apples are on average. Choose the procedure, then construct and interpret a 95% interval. Round the endpoints to 1 decimal place.

1. Question 1: weight in grams is quantitative, so the parameter is a mean.
2. Question 2: the two apples come from the SAME tree, so the data are paired; we analyze the 10 differences.
3. Question 3: "how much heavier" asks for an estimate, so we build an interval. Procedure: paired t interval for μ_d.
4. State: estimate μ_d, the true mean difference in apple weight (south minus north, grams) for all trees in the east block, with 95% confidence.
5. Plan: Random: 10 randomly chosen trees. 10%: 10 is less than 10% of more than 200 trees. Normal/Large Sample: the dotplot of differences shows no strong skewness or outliers.
6. Do: $SE = \frac{8}{\sqrt{10}} \approx 2.52982$; df = 9, $t^* = 2.262$; margin $2.262 \times 2.52982 \approx 5.72246$; interval $12 \pm 5.72246 = (6.27754, 17.72246) \approx (6.3, 17.7)$ grams.
7. Conclude: We are 95% confident that apples on the sunny south side of trees in the east block weigh between 6.3 and 17.7 grams more, on average, than apples on the shaded north side.

Answer: Paired t interval: $(6.3, 17.7)$ grams. We are 95% confident that south-side apples in the east block weigh between 6.3 and 17.7 grams more on average than north-side apples.

Example 2. A county extension agent takes a random sample of 250 eggs from Farm A and a separate random sample of 200 eggs from Farm B and grades each egg. The results are in the table. Is there convincing evidence at the $\alpha = 0.05$ level that the proportions of Grade AA eggs differ at the two farms? Choose the procedure and carry it out.

Egg grades at two farms

Farm	Grade AA	Not Grade AA	Total
Farm A	205	45	250
Farm B	150	50	200
Total	355	95	450

1. Question 1: each egg is Grade AA or not, a categorical variable, so the parameter is a proportion.
2. Question 2: two separate random samples from two farms, with no link between particular eggs: two independent samples.
3. Question 3: "is there convincing evidence" calls for a test. Procedure: two-proportion z test.
4. State: $H_0 : p_A = p_B$ and $H_a : p_A \neq p_B$, where p_A and p_B are the true proportions of Grade AA eggs at Farm A and Farm B. $\alpha = 0.05$.
5. Plan: Random: independent random samples. 10%: each farm produces many thousands of eggs. Large Counts: $\hat{p}_c = \dfrac{355}{450} \approx 0.78889$; $250(0.78889) \approx 197$, $250(0.21111) \approx 53$, $200(0.78889) \approx 158$, $200(0.21111) \approx 42$, all at least 10.
6. Do: $\hat{p}_A = \dfrac{205}{250} = 0.82$, $\hat{p}_B = \dfrac{150}{200} = 0.75$; $SE = \sqrt{0.78889(0.21111)\left(\dfrac{1}{250} + \dfrac{1}{200}\right)} \approx \sqrt{0.0014989} \approx 0.038716$.
7. $z = \dfrac{0.82 - 0.75}{0.038716} \approx 1.81$; two-sided P-value ≈ 0.0706 (table with $z = 1.81$: 0.0702).
8. Conclude: Because $0.0706 > 0.05$, we fail to reject H_0. There is not convincing evidence that the proportions of Grade AA eggs differ at the two farms. (This does not prove they are equal; a 7-point difference could be real but too small to detect with these samples.)

Answer: Two-proportion z test: $z \approx 1.81$, P-value ≈ 0.0706; fail to reject H_0. There is not convincing evidence that the farms' proportions of Grade AA eggs differ.

PRACTICE

1. EASY Is the variable categorical or quantitative? (a) Whether each seed in a sample germinates. (b) The weight of each jar of honey in a sample.

2. EASY Are the data paired? Answer yes or no for each. (a) The scores of the same 15 students on a quiz before and after a review session. (b) Temperatures on 12 randomly chosen July days from Town A's records and on a separate random sample of 12 July days from Town B's records.

3. EASY Does each question call for a significance test? Answer yes or no. (a) "How many more bushels per acre does the new corn hybrid yield, on average?" (b) "Is there evidence that the new seed germinates at a higher rate than the old seed?"

4. EASY A sawmill foreman measures a random sample of 20 boards cut today to estimate the mean length of all boards cut today. Name the procedure.

5. MEDIUM A dairy researcher takes a random sample of 60 Holstein cows and a separate random sample of 50 Jersey cows and records each cow's daily milk yield. She wants to estimate how much more milk Holsteins give on average. Name the procedure and the parameter, and explain your choice.

6. MEDIUM Each of 20 choir members sang a long note twice, once after a breathing exercise and once without it, in a random order, and the director timed each note. The director wants to know whether the exercise lengthens the mean time. Name the procedure and explain why it is not a two-sample t test.

7. MEDIUM A random sample of 16 jars of honey from a large batch has mean weight $\bar{x} = 15.92$ ounces and $s = 0.12$ ounce, with no outliers or strong skewness. Construct a 95% confidence interval for the mean weight of all jars in the batch. Round the endpoints to 3 decimal places.

8. MEDIUM A seed company claims that 90% of its pepper seeds germinate. Grace plants a random sample of 200 seeds and 170 germinate. Test $H_0 : p = 0.90$ against $H_a : p < 0.90$ at the $\alpha = 0.05$ level. Give z (2 decimal places), the P-value (4 decimal places) and your decision.

9. MEDIUM Henry measures the moisture content, as a percent, of 12 randomly chosen hay bales. He says, "The answer is a percent, so I should use a procedure for proportions." Explain his mistake and name the correct procedure for estimating the average moisture.

10. HARD A bakery labels its loaves as 16 ounces. An inspector weighs 7 randomly chosen loaves: 15.8, 16.1, 15.7, 15.9, 15.6, 15.7, 15.8 ounces. Choose the procedure and test at the $\alpha = 0.05$ level whether the mean weight of the bakery's loaves is less than 16 ounces. Give t (2 decimal places), df and the P-value (4 decimal places).

11. HARD In a random sample of 300 households in one rural county, 96 heat mainly with wood; in a random sample of 250 households in a neighboring county, 60 do. Choose the procedure and construct a 95% confidence interval for the difference in the two counties' proportions (first county minus second). Round the endpoints to 3 decimal places.

12. HARD A 4-H club measured random samples of sunflowers from two large fields: Field A, $n = 31$, $\bar{x} = 168$ cm, $s = 15$ cm; Field B, $n = 41$, $\bar{x} = 159$ cm, $s = 18$ cm. (a) Name the procedure for estimating $\mu_A - \mu_B$ and explain why the Normal/Large Sample condition is met without seeing the data. (b) Construct a 95% confidence interval. Round the endpoints to 1 decimal place.

13. **CHALLENGE** Ruth chose 7 hens at random from her flock of 80 and counted the eggs each laid in March and in April (table). Days grow longer in April, and she wants to know whether hens like hers lay more eggs on average in April. Choose the procedure and carry out a full four-step test at the $\alpha = 0.05$ level, giving t (2 decimal places) and the P-value (4 decimal places).

Eggs laid by 7 hens

Hen	March	April
Hen 1	24	27
Hen 2	22	23
Hen 3	26	29
Hen 4	20	24
Hen 5	25	26
Hen 6	23	23
Hen 7	21	23

14. **CHALLENGE** A parish secretary has the attendance records of ALL 240 families in the parish for last year. She finds that 62% of families with children attended the summer picnic and 48% of families without children did. She plans to run a two-proportion z test to see whether the difference is significant. Is a significance test needed here? Answer yes or no and explain.

COMMON MISTAKES

- Choosing a procedure from a key word alone ("percent" means proportion). Ask what each individual gives: a category means a proportion; a measured number means a mean.
- Using a two-sample procedure when the data are paired, or a paired procedure for two unrelated groups. Check whether each value has a particular partner in the other group.
- Calculating before stating the parameter in words. Define p, μ, μ_d, $p_1 - p_2$ or $\mu_1 - \mu_2$ first; it makes the hypotheses and conclusion clear.
- Running inference on data that came from neither random sampling nor random assignment, or on a complete census. Inference measures chance variation; without a chance process (or with the whole population) it has nothing to measure.

ANSWERS TO PRACTICE

1. (a) Categorical (germinated or not), so the parameter is a proportion. (b) Quantitative (ounces), so the parameter is a mean.
2. (a) Yes: each student gives two linked scores. (b) No: the days were chosen separately, so no day in one sample is linked to a day in the other.

3. (a) No: "how many more" asks for an estimate, so use a confidence interval. (b) Yes: "is there evidence" asks for a test.
4. A one-sample t interval for μ, the true mean length of all boards cut today: one group, a quantitative variable, and the goal is to estimate.
5. A two-sample t interval for $\mu_H - \mu_J$, the difference in true mean daily milk yield. Milk yield is quantitative (means), the two breeds are separate, unlinked samples (independent), and "how much more" calls for an estimate.
6. A paired t test on the differences (with exercise minus without), $H_0 : \mu_d = 0$, $H_a : \mu_d > 0$. Each singer gives two linked times, so the data are paired; a two-sample test would wrongly treat them as independent and would let the large differences between singers hide the effect of the exercise.
7. One-sample t interval: $SE = \dfrac{0.12}{\sqrt{16}} = 0.03$; df = 15, $t^* = 2.131$; $15.92 \pm 2.131(0.03) = 15.92 \pm 0.06393$, so $(15.856, 15.984)$ ounces.
8. One-proportion z test: $\hat{p} = 0.85$; $\sigma_{\hat{p}} = \sqrt{\dfrac{0.9(0.1)}{200}} \approx 0.021213$; $z = \dfrac{0.85 - 0.90}{0.021213} \approx -2.36$; P-value ≈ 0.0092 (table: 0.0091). Since $0.0092 < 0.05$, reject H_0: there is convincing evidence that fewer than 90% of the company's pepper seeds germinate.
9. Each bale gives a measured number (for example 14.2% moisture), so the variable is quantitative even though its units are percent. A proportion procedure is for a categorical variable (each bale is "too wet" or "not too wet"). To estimate the average moisture he should use a one-sample t interval for μ.
10. One sample, quantitative, test: one-sample t test. $H_0 : \mu = 16$, $H_a : \mu < 16$. $\bar{x} = \dfrac{110.6}{7} = 15.8$; squared deviations add to 0.16, $s = \sqrt{\dfrac{0.16}{6}} \approx 0.163299$; $SE = \dfrac{0.163299}{\sqrt{7}} \approx 0.061721$; $t = \dfrac{15.8 - 16}{0.061721} \approx -3.24$, df = 6; P-value ≈ 0.0088 (table: between 0.005 and 0.01). Since $0.0088 < 0.05$, reject H_0: there is convincing evidence that the bakery's loaves weigh less than 16 ounces on average.
11. Categorical, two independent samples, estimate: two-proportion z interval. $\hat{p}_1 = 0.32$, $\hat{p}_2 = 0.24$; $SE = \sqrt{\dfrac{0.32(0.68)}{300} + \dfrac{0.24(0.76)}{250}} = \sqrt{0.00145493} \approx 0.038144$; $0.08 \pm 1.96(0.038144) = 0.08 \pm 0.074761$, so $(0.005, 0.155)$.
12. (a) Two-sample t interval; both samples have at least 30 plants ($31 \geq 30$ and $41 \geq 30$), so each sample mean is approximately normal by the central limit theorem. (b) $SE = \sqrt{\dfrac{225}{31} + \dfrac{324}{41}} = \sqrt{15.16050} \approx 3.89365$; df = 30, $t^* = 2.042$; 9 ± 7.95083, so $(1.0, 17.0)$ cm.

13. Each hen is counted twice, so the data are paired: paired t test on d = April minus March: 3, 1, 3, 4, 1, 0, 2. State: $H_0 : \mu_d = 0$, $H_a : \mu_d > 0$, where μ_d is the true mean increase in eggs per hen from March to April for the flock; $\alpha = 0.05$. Plan: random sample of 7 hens; 7 is at most 10% of 80; the differences show no strong skewness or outliers. Do: $\bar{x}_d = \frac{14}{7} = 2$; squared deviations add to 12, $s_d = \sqrt{2} \approx 1.41421$; $SE = \frac{1.41421}{\sqrt{7}} \approx 0.53452$; $t = \frac{2}{0.53452} \approx 3.74$, df = 6; P-value ≈ 0.0048 (table: between 0.0025 and 0.005). Conclude: $0.0048 < 0.05$, so reject H_0. There is convincing evidence that hens in Ruth's flock lay more eggs on average in April than in March.
14. No. These records cover every family in the parish (a census), so 62% and 48% ARE the parameters for last year, not estimates. Inference is for reasoning from a sample to a larger population or from random assignment to cause; here there is no sampling variability to account for, and the difference of 14 percentage points is simply the true difference.

Unit 12 test

You have 60 minutes. You may use a scientific or graphing calculator and the printed standard normal and t tables. Show your work. For full inference problems, show all four steps (State, Plan, Do, Conclude) and check the conditions. For two-sample t procedures use the conservative df = the smaller of $n_1 - 1$ and $n_2 - 1$. Keep full precision until the last step; round z and t to 2 decimal places and P-values to 4 decimal places unless the problem says otherwise. Answers from the tables may differ from technology in the last digit.

1. **5 PTS** A gardener opened a random sample of 80 pea pods of variety X and found that 56 held at least 8 peas; in a random sample of 120 pods of variety Y, 72 held at least 8 peas. Find (a) $\hat{p}_X - \hat{p}_Y$ and (b) the pooled proportion $\hat{p}_c$.

2. **5 PTS** Random samples of pumpkins from two patches gave: Patch 1, $n = 10$, $s = 3.0$ pounds; Patch 2, $n = 18$, $s = 4.2$ pounds. Find (a) the standard error of $\bar{x}_1 - \bar{x}_2$, rounded to 4 decimal places, and (b) the conservative df.

3. **5 PTS** Are the data paired? Answer yes or no for each. (a) The grip strength of the left hand and of the right hand for each of 12 students. (b) The mile times of 10 randomly chosen sophomores and 10 randomly chosen seniors.

4. **6 PTS** In one large town 60% of households own a dog; in another, 45% do. Random samples of 150 households are taken from each town. Find the mean and standard deviation of the sampling distribution of $\hat{p}_1 - \hat{p}_2$ (first town minus second). Round the standard deviation to 4 decimal places.

5. **6 PTS** A farmer randomly assigned rows of a field to a new and an old watermelon variety. A 95% confidence interval for $\mu_{new} - \mu_{old}$, the difference in mean melon weight, is $(0.8, 3.4)$ pounds. (a) Interpret the interval in context. (b) At the $\alpha = 0.05$ level, would a two-sided test give convincing evidence that the mean weights differ? Answer yes or no.

6. **6 PTS** Name the correct inference procedure for each situation and give a one-sentence reason. (a) A beekeeper weighs the honey from the same 12 hives last year and this year to see whether the mean harvest changed. (b) A library takes random samples of 150 sophomores and 150 seniors at a large school, asks each whether they read a book for pleasure last month, and wants to estimate the difference in the two proportions.

7. **8 PTS** Henry tested $H_0 : p_1 = p_2$ for the germination rates of two seed lots. In his z statistic he used the standard error $\sqrt{\dfrac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \dfrac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$. Which standard error should the test use, and why? When is Henry's standard error the right one to use?

8. **8 PTS** A bird club randomly assigned 220 new nesting boxes to be painted green (120 boxes) or left unpainted (100 boxes). By June, 78 green boxes and 52 unpainted boxes were occupied. Check the conditions and construct a 90% confidence interval for $p_G - p_U$. Round the endpoints to 3 decimal places.

9. **8 PTS** In a random sample of 400 households in a county with about 12,000 households, 248 keep a vegetable garden. Name the procedure, check the conditions, and construct a 99% confidence interval for the proportion of all households in the county that keep a garden. Round the endpoints to 3 decimal places.

10. **8 PTS** A carpenter compared an old and a new saw blade by cutting each of 6 different boards once with each blade, in a random order, and timing the cuts (table). Construct a 90% confidence interval for the mean difference in cutting time (old minus new). Round the endpoints to 1 decimal place.

Cutting times with two saw blades (seconds)

Board	Old blade	New blade
1	31	27
2	25	23
3	40	35
4	28	25
5	45	39
6	35	31

11. **9 PTS** A random sample of 180 winter days from Town A's weather records and a separate random sample of 150 winter days from Town B's records were checked for measurable snowfall (table). Do the data give convincing evidence at the $\alpha = 0.05$ level that Town A has a higher proportion of snowy winter days? Use the four-step process.

Winter days with measurable snowfall

Station	Snow	No snow	Total days sampled
Town A	54	126	180
Town B	30	120	150
Total	84	246	330

12. **9 PTS** A tree farmer measured random samples of five-year-old Fraser firs and Balsam firs from two large plantings (table). Histograms of both samples show no strong skewness or outliers. Test at the $\alpha = 0.05$ level whether the mean heights of the two species differ. Use the four-step process with the conservative df.

Heights of five-year-old Christmas trees

Species	Number of trees	Mean height (cm)	SD (cm)
Fraser fir	15	182	12
Balsam fir	21	171	16

13. **9 PTS** Nine members of an astronomy club each counted the stars visible in the same patch of sky twice on one night, once with a light-pollution filter and once without, in a random order. The differences (with filter minus without) had $\bar{x}_d = 6.0$ stars and $s_d = 7.5$ stars, and a dotplot of the differences showed no outliers. Do the data give convincing evidence at the $\alpha = 0.05$ level that the filter increases the mean count? Use the four-step process.

14. **8 PTS** A track coach wants to compare two kinds of starting blocks using 16 sprinters. Plan 1: randomly assign 8 sprinters to each kind of block and compare mean 40-meter times. Plan 2: each sprinter runs once with each kind of block, in a random order. (a) Which plan is better, and why? (b) Which inference procedure goes with each plan?

UNIT 12 TEST: ANSWER KEY

1. (a) $\frac{56}{80} - \frac{72}{120} = 0.70 - 0.60 = 0.10$ (b) $\hat{p}_c = \frac{56 + 72}{80 + 120} = \frac{128}{200} = 0.64$
2. (a) $SE = \sqrt{\frac{9}{10} + \frac{17.64}{18}} = \sqrt{0.9 + 0.98} = \sqrt{1.88} \approx 1.3711$ pounds (b) df = smaller of 9 and 17 = 9
3. (a) Yes: each student gives two linked measurements. (b) No: the two groups are separate students with no partners.
4. Mean $= 0.60 - 0.45 = 0.15$; SD $= \sqrt{\frac{0.6(0.4)}{150} + \frac{0.45(0.55)}{150}} = \sqrt{0.00325} \approx 0.0570$
5. (a) We are 95% confident that melons of the new variety weigh between 0.8 and 3.4 pounds more on average than melons of the old variety grown in rows like these. (b) Yes: 0 is not in the interval, so a two-sided test at $\alpha = 0.05$ would reject $H_0 : \mu_{new} = \mu_{old}$.
6. (a) Paired t test on the differences (this year minus last year): the variable is quantitative and each hive is measured twice. (b) Two-proportion z interval: each answer is yes or no (categorical), the two samples are independent, and the goal is to estimate.
7. A test computes its P-value assuming $H_0 : p_1 = p_2$ is true. Under that assumption both groups share one proportion, and the best estimate of it combines all the successes and all the individuals: $\hat{p}_c = \frac{X_1 + X_2}{n_1 + n_2}$. So the test should use $\sqrt{\hat{p}_c(1 - \hat{p}_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$. Henry's unpooled standard error is the right one for a confidence interval for $p_1 - p_2$, which assumes nothing about the two proportions being equal; it estimates how far apart they are, so it uses each group's own $\hat{p}$.
8. Random assignment; Large Counts: 78, 42, 52 and 48 are all at least 10. $\hat{p}_G = 0.65$, $\hat{p}_U = 0.52$; $SE = \sqrt{\frac{0.65(0.35)}{120} + \frac{0.52(0.48)}{100}} = \sqrt{0.0043918} \approx 0.066271$; $0.13 \pm 1.645(0.066271) = 0.13 \pm 0.109016$, so $(0.021, 0.239)$. We are 90% confident that painting boxes green raises the occupancy rate by between 2.1 and 23.9 percentage points.
9. One-proportion z interval (one sample, categorical, estimate). Random sample; 400 is less than 10% of the county's 12,000 households; 248 and 152 are at least 10. $\hat{p} = 0.62$; $SE = \sqrt{\frac{0.62(0.38)}{400}} \approx 0.024269$; $0.62 \pm 2.576(0.024269) = 0.62 \pm 0.062518$, so $(0.557, 0.683)$.
10. Paired t interval. d = old minus new: 4, 2, 5, 3, 6, 4; $\bar{x}_d = \frac{24}{6} = 4$; squared deviations add to 10, $s_d = \sqrt{2} \approx 1.41421$; $SE = \frac{1.41421}{\sqrt{6}} \approx 0.57735$; df = 5, $t^* = 2.015$; 4 ± 1.16336, so $(2.8, 5.2)$ seconds. We are 90% confident that the new blade cuts boards like these between 2.8 and 5.2 seconds faster on average.

11. State: $H_0 : p_A = p_B$, $H_a : p_A > p_B$, where p is the true proportion of winter days with measurable snow at each town; $\alpha = 0.05$. Plan: two-proportion z test; independent random samples; each sample is less than 10% of all winter days on record; $\hat{p}_c = \dfrac{84}{330} \approx 0.254545$, and $180(0.2545) \approx 45.8$, $180(0.7455) \approx 134.2$, $150(0.2545) \approx 38.2$, $150(0.7455) \approx 111.8$ are all at least 10. Do: $\hat{p}_A = 0.30$, $\hat{p}_B = 0.20$; $SE = \sqrt{0.254545(0.745455)\left(\dfrac{1}{180} + \dfrac{1}{150}\right)} \approx 0.048158$; $z = \dfrac{0.10}{0.048158} \approx 2.08$; P-value ≈ 0.0189 (table: 0.0188). Conclude: $0.0189 < 0.05$, so reject H_0. There is convincing evidence that a higher proportion of winter days bring measurable snow in Town A than in Town B.

12. State: $H_0 : \mu_F = \mu_B$, $H_a : \mu_F \neq \mu_B$, where μ is the true mean height of five-year-old trees of each species in these plantings; $\alpha = 0.05$. Plan: two-sample t test; independent random samples; each sample is less than 10% of a large planting; histograms show no strong skewness or outliers. Do: $SE = \sqrt{\dfrac{144}{15} + \dfrac{256}{21}} = \sqrt{21.790476} \approx 4.66803$; $t = \dfrac{182 - 171}{4.66803} \approx 2.36$; df = smaller of 14 and 20 = 14; two-sided P-value ≈ 0.0335 (table: between 0.02 and 0.04). Conclude: $0.0335 < 0.05$, so reject H_0. There is convincing evidence that the mean heights of five-year-old Fraser and Balsam firs in these plantings differ; the Frasers are taller on average.

13. State: $H_0 : \mu_d = 0$, $H_a : \mu_d > 0$, where μ_d is the true mean difference in star count (with filter minus without) for observers like these; $\alpha = 0.05$. Plan: paired t test (each observer counts twice); random order of the two conditions; the dotplot of differences shows no outliers or strong skewness. Do: $SE = \dfrac{7.5}{\sqrt{9}} = 2.5$; $t = \dfrac{6.0}{2.5} = 2.40$, df = 8; P-value ≈ 0.0216 (table: between 0.02 and 0.025). Conclude: $0.0216 < 0.05$, so reject H_0. There is convincing evidence that the filter increases the mean number of stars counted by observers like these.

14. (a) Plan 2. Sprinters differ a great deal in speed, and in Plan 1 that sprinter-to-sprinter variation goes into the standard error and can hide a small difference between the blocks. In Plan 2 each sprinter serves as their own comparison, so subtracting within each sprinter removes that variation; the differences vary much less, giving a more powerful test. (b) Plan 1: two-sample t procedures. Plan 2: paired t procedures on the 16 differences.

UNIT 13

Chi-Square Tests

Units 11 and 12 tested claims about one or two proportions and one or two means. But many questions involve categorical variables with more than two categories: Do a hatchery's chicks come in the breed mix it claims? Is there an association between morning clouds and afternoon rain? Do three seed suppliers have the same germination rates? This unit introduces one statistic, chi-square, that compares observed counts with the counts we would expect if the null hypothesis were true, and it explains why the statistic squares each difference and divides by the expected count.

Lesson 13.1 uses chi-square to test whether one categorical variable fits a claimed distribution (the goodness-of-fit test, including Mendel's famous peas). Lesson 13.2 applies the same idea to two-way tables, first met in Lesson 1.3, to test for homogeneity among several groups or for independence between two variables. Both lessons use the four-step process (State, Plan, Do, Conclude) from Units 10 to 12, and the conclusions are worded with the same care: convincing evidence or not, never proof.

Lesson 13.3 closes the course. It gathers the year's lessons into a habit of judgment: asking how data were collected, compared to what, whether chance could explain a result, whether it shows cause, and whether it matters. It also treats honesty with data as a duty. The statistics project that follows gives students the chance to put all of this into practice with their own data.

PREREQUISITES

- Two-way tables and conditional distributions (Lesson 1.3)
- Independence of events and P(A and B) = P(A)·P(B) (Lesson 6.4)
- The logic of significance tests: hypotheses, P-values, the significance level α, and why we fail to reject rather than accept H_0 (Lessons 11.1 and 11.2)
- The four-step process State, Plan, Do, Conclude, with conditions checked first (Units 10 to 12)
- The pooled two-proportion z test (Lesson 12.1)
- Random sampling, random assignment and the scope of conclusions (Unit 5)

MASTERY CHECKLIST

☐ I can state hypotheses for a chi-square goodness-of-fit test in symbols and in words.
☐ I can compute expected counts for a claimed distribution and check that they add to n.
☐ I can compute the chi-square statistic and explain why it squares each difference and divides by the expected count.
☐ I can find the degrees of freedom and the P-value (always the right tail) from the chi-square table or technology.
☐ I can check the Random, 10% and Large Counts conditions, using expected counts for Large Counts.
☐ I can tell a test for homogeneity from a test for independence by how the data were produced.

☐ I can compute expected counts in a two-way table with (row total)(column total)/(table total) and explain why the formula works.
☐ I can explain why a two-way table has (r − 1)(c − 1) degrees of freedom.
☐ I can carry out a chi-square test with the four-step process and follow up by looking at the largest components.
☐ I can evaluate a statistical claim by asking how the data were collected, compared to what, chance or real, correlation or cause, and whether it matters in practice.
☐ I can explain why cherry-picking and running many tests mislead, and report results honestly.

LESSON 13.1 · 2 SESSIONS OF 45 MINUTES

The Chi-Square Goodness-of-Fit Test

OBJECTIVES

- Compute expected counts for a claimed distribution and check the conditions for a chi-square test.
- Calculate the chi-square statistic and explain why it squares each difference and divides by the expected count.
- Find the degrees of freedom and the P-value from the chi-square table or technology.
- Carry out a goodness-of-fit test with the four-step process, interpret the result in context, and follow up by examining the largest components.

VIDEO

WATCH

Chi-square Tests for One-way Tables

jbstatistics · 9:07

When to watch: Before Example 1: the chi-square goodness-of-fit test.

https://www.youtube.com/watch?v=gkgyg-eR0TQ

BACKUP

Chi Square Goodness of Fit Test

ProfRobBob · 20:34

When to watch: Extra practice: a goodness-of-fit test, worked on a calculator.

https://www.youtube.com/watch?v=Op4BBNj-ML4

KEY TERMS

goodness-of-fit test	A significance test of whether the distribution of one categorical variable in a population matches a claimed (hypothesized) distribution.
expected count	The count we would expect in a category if the null hypothesis were exactly true: the sample size times the claimed proportion for that category.
chi-square statistic	The number $\chi^2 = \sum \dfrac{(\text{observed} - \text{expected})^2}{\text{expected}}$, which measures how far all the observed counts are from the expected counts, taken together.

component	One term $\dfrac{(\text{observed} - \text{expected})^2}{\text{expected}}$ of the chi-square sum; it shows how much one category contributes to the total.
chi-square distribution	A family of right-skewed density curves that take only values of 0 or more; each one is named by its degrees of freedom, and its mean equals its degrees of freedom.
degrees of freedom (goodness of fit)	The number of categories minus 1: the number of counts that are free to vary once the total n is fixed.

CORE IDEAS

In Units 11 and 12 you tested claims about one proportion or compared two. Many categorical variables, though, have more than two categories: the breed of a chick, the color of a flower, the day of the week a board is rejected at a sawmill. A **goodness-of-fit test** asks whether the distribution of such a variable in the population fits a claimed distribution, or whether the sample's mismatch is too large to blame on chance alone. The null hypothesis lists the claimed proportions, for example H_0: $p_1 = 0.40$, $p_2 = 0.35$, $p_3 = 0.25$. The alternative is simply H_a: at least one of the proportions differs from its claimed value. There is no one-sided version: any kind of mismatch counts against H_0.

First find the **expected counts**, the counts we would expect if H_0 were exactly true: expected count $= n \cdot p_i$. For example, a wildflower seed mix is advertised as 40% poppies, 35% cornflowers and 25% daisies. Grace identifies 120 seedlings grown from a random scoop of the mix. The expected counts are $120 \times 0.40 = 48$, $120 \times 0.35 = 42$ and $120 \times 0.25 = 30$. Expected counts need not be whole numbers (31.5 is fine), and they always add to n, just as the observed counts do. Grace's observed counts are in the table.

Grace's wildflower seedlings

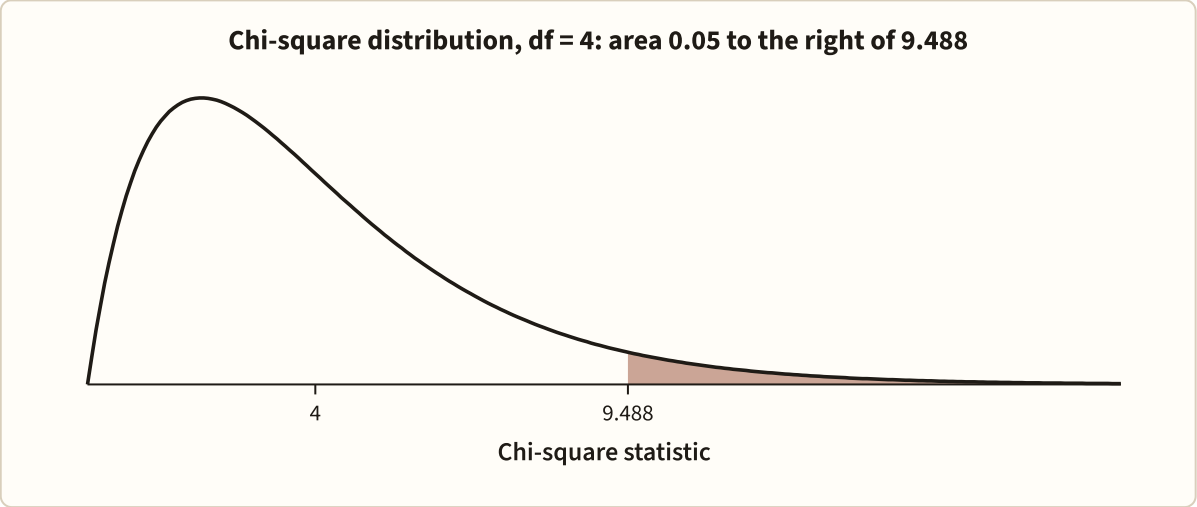
Flower	Observed	Expected	O - E	(O - E)^2 / E
Poppy	54	48	6	0.75
Cornflower	33	42	-9	1.93
Daisy	33	30	3	0.3
Total	120	120	0	2.98

How should three misses be combined into one number? Adding the differences $O - E$ is useless: they always add to 0 (here $6 - 9 + 3 = 0$), because both columns add to n. As with deviations from the mean in Lesson 2.4, we square the differences so that misses in either direction count. Then we divide each square by its expected count, because a miss only means something compared with what we expected. Missing by 9 when you expected 42 is a noticeable miss; missing by 9 when you expected 1,000 is almost nothing. Each term $(O - E)^2/E$ is a **component**, and their sum is the **chi-square**

statistic (chi is the Greek letter χ, pronounced *kye*). For the wildflowers, $\chi^2 = 0.75 + 1.93 + 0.30 = 2.98$. A χ^2 of 0 would mean a perfect fit; the larger χ^2 is, the worse the fit.

Is 2.98 large? If H_0 is true and the expected counts are not too small, the statistic χ^2 varies from sample to sample according to a **chi-square distribution**. There is a different curve for each number of **degrees of freedom**, and for a goodness-of-fit test $df = (\text{number of categories}) - 1$. Why one less? The counts must add to n, so once you know all but one of them, the last one is forced. Only $k - 1$ of the k counts are free to vary, just as only $n - 1$ deviations from $\bar{x}$ are free. The wildflower test has $df = 3 - 1 = 2$.

Every chi-square curve starts at 0 (a sum of squares can never be negative), is skewed right, and has a mean equal to its degrees of freedom. So a χ^2 near its df is just what chance produces, and only a χ^2 well above its df is surprising. As df grows, the curve moves to the right and becomes more symmetric. In the curve below (df = 4) the shaded tail to the right of 9.488 has area 0.05: if H_0 were true, a χ^2 above 9.488 would happen only 5% of the time.



Because every kind of mismatch makes χ^2 bigger, the P-value is always the area to the RIGHT of the observed χ^2 under the chi-square curve with the correct df. With technology, use a chi-square cdf command, such as χ^2cdf(2.98, 1E99, 2) on a TI-84. With the table, go to the row for your df and see between which entries your χ^2 falls; the column headings give bounds for the P-value. For the wildflowers, 2.98 is less than 4.605 (the 0.10 entry for df = 2), so P > 0.10; technology gives P = 0.2255. A mismatch this size happens often by chance alone, so Grace's seedlings give no convincing evidence against the advertised mix.

The chi-square curve is only an approximation to the true sampling distribution of χ^2, so check the conditions first. **Random:** the data come from a random sample (or a random chance process). **10%:** when sampling without replacement, n is at most 10% of the population. **Large Counts:** every EXPECTED count is at least 5; with smaller expected counts, a tiny category can produce a huge component and the curve gives wrong P-values. If an expected count is too small, take a larger sample or combine categories. Finally, when you reject H_0, the test tells you THAT the distribution differs but

not WHERE. Follow up by finding the largest components and comparing observed with expected counts in those categories.

KEY FORMULAS

$$\text{expected count} = n \cdot p_i$$

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$df = \text{number of categories} - 1$$

$$\text{P-value} = P(\chi^2 \geq \text{observed value}), \text{ the area to the right}$$

From history. Karl Pearson, an English mathematician, introduced the chi-square test in 1900, giving scientists their first general method for judging whether observed counts fit a theory. A classic example is older: Gregor Mendel, an Augustinian friar who bred pea plants in the garden of his monastery in Brno, published results in 1866 showing that his theory of heredity predicted a 9:3:3:1 ratio of seed types in his crosses. Students still test Mendel's counts against that ratio (practice problem 12).

Practice habit. Before computing anything, make a table with the observed and expected counts side by side. Check that both columns add to n and that every expected count is at least 5.

WORKED EXAMPLES

Example 1. A hatchery says the chicks in its mixed-breed orders are 50% Rhode Island Reds, 30% Leghorns and 20% Plymouth Rocks. Henry takes a random sample of 200 chicks from the hatchery's spring orders; the counts are in the table. Is there convincing evidence at the $\alpha = 0.05$ level that the breed distribution differs from the hatchery's claim? Give χ^2 to 2 decimal places and the P-value to 4 decimal places.

Henry's sample of 200 chicks

Breed	Claimed proportion	Observed count
Rhode Island Red	0.5	112
Leghorn	0.3	50
Plymouth Rock	0.2	38
Total	1	200

1. State: H_0: $p_R = 0.50$, $p_L = 0.30$, $p_P = 0.20$ (the hatchery's claim is correct). H_a: at least one of these proportions is different. Use $\alpha = 0.05$.
2. Plan: chi-square goodness-of-fit test. Random: Henry took a random sample. 10%: 200 chicks is less than 10% of all the chicks the hatchery sells in spring.
3. Expected counts: $200 \times 0.50 = 100$, $200 \times 0.30 = 60$, $200 \times 0.20 = 40$. They add to 200, and all are at least 5, so the Large Counts condition is met.
4. Do: components $\dfrac{(112 - 100)^2}{100} = 1.44$, $\dfrac{(50 - 60)^2}{60} = 1.667$, $\dfrac{(38 - 40)^2}{40} = 0.10$.
5. $\chi^2 = 1.44 + 1.667 + 0.10 = 3.21$, with $df = 3 - 1 = 2$.
6. P-value $= P(\chi^2 \geq 3.21) = 0.2012$ (table: 3.21 is less than 4.605, so P > 0.10).
7. Conclude: because $0.2012 > 0.05$, we fail to reject H_0. There is not convincing evidence that the breed distribution of the hatchery's chicks differs from the claimed 50%, 30% and 20%. (This does not prove the claim is exactly right; it only says the data are consistent with it.)

Answer: $\chi^2 = 3.21$, $df = 2$, P-value $= 0.2012$. Fail to reject H_0: there is not convincing evidence that the breed mix differs from the hatchery's claim.

Example 2. A sawmill foreman wonders whether boards rejected for defects are equally likely to come from each of the five weekdays. A random sample of 100 rejected boards from the past year's records gives the counts in the table. Do these data give convincing evidence at the $\alpha = 0.05$ level that rejected boards are not equally spread across the weekdays? If so, which day contributes most to the difference? Give χ^2 to 2 decimal places and the P-value to 4 decimal places.

Rejected boards by weekday

Day	Mon	Tue	Wed	Thu	Fri	Total
Rejected boards	32	18	15	14	21	100

1. State: H_0: $p_{\text{Mon}} = p_{\text{Tue}} = p_{\text{Wed}} = p_{\text{Thu}} = p_{\text{Fri}} = 0.20$ (rejected boards are equally likely to come from each weekday). H_a: at least one of these proportions differs from 0.20. Use $\alpha = 0.05$.
2. Plan: chi-square goodness-of-fit test. Random: a random sample of rejected boards. 10%: the mill rejected well over 1,000 boards during the year, so 100 is less than 10%.
3. Expected counts: $100 \times 0.20 = 20$ for each day, all at least 5.
4. Do: components Mon $\dfrac{(32-20)^2}{20} = 7.20$; Tue $\dfrac{(18-20)^2}{20} = 0.20$; Wed $\dfrac{(15-20)^2}{20} = 1.25$; Thu $\dfrac{(14-20)^2}{20} = 1.80$; Fri $\dfrac{(21-20)^2}{20} = 0.05$.
5. $\chi^2 = 7.20 + 0.20 + 1.25 + 1.80 + 0.05 = 10.50$, with $df = 5 - 1 = 4$.
6. P-value $= P(\chi^2 \geq 10.50) = 0.0328$ (table: 10.50 is between 9.488 and 11.143, so $0.025 < P < 0.05$).
7. Conclude: because $0.0328 < 0.05$, we reject H_0. There is convincing evidence that rejected boards are not equally spread across the five weekdays.
8. Follow-up: Monday's component, 7.20, is by far the largest (about 69% of the total 10.50). Monday had 32 rejected boards when only 20 were expected, so the foreman should look at what is different about Monday's work.

Answer: $\chi^2 = 10.50$, $df = 4$, P-value $= 0.0328$. Reject H_0: there is convincing evidence that rejected boards are not equally spread across weekdays. Monday contributes most (component 7.20): 32 rejects observed versus 20 expected.

PRACTICE

1. EASY A garden center says its bags of onion sets are 50% yellow, 30% red and 20% white. Ruth plants 80 sets chosen at random from a large bin. If the claim is true, how many sets of each color should Ruth expect?

2. EASY To check whether a six-sided board-game die is fair, Samuel rolls it 120 times and counts how often each face comes up. (a) What is the expected count for each face if the die is fair? (b) How many degrees of freedom does the goodness-of-fit test have?

3. EASY Grace wonders whether the birds that visit her feeder are equally divided among four kinds: cardinals, chickadees, finches and sparrows. She records the kind of bird at randomly chosen times. State the hypotheses for a goodness-of-fit test in symbols and in words.

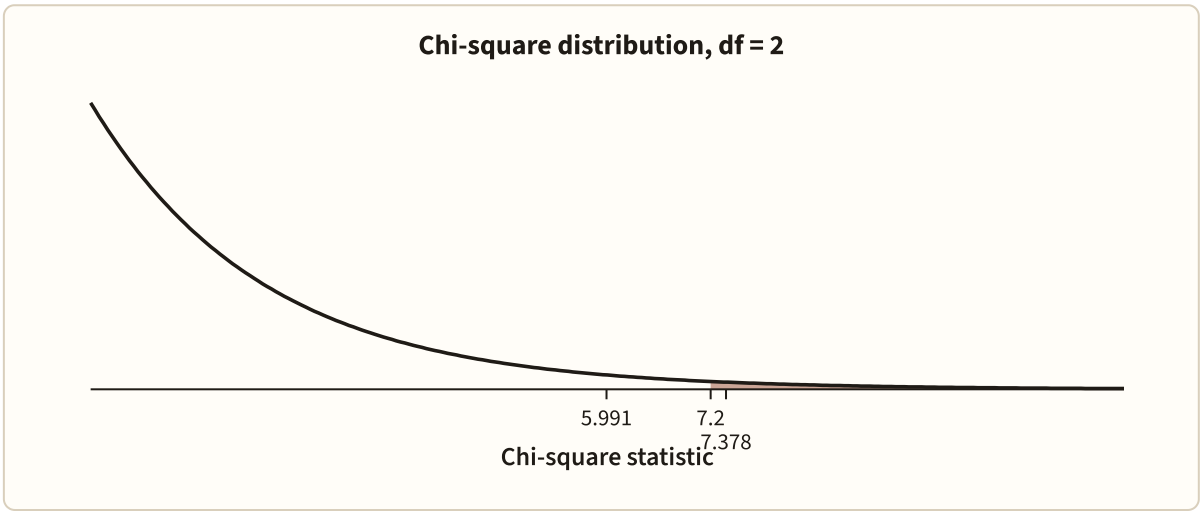
4. EASY A bakery owner believes that $\frac{1}{3}$ of the loaves sold are white bread, $\frac{5}{12}$ are whole wheat and $\frac{1}{4}$ are rye. The table shows a random sample of 60 loaves sold. Find the expected counts and the chi-square statistic (2 decimal places).

Random sample of 60 loaves sold

Bread	White	Whole wheat	Rye	Total
Loaves sold	18	30	12	60

5. MEDIUM A forester says the trees in a woodlot are 50% oak, 30% maple, 10% hickory and 10% walnut. Samuel plans to identify 40 randomly chosen trees and test the claim. Is the Large Counts condition met? Explain, and suggest a fix if it is not.

6. MEDIUM A weather station classifies summer days as clear, partly cloudy or rainy, and a goodness-of-fit test compares a random sample of days with the long-term percentages. The test gives $\chi^2 = 7.20$. (a) How many degrees of freedom does the test have? (b) Find the P-value (4 decimal places); it is the shaded area in the graph.



7. MEDIUM A soccer goalkeeping coach claims that penalty kicks in a high school league go 45% to the keeper's left, 20% to the center and 35% to the keeper's right. Game films of a random sample of 100 penalty kicks from the league give the counts in the table. Carry out a test at the $\alpha = 0.05$ level. Give χ^2 to 2 decimal places and the P-value to 4 decimal places.

Direction of 100 penalty kicks

Direction	Left	Center	Right	Total
Claimed proportion	0.45	0.2	0.35	1
Observed count	52	12	36	100

8. MEDIUM In two different goodness-of-fit tests, one category's observed count is 10 more than its expected count. In the first test that expected count is 20; in the second it is 500. (a) Find the component for each test. (b) Explain why it makes sense that the two components are so different.

9. MEDIUM A forester compares the species of 150 randomly chosen trees in a state forest with the mix of five species recorded in a survey 30 years ago. The goodness-of-fit test gives $\chi^2 = 12.90$ with $df = 4$ and P-value $= 0.0118$. Interpret the P-value in context.

10. HARD A board game's spinner has five equal sections numbered 1 to 5. Grace suspects it is not fair and spins it 150 times, with the results in the table. Carry out a test at the $\alpha = 0.05$ level. Give χ^2 to 2 decimal places and the P-value to 4 decimal places.

150 spins of a five-section spinner

Section	1	2	3	4	5	Total
Count	38	24	33	27	28	150

11. HARD A plant breeder crosses two tomato plants. Genetic theory predicts offspring that are tall with red fruit, tall with yellow fruit, short with red fruit and short with yellow fruit in the ratio 9:3:3:1. The 160 offspring (treated as a random sample of possible offspring) are counted in the table. Test at the $\alpha = 0.05$ level whether the offspring fit the 9:3:3:1 ratio, and if you reject, name the category that contributes most. Give χ^2 to 2 decimal places and the P-value to 4 decimal places.

Offspring of a tomato cross

Type	Tall, red	Tall, yellow	Short, red	Short, yellow	Total
Plants	81	30	30	19	160

12. **HARD** In one experiment Gregor Mendel classified 556 pea seeds, with the counts in the table. His theory predicted the ratio 9:3:3:1 for round yellow, round green, wrinkled yellow and wrinkled green. (a) Find χ^2 (2 decimal places). (b) Using the chi-square table, what is the decision at the $\alpha = 0.05$ level?

Mendel's pea seeds

Seed type	Round yellow	Round green	Wrinkled yellow	Wrinkled green	Total
Observed	315	108	101	32	556

13. **CHALLENGE** A seed company claims that 75% of its seeds germinate. In a random sample of 300 seeds, 210 germinate. (a) Treat this as a goodness-of-fit test with two categories (germinate, do not germinate) and find χ^2. (b) Find the z statistic for the one-sample z test of H_0: $p = 0.75$ against H_a: $p \neq 0.75$ (Lesson 11.2). (c) Check that $z^2 = \chi^2$, and give the P-value (4 decimal places) that both tests share.

COMMON MISTAKES

- Using percents or proportions instead of counts in the formula. The chi-square statistic must be computed from counts; convert any percents to counts (percent of n) before you begin.
- Checking the OBSERVED counts for the Large Counts condition. It is the EXPECTED counts that must all be at least 5.
- Using df = n − 1. For goodness of fit, df = (number of categories) − 1: a sample of 200 chicks in 3 breeds has df = 2, not 199.
- Using the left tail or doubling the P-value. Only a large χ^2 is evidence against H_0, so the P-value is always the area to the right.

ANSWERS TO PRACTICE

1. Yellow $80 \times 0.50 = 40$, red $80 \times 0.30 = 24$, white $80 \times 0.20 = 16$ (they add to 80).
2. (a) $120 \times \dfrac{1}{6} = 20$ for each face. (b) $df = 6 - 1 = 5$.
3. H_0: $p_{\text{card}} = p_{\text{chick}} = p_{\text{finch}} = p_{\text{sparrow}} = 0.25$; the four kinds of birds visit the feeder equally often. H_a: at least one of these proportions differs from 0.25; the four kinds do not visit equally often.
4. Expected: $60 \times \dfrac{1}{3} = 20$, $60 \times \dfrac{5}{12} = 25$, $60 \times \dfrac{1}{4} = 15$. $\chi^2 = \dfrac{(18-20)^2}{20} + \dfrac{(30-25)^2}{25} + \dfrac{(12-15)^2}{15} = 0.20 + 1.00 + 0.60 = 1.80$.
5. No. The expected counts are $40 \times 0.50 = 20$, $40 \times 0.30 = 12$, $40 \times 0.10 = 4$ and 4. Two expected counts are below 5 (the observed counts do not matter for this check). Samuel could sample at least 50 trees (then $50 \times 0.10 = 5$) or combine hickory and walnut into one category with expected count 8.

6. (a) $df = 3 - 1 = 2$. (b) P-value $= P(\chi^2 \geq 7.20) = 0.0273$ (table: 7.20 is between 5.991 and 7.378, so $0.025 < P < 0.05$).

7. State: H_0: $p_L = 0.45$, $p_C = 0.20$, $p_R = 0.35$; H_a: at least one proportion differs; $\alpha = 0.05$. Plan: goodness-of-fit test; random sample; 100 is less than 10% of all penalty kicks in the league; expected counts 45, 20, 35 are all at least 5. Do: $\chi^2 = \frac{7^2}{45} + \frac{(-8)^2}{20} + \frac{1^2}{35} = 1.089 + 3.200 + 0.029 = 4.32$, $df = 2$, P-value $= 0.1155$. Conclude: $0.1155 > 0.05$, so fail to reject H_0. There is not convincing evidence that the directions of penalty kicks in this league differ from the coach's claim.

8. (a) $\frac{10^2}{20} = 5$ and $\frac{10^2}{500} = 0.20$. (b) A miss of 10 is large when you expected only 20 (half again as many), but tiny when you expected 500 (a 2% difference). Dividing by the expected count measures each miss relative to the size of what was expected.

9. If the species mix in the forest were still the same as 30 years ago, there would be only about a 1.2% chance of getting a random sample of 150 trees whose counts are this far or farther from the old mix (a χ^2 of 12.90 or more) just by chance. That is good evidence that the mix has changed.

10. State: H_0: $p_1 = p_2 = p_3 = p_4 = p_5 = 0.20$ (the spinner is fair); H_a: at least one proportion differs from 0.20; $\alpha = 0.05$. Plan: goodness-of-fit test; the spins are random and independent (no 10% condition is needed, since we are not sampling from a finite population); expected counts $150 \times 0.20 = 30$ each, all at least 5. Do: $\chi^2 = \frac{8^2 + (-6)^2 + 3^2 + (-3)^2 + (-2)^2}{30} = \frac{122}{30} = 4.07$, $df = 4$, P-value $= 0.3971$. Conclude: $0.3971 > 0.05$, so fail to reject H_0. There is not convincing evidence that the spinner is unfair. (This does not prove it is fair; many more spins might detect a small bias.)

11. State: H_0: the proportions are $\frac{9}{16}, \frac{3}{16}, \frac{3}{16}, \frac{1}{16}$; H_a: at least one differs; $\alpha = 0.05$. Plan: goodness-of-fit test; expected counts $160 \times \frac{9}{16} = 90, 30, 30, 10$, all at least 5. Do: $\chi^2 = \frac{(-9)^2}{90} + 0 + 0 + \frac{9^2}{10} = 0.90 + 8.10 = 9.00$, $df = 3$, P-value $= 0.0293$ (table: between 7.815 and 9.348). Conclude: $0.0293 < 0.05$, so reject H_0; there is convincing evidence that the offspring do not follow the 9:3:3:1 ratio. Short plants with yellow fruit contribute most (8.10 of the 9.00): 19 were observed but only 10 were expected.

12. (a) Expected: $556 \times \frac{9}{16} = 312.75$, $556 \times \frac{3}{16} = 104.25$, 104.25, $556 \times \frac{1}{16} = 34.75$. $\chi^2 = \frac{2.25^2}{312.75} + \frac{3.75^2}{104.25} + \frac{(-3.25)^2}{104.25} + \frac{(-2.75)^2}{34.75} = 0.016 + 0.135 + 0.101 + 0.218 = 0.47$. (b) With $df = 3$, 0.47 is far below 6.251, so P > 0.10 (technology: about 0.93). Fail to reject H_0: Mendel's counts are very consistent with the 9:3:3:1 ratio.

13. (a) Expected: 225 and 75. $\chi^2 = \dfrac{(-15)^2}{225} + \dfrac{15^2}{75} = 1.00 + 3.00 = 4.00$. (b) $\hat{p} = 0.70$; $z = \dfrac{0.70 - 0.75}{\sqrt{\dfrac{0.75 \times 0.25}{300}}} = \dfrac{-0.05}{0.025} = -2.00$. (c) $z^2 = (-2.00)^2 = 4.00 = \chi^2$. Two-sided P-value $= 2(0.02275) = 0.0455$, which equals $P(\chi^2 \geq 4.00)$ with $df = 1$. With two categories, the chi-square test is the two-sided z test in disguise.

LESSON 13.2 · 2 SESSIONS OF 45 MINUTES

Chi-Square Tests for Two-Way Tables

OBJECTIVES

- Distinguish a chi-square test for homogeneity from a test for independence by how the data were produced, and state the hypotheses in context.
- Compute expected counts with (row total)(column total)/(table total) and explain why this formula follows from the null hypothesis.
- Find df = (r − 1)(c − 1), check the conditions, and carry out a chi-square test for a two-way table with the four-step process.
- Interpret the conclusion in context, using the largest components and conditional percents to describe any association.

VIDEO

WATCH

Chi-square Tests of Independence (Chi-square Tests for Two-Way Tables)

jbstatistics · 9:54

When to watch: Before Example 1: chi-square tests for two-way tables.

https://www.youtube.com/watch?v=L1QPBGoDmT0

BACKUP

Chi Square Test for Independence & Homogeneity

ProfRobBob · 18:38

When to watch: Extra practice: tests of independence and homogeneity.

https://www.youtube.com/watch?v=oZKzi2YcPIE

KEY TERMS

chi-square test for homogeneity	A test that compares the distribution of one categorical variable across several populations or treatments, using separate random samples or groups formed by random assignment. H_0: the distribution is the same in every population (or for every treatment).
chi-square test for independence	A test that uses ONE random sample in which two categorical variables are recorded for each individual. H_0: there is no association between the two variables in the population.

expected count (two-way table)	The count expected in a cell if H_0 is true: $\dfrac{(\text{row total})(\text{column total})}{\text{table total}}$.
degrees of freedom (two-way table)	$df = (r-1)(c-1)$ for a table with r rows and c columns: the number of cells that can be filled in freely once all the row and column totals are fixed.
association	Two categorical variables are associated when knowing the value of one changes the distribution of the other (Lesson 1.3).

CORE IDEAS

In Lesson 1.3 you compared conditional distributions in a two-way table to look for an association. At a county fair, the organizers took a random sample of 60 visitors on Friday and a separate random sample of 90 visitors on Saturday and asked which of three food stands each liked best. Kettle corn was the favorite of 40% of the Friday sample (24 of 60) but of only about 23% of the Saturday sample (21 of 90). Samples always differ a little, so the question is the familiar one from Unit 11: are these differences too large to blame on chance alone?

Favorite food stand (observed counts)

Day	Kettle corn	Pie	Barbecue	Total
Friday	24	20	16	60
Saturday	21	40	29	90
Total	45	60	45	150

The chi-square statistic answers two different questions about two-way tables. A **test for homogeneity** compares the distribution of ONE variable across several populations or treatments (Friday visitors versus Saturday visitors; three seed suppliers; three feeds assigned at random). Its H_0 says the distribution is the same for every group. A **test for independence** takes ONE random sample and records TWO variables on each individual (age group and favorite pie for each customer). Its H_0 says the two variables are not associated. To tell them apart, ask how the data were produced: if the design fixed the group sizes in advance (separate samples or treatment groups), it is homogeneity; if one sample was split into categories afterward, it is independence. The arithmetic is identical; only the hypotheses and the wording of the conclusion differ.

Why is the expected count (row total)(column total)/(table total)? Overall, 45 of the 150 visitors (30%) chose kettle corn. If Friday and Saturday visitors really have the same preferences, then 30% of EACH day's sample should choose it: $60 \times \dfrac{45}{150} = 18$ on Friday and $90 \times \dfrac{45}{150} = 27$ on Saturday. That is row total $\times \dfrac{\text{column total}}{\text{table total}}$. For a test of independence the reasoning leads to the same formula: if the

variables are independent, $P(A \text{ and } B) = P(A) \cdot P(B)$, so the expected count is $n \cdot \frac{\text{row total}}{n} \cdot \frac{\text{column total}}{n} = \frac{(\text{row total})(\text{column total})}{n}$. Notice that the expected table has exactly the same row and column totals as the observed table.

Expected counts if both days have the same preferences

Day	Kettle corn	Pie	Barbecue	Total
Friday	18	24	18	60
Saturday	27	36	27	90
Total	45	60	45	150

The chi-square statistic is computed just as in Lesson 13.1, adding $\frac{(O-E)^2}{E}$ over every cell of the table (never the totals). For the fair, $\chi^2 = \frac{6^2}{18} + \frac{(-4)^2}{24} + \frac{(-2)^2}{18} + \frac{(-6)^2}{27} + \frac{4^2}{36} + \frac{2^2}{27} = 2.000 + 0.667 + 0.222 + 1.333 + 0.444 + 0.148 = 4.81$. The degrees of freedom are $df = (r-1)(c-1)$. Why? Once the row and column totals are fixed, choose the first two cells of the Friday row; the third Friday cell is forced by the row total, and the whole Saturday row is forced by the column totals. Only $2 = (2-1)(3-1)$ cells are free.

Only 2 cells are free once the totals are fixed

Day	Kettle corn	Pie	Barbecue	Total
Friday	24	20	forced	60
Saturday	forced	forced	forced	90
Total	45	60	45	150

For the fair, $df = 2$ and the P-value is $P(\chi^2 \geq 4.81) = 0.0900$. Differences this large would show up about 9% of the time by chance alone, so at $\alpha = 0.05$ they are not convincing evidence that Friday and Saturday visitors differ in their favorite stand. The conditions are the same as before. **Random:** one random sample (independence) or independent random samples or random assignment (homogeneity). **10%:** each sample is at most 10% of its population when sampling without replacement. **Large Counts:** every expected count is at least 5.

Rejecting H_0 says the association (or the difference among groups) is too large to blame on chance. It does NOT say the association is strong, and it does not show cause and effect unless the data come from a randomized experiment. Follow up by comparing conditional percents and by finding the largest components, which point to the cells where observed and expected counts disagree most. One more connection: a 2×2 table has $df = 1$, and its χ^2 is exactly the square of the pooled two-proportion

z statistic from Lesson 12.1, with the same P-value as the two-sided z test. The chi-square test extends that test to more than two groups or more than two outcomes.

KEY FORMULAS

$$\text{expected count} = \frac{(\text{row total})(\text{column total})}{\text{table total}}$$

$$\chi^2 = \sum_{\text{all cells}} \frac{(\text{observed} - \text{expected})^2}{\text{expected}}$$

$$df = (\text{number of rows} - 1)(\text{number of columns} - 1)$$

Homogeneity: H_0: the distribution of the variable is the same in every population.
Independence: H_0: there is no association between the two variables.

From history. Karl Pearson's chi-square test of 1900 was soon applied to two-way tables, but the right degrees of freedom were disputed. In 1922 Ronald A. Fisher (whom you met in Lesson 5.5) showed that a table with r rows and c columns has (r − 1)(c − 1) degrees of freedom, the rule used ever since.

Practice habit. Before you compute, ask how the data were produced: one sample classified two ways (independence) or separate samples or treatment groups (homogeneity). Then write H_0 and H_a in words that name the variables and the populations.

WORKED EXAMPLES

Example 1. A gardening club buys bean seeds from three suppliers. Members plant a random sample of 100 seeds from each supplier and record whether each seed germinates; the results are in the table. Is there convincing evidence at the $\alpha = 0.05$ level that the germination rate differs among the three suppliers? Give χ^2 to 2 decimal places and the P-value to 4 decimal places.

Bean seeds from three suppliers

Supplier	Germinated	Did not germinate	Total
A	86	14	100
B	80	20	100
C	71	29	100
Total	237	63	300

1. State: H_0: the proportion of seeds that germinate is the same for all three suppliers. H_a: the proportions are not all the same. Use $\alpha = 0.05$.
2. Plan: chi-square test for homogeneity (three separate random samples, one variable). Random: random samples from each supplier. 10%: 100 seeds is less than 10% of each supplier's stock.
3. Expected counts: each row total is 100, so germinated $= \dfrac{100 \times 237}{300} = 79$ and did not germinate $= \dfrac{100 \times 63}{300} = 21$ in every row. All are at least 5.
4. Do: components A: $\dfrac{7^2}{79} = 0.620$, $\dfrac{(-7)^2}{21} = 2.333$; B: $\dfrac{1^2}{79} = 0.013$, $\dfrac{(-1)^2}{21} = 0.048$; C: $\dfrac{(-8)^2}{79} = 0.810$, $\dfrac{8^2}{21} = 3.048$.
5. $\chi^2 = 0.620 + 2.333 + 0.013 + 0.048 + 0.810 + 3.048 = 6.87$, with $df = (3-1)(2-1) = 2$.
6. P-value $= P(\chi^2 \geq 6.87) = 0.0322$ (table: between 5.991 and 7.378, so $0.025 < P < 0.05$).
7. Conclude: because $0.0322 < 0.05$, we reject H_0. There is convincing evidence that the germination rates of the three suppliers' bean seeds are not all the same.
8. Follow-up: the largest components are in the did-not-germinate column. Supplier C had 29 failures where 21 were expected (component 3.05), and supplier A had only 14 (component 2.33). Supplier A's sample germinated best (86%) and supplier C's worst (71%).

Answer: $\chi^2 = 6.87$, $df = 2$, P-value $= 0.0322$. Reject H_0: there is convincing evidence that germination rates differ among the suppliers; C's seeds failed more often than expected and A's less often.

Example 2. An orchard's farm market asked a random sample of 200 of its customers for their age group and their favorite kind of pie; the results are in the table. Do these data give convincing evidence at the $\alpha = 0.01$ level of an association between age group and favorite pie among the market's customers? Give χ^2 to 2 decimal places and the P-value to 4 decimal places.

Farm market customers: age group and favorite pie

Age group	Apple	Cherry	Pecan	Total
Under 30	14	25	11	50
30 to 54	34	22	24	80
55 and over	32	13	25	70
Total	80	60	60	200

1. State: H_0: there is no association between age group and favorite pie among the market's customers. H_a: there is an association. Use $\alpha = 0.01$.
2. Plan: chi-square test for independence (one random sample, two variables). Random: a random sample of customers. 10%: 200 is less than 10% of all the market's customers.
3. Expected counts, for example under 30 and apple: $\dfrac{50 \times 80}{200} = 20$. In full: under 30: 20, 15, 15; 30 to 54: 32, 24, 24; 55 and over: 28, 21, 21. All are at least 5.
4. Do: components under 30: $\dfrac{(-6)^2}{20} = 1.800$, $\dfrac{10^2}{15} = 6.667$, $\dfrac{(-4)^2}{15} = 1.067$; 30 to 54: $\dfrac{2^2}{32} = 0.125$, $\dfrac{(-2)^2}{24} = 0.167$, 0; 55 and over: $\dfrac{4^2}{28} = 0.571$, $\dfrac{(-8)^2}{21} = 3.048$, $\dfrac{4^2}{21} = 0.762$.
5. $\chi^2 = 14.21$, with $df = (3-1)(3-1) = 4$.
6. P-value $= P(\chi^2 \geq 14.21) = 0.0067$ (table: between 13.277 and 14.860, so $0.005 < P < 0.01$).
7. Conclude: because $0.0067 < 0.01$, we reject H_0. There is convincing evidence of an association between age group and favorite pie among the market's customers.
8. Follow-up: the largest component (6.67) is customers under 30 choosing cherry: 25 did so where 15 were expected (50% of that age group versus 30% of all customers). Customers 55 and over chose cherry less often than expected (13 versus 21). This is an observational study, so it shows an association, not that age causes pie preferences.

Answer: $\chi^2 = 14.21$, $df = 4$, P-value $= 0.0067$. Reject H_0: there is convincing evidence of an association between age group and favorite pie; younger customers favor cherry more than expected.

PRACTICE

1. EASY Which chi-square test fits each study: homogeneity or independence? Explain briefly. (a) A random sample of 250 members of a county 4-H program is asked each member's age group and main project area. (b) Random samples of 100 eggs from each of three farms are graded small, medium, large or extra large.

2. EASY In a two-way table of randomly chosen days at a weather station, the row Windy morning has total 60, the column Afternoon rain has total 45, and the table total is 150. Find the expected count for windy mornings followed by afternoon rain.

3. EASY Find the degrees of freedom for a chi-square test on (a) a table with 3 rows and 4 columns, (b) a table with 2 rows and 5 columns.

4. EASY A choir director randomly assigns 90 singers to three warm-up routines, 30 singers each, and after rehearsal rates each singer's voice as *strong*, *fair* or *weak*. State the hypotheses for the chi-square test, and name the test.

5. MEDIUM The table classifies a random sample of 40 students at a large homeschool co-op by whether they play an instrument and whether they sing in the choir. Find the four expected counts (in the order: plays and sings, plays and does not sing, does not play and sings, does not play and does not sing) and say whether the Large Counts condition is met.

Co-op students

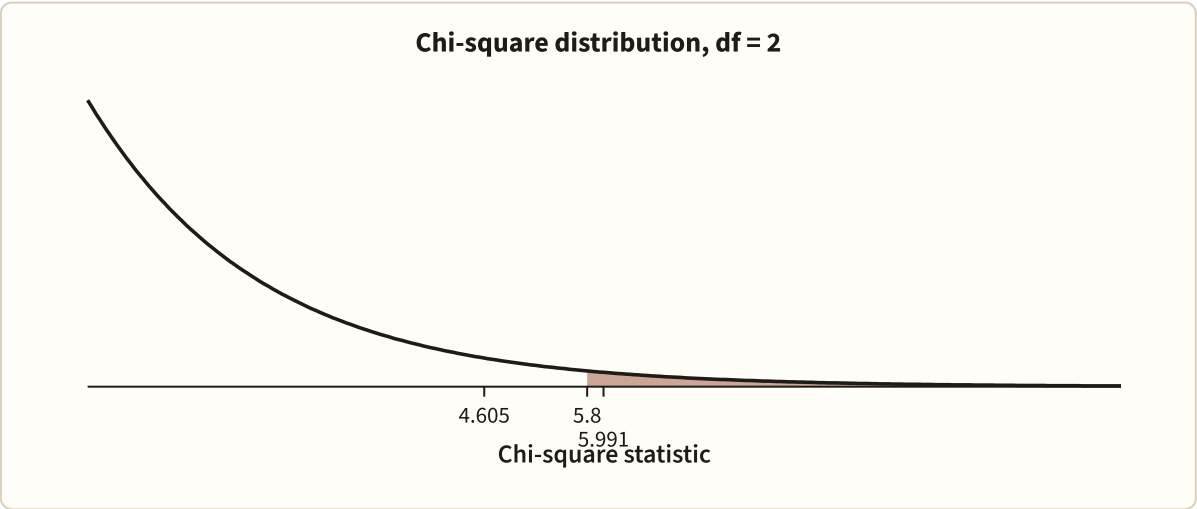
	Sings in choir	Does not sing in choir	Total
Plays an instrument	9	6	15
Does not play	7	18	25
Total	16	24	40

6. MEDIUM A grower plants 200 randomly chosen seeds of each of two brands of tomato and records which plants develop leaf spot (table). Test at the $\alpha = 0.05$ level whether the proportion of plants with leaf spot differs between the brands. Give χ^2 to 2 decimal places and the P-value to 4 decimal places.

Tomato plants with leaf spot

Seed brand	Leaf spot	No leaf spot	Total
Brand A	48	152	200
Brand B	32	168	200
Total	80	320	400

7. MEDIUM A chi-square test on a table with 2 rows and 3 columns gives $\chi^2 = 5.80$. (a) How many degrees of freedom does it have? (b) Find the P-value (4 decimal places); it is the shaded area in the graph. (c) What is the decision at the $\alpha = 0.05$ level?



8. MEDIUM A random sample of 120 hikers in a state park includes 40 hikers on Trail A; 30 of the 120 hikers carry trekking poles. (a) Find the expected number of Trail A hikers with poles for a test of independence between trail and carrying poles. (b) Explain in context why this is the right expected count.

9. MEDIUM After the farm-market pie study in worked example 2, a student writes: 'The chi-square test proves that getting older causes people to like cherry pie less.' Explain two things wrong with this sentence.

10. HARD A weather watcher picks a random sample of 240 summer days from 30 years of records at a station and classifies each day by the morning sky and the afternoon rainfall (table). Test at the $\alpha = 0.01$ level whether there is an association between morning sky and afternoon rain. Give χ^2 to 2 decimal places and the P-value to 4 decimal places.

240 randomly chosen summer days

Morning sky	No rain	Light rain	Heavy rain	Total
Clear	100	30	10	140
Cloudy	50	30	20	100
Total	150	60	30	240

11. **HARD** A gardener randomly assigns 180 pepper seeds to three starting methods, 60 seeds each, and records whether each produces a strong seedling, a weak seedling or does not sprout (table). Test at the $\alpha = 0.05$ level whether the distribution of outcomes differs among the methods. Give χ^2 to 2 decimal places and the P-value to 4 decimal places.

Pepper seeds: three starting methods

Method	Strong seedling	Weak seedling	Did not sprout	Total
Peat pots	38	14	8	60
Soil blocks	30	19	11	60
Paper towel	22	21	17	60
Total	90	54	36	180

12. **HARD** A random sample of 100 hikers on two trails was classified by trail and hiking time. The table gives all the totals and two cells. (a) Fill in the four missing cells, in the order North over 2 hours, South under 1 hour, South 1 to 2 hours, South over 2 hours. (b) Explain how this shows why the table has 2 degrees of freedom.

Hiking times on two trails

Trail	Under 1 hour	1 to 2 hours	Over 2 hours	Total
North	15	18	?	40
South	?	?	?	60
Total	30	50	20	100

13. **CHALLENGE** Random samples of 200 apples from each of two orchards are checked for blemishes (table). (a) Find χ^2 for the 2×2 table. (b) Find the pooled two-proportion z statistic for H_0: $p_1 = p_2$ (Hillside minus Valley; Lesson 12.1). (c) Check that $z^2 = \chi^2$ and give the two-sided P-value (4 decimal places).

Blemished apples from two orchards

Orchard	Blemished	Not blemished	Total
Hillside	50	150	200
Valley	30	170	200
Total	80	320	400

COMMON MISTAKES

- Including the totals as cells. Add components only over the inside cells of the table, never the row, column or grand totals.
- Using df = (number of cells) − 1. For a two-way table, df = (r − 1)(c − 1); a 3 × 3 table has df = 4, not 8.
- Mixing up the two tests. Decide by how the data were produced (one sample classified two ways, or separate samples or treatment groups), and word H_0 to match: no association, or the same distribution in every group.
- Claiming cause from an observational study, or saying the test proves something. Only a randomized experiment supports cause and effect, and a test gives evidence, not proof.

ANSWERS TO PRACTICE

1. (a) Independence: one random sample with two variables (age group, project area) recorded for each member. (b) Homogeneity: three separate samples, one variable (egg grade), comparing its distribution across the three farms.
2. $\frac{60 \times 45}{150} = \frac{2700}{150} = 18$ days.
3. (a) $(3-1)(4-1) = 6$. (b) $(2-1)(5-1) = 4$.
4. H_0: the distribution of voice ratings (strong, fair, weak) is the same for all three warm-up routines. H_a: the distribution of ratings is not the same for all three routines. This is a test for homogeneity, because the three groups were formed by random assignment and one variable (the rating) is compared across them.
5. $\frac{15 \times 16}{40} = 6$, $\frac{15 \times 24}{40} = 9$, $\frac{25 \times 16}{40} = 10$, $\frac{25 \times 24}{40} = 15$. All are at least 5, so yes, the condition is met.
6. State: H_0: the proportion of plants with leaf spot is the same for both brands; H_a: the proportions differ; $\alpha = 0.05$. Plan: test for homogeneity; random samples of each brand's seeds; 200 is less than 10% of each brand's seeds; expected counts $\frac{200 \times 80}{400} = 40$ and $\frac{200 \times 320}{400} = 160$ in each row, all at least 5. Do: $\chi^2 = \frac{8^2}{40} + \frac{(-8)^2}{160} + \frac{(-8)^2}{40} + \frac{8^2}{160} = 1.6 + 0.4 + 1.6 + 0.4 = 4.00$, $df = 1$, P-value $= 0.0455$. Conclude: $0.0455 < 0.05$, so reject H_0. There is convincing evidence that the proportion of plants with leaf spot differs between the brands (24% for Brand A versus 16% for Brand B).
7. (a) $df = (2-1)(3-1) = 2$. (b) $P(\chi^2 \geq 5.80) = 0.0550$ (table: between 4.605 and 5.991, so $0.05 < P < 0.10$). (c) $0.0550 > 0.05$, so fail to reject H_0.
8. (a) $\frac{40 \times 30}{120} = 10$. (b) Overall, $\frac{30}{120} = 25\%$ of the hikers carry poles. If carrying poles has no association with the trail, Trail A hikers should carry poles at the same 25% rate, and 25% of 40 hikers is 10.

9. First, a significance test never proves anything; it gives convincing evidence of an association, with some chance of error. Second, the data come from an observational study, not an experiment, so they cannot show cause and effect; lurking variables (such as which pies were common when each group was young) could explain the association. A better sentence: there is convincing evidence of an association between age group and favorite pie among the market's customers.

10. State: H_0: no association between morning sky and afternoon rain on summer days at this station; H_a: there is an association; $\alpha = 0.01$. Plan: test for independence; random sample of days; 240 is less than 10% of the roughly 2,760 summer days in 30 years; expected counts clear: 87.5, 35, 17.5; cloudy: 62.5, 25, 12.5, all at least 5. Do: $\chi^2 = 1.786 + 0.714 + 3.214 + 2.500 + 1.000 + 4.500 = 13.71$, $df = 2$, P-value $= 0.0011$ (table: between 10.597 and 13.816). Conclude: $0.0011 < 0.01$, so reject H_0. There is convincing evidence of an association between morning sky and afternoon rain. The largest component (4.50): cloudy mornings were followed by heavy rain on 20 days where only 12.5 were expected.

11. State: H_0: the distribution of outcomes is the same for all three methods; H_a: the distributions are not all the same; $\alpha = 0.05$. Plan: test for homogeneity; random assignment of seeds to methods; expected counts in every row $\frac{60 \times 90}{180} = 30$, $\frac{60 \times 54}{180} = 18$, $\frac{60 \times 36}{180} = 12$, all at least 5. Do: $\chi^2 = 2.133 + 0.889 + 1.333 + 0 + 0.056 + 0.083 + 2.133 + 0.500 + 2.083 = 9.21$, $df = 4$, P-value $= 0.0560$ (table: between 7.779 and 9.488, so $0.05 < P < 0.10$). Conclude: $0.0560 > 0.05$, so fail to reject H_0. There is not convincing evidence that the distribution of outcomes differs among the three methods. (The result is close; a larger experiment might detect a real difference.)

12. (a) North over 2 hours: $40 - 15 - 18 = 7$; South under 1 hour: $30 - 15 = 15$; South 1 to 2 hours: $50 - 18 = 32$; South over 2 hours: $20 - 7 = 13$ (check: $15 + 32 + 13 = 60$). (b) Once the totals are fixed, only 2 cells could be chosen freely; every other cell was forced. That number of free cells, $(2 - 1)(3 - 1) = 2$, is the degrees of freedom.

13. (a) Expected counts 40 and 160 in each row; $\chi^2 = \frac{10^2}{40} + \frac{(-10)^2}{160} + \frac{(-10)^2}{40} + \frac{10^2}{160} = 2.5 + 0.625 + 2.5 + 0.625 = 6.25$. (b) $\hat{p}_1 = 0.25$, $\hat{p}_2 = 0.15$, $\hat{p}_c = \frac{80}{400} = 0.20$; $z = \frac{0.25 - 0.15}{\sqrt{0.20 \times 0.80\left(\frac{1}{200} + \frac{1}{200}\right)}} = \frac{0.10}{0.04} = 2.50$. (c) $z^2 = 2.50^2 = 6.25 = \chi^2$. P-value $= 2(0.00621) = 0.0124$, the same as $P(\chi^2 \geq 6.25)$ with $df = 1$.

LESSON 13.3 · 45 MINUTES

Thinking Wisely about Statistics

OBJECTIVES

- Evaluate a statistical claim by asking how the data were collected, what was measured, how large the sample was, whether chance could explain the result, whether it shows cause, and whether it matters in practice.
- Explain why cherry-picking and running many tests mislead, and compute the chance of at least one false alarm among several tests.
- Choose and carry out an appropriate procedure from any unit of the course, and state the conclusion honestly in context.

VIDEO

WATCH

Survivorship Bias

Eddie Woo · 6:09

When to watch: Before Example 1: survivorship bias.

https://www.youtube.com/watch?v=P9WFpVsRtQg

BACKUP

Simpson's Paradox

ProfRobBob · 5:26

When to watch: After Example 1: Simpson's paradox.

https://www.youtube.com/watch?v=fi9iCM-6xxk

KEY TERMS

cherry-picking	Reporting only the data or results that support the conclusion you want, and leaving out the rest.
practical importance	Whether an effect is large enough to matter in real life. A result can be statistically significant (too large to be explained easily by chance alone) and still be too small to matter.
multiple testing	Running many significance tests on the same data or question; the more tests you run, the more likely at least one will be significant by chance alone.

Simpson's paradox A reversal in which a comparison that holds within every group turns the other way when the groups are combined, because a lurking variable (the variable that defines the groups) is combined with very unequal group sizes.

CORE IDEAS

You now have tools to describe data, to model chance and to draw conclusions with known error rates. The last skill is judgment: knowing when a statistical claim deserves belief. Claims meet you everywhere, in seed catalogs, sports broadcasts, advertisements and science-fair posters. Before believing one, run through a short set of questions. Each points back to something you learned this year.

Questions to ask about a statistical claim

Question	Why it matters	Where you learned it
How were the data collected?	Bias cannot be fixed by a bigger sample or a formula	Unit 5
What exactly was measured?	Vague or slanted measures mislead	Lessons 1.1, 5.3
Compared to what?	An effect needs a fair comparison group	Lessons 5.4, 5.5
Could chance explain it?	Small samples vary a lot; ask for an interval or P-value	Units 9 to 13
Correlation or cause?	Only randomized experiments show cause	Lessons 4.5, 5.6
Is it big enough to matter?	Significant is not the same as important	Lesson 11.5

How were the data collected? This comes first because no formula can repair bad data. A voluntary response or convenience sample is biased, and a larger biased sample only gives a more precise wrong answer (Lessons 5.1 and 5.3). The margin of error from Unit 10 covers random sampling error only, never bias. **Compared to what?** A claim such as "plants grew 30% taller" means nothing without a fair comparison group, and a fair comparison means random assignment (Lesson 5.5).

Chance or a real effect? Small samples vary a great deal, so a striking result from 8 tomato plants may be luck; ask for a P-value or, better, a confidence interval. **Significant or important?** With a huge sample, even a tiny difference can be statistically significant (Lesson 11.5); the interval shows how big the effect really is. **Correlation or cause?** An association in observational data may be due to a lurking variable (Lesson 4.5). Lurking variables can even reverse a comparison when groups are combined, a surprise called **Simpson's paradox** (practice problem 13).

Honesty with data is a duty, not only a technique. Statistics is a tool for telling the truth about God's world, and it works only when the people using it are truthful; the command not to bear false witness applies to graphs and P-values as much as to words. Honest work means reporting what you found even when it is disappointing, showing how the data were collected, admitting the limits of your study, and never **cherry-picking**.

Cherry-picking is easy to fall into by running many tests. If you test 20 products that do nothing, each at $\alpha = 0.05$, you expect $20 \times 0.05 = 1$ to look significant by chance alone, and if the tests are independent, $P(\text{at least one}) = 1 - 0.95^{20} = 0.6415$. Reporting only that one 'winner' is a false witness, even if every calculation is correct. Honest reports say how many tests were run.

Your statistics project is the place to practice all of this: ask a clear question, collect data carefully, plot it first, choose a procedure whose conditions you can check, and write a conclusion that says exactly what your data can support and no more.

KEY FORMULAS

$$P(\text{at least one false alarm in } k \text{ independent tests}) = 1 - (1 - \alpha)^k$$

Practice habit. Before you believe any statistical claim, ask two questions: How were the data collected? Compared to what?

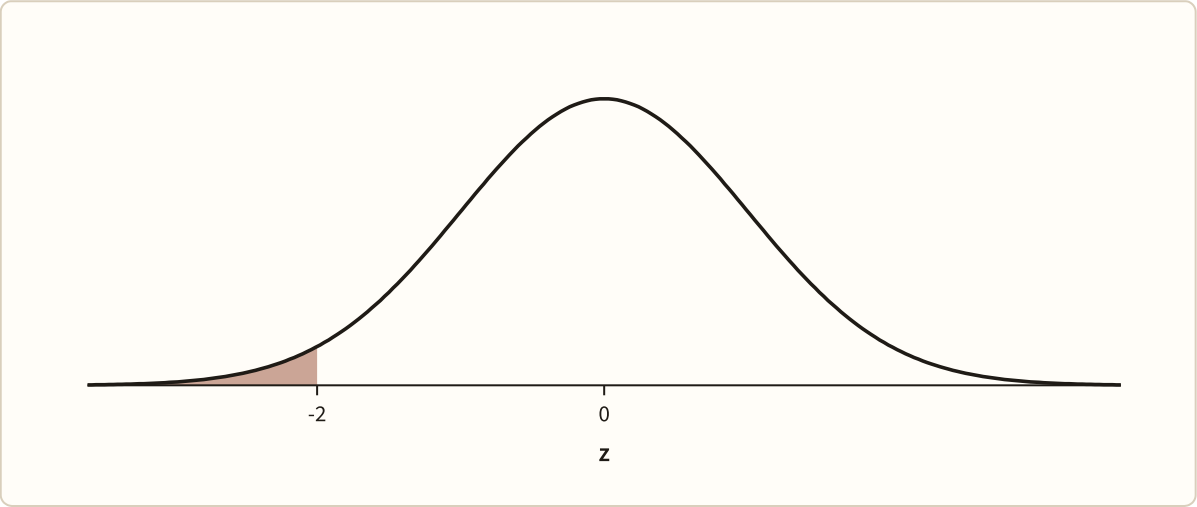
WORKED EXAMPLES

Example 1. A garden-supply catalog prints: 'Tomatoes grown with Super-Grow plant food average 30% heavier! Based on 85 letters from our customers.' Evaluate this claim using the questions in the table of this lesson, and describe a better study.

1. How were the data collected? The company used letters from customers who chose to write: a voluntary response sample. Pleased customers write; disappointed ones mostly do not, so the sample is biased toward success.
2. What was measured? '30% heavier' than what? Letter writers were probably comparing with their memory of other years, not with weighed tomatoes.
3. Compared to what? There is no control group. Weather, soil and tomato variety differ from garden to garden and from year to year.
4. Correlation or cause? Gardeners who buy plant food may also water and weed more carefully (a lurking variable), so even a real difference would not show that Super-Grow caused it.
5. Could chance explain it? No measurements, no measure of variation, and no interval or P-value are given.
6. A better study: in one garden, randomly assign tomato plants (or pairs of neighboring plants, one of each pair chosen at random) to receive Super-Grow or no plant food, weigh all the tomatoes, and compare the groups with a two-sample or paired t procedure.

Answer: The claim is not trustworthy: it rests on a biased voluntary response sample with no control group, no clear measurement and no allowance for chance. Only a randomized comparative experiment could show that Super-Grow makes tomatoes heavier.

Example 2. A seed company advertises that 90% of its carrot seeds germinate. A garden club plants a random sample of 400 of the seeds, and 348 germinate. Is there convincing evidence at the $\alpha = 0.05$ level that fewer than 90% of the company's carrot seeds germinate? Then comment on whether the difference matters in practice. Give z to 2 decimal places and the P-value to 4 decimal places.



1. Choose the procedure: one categorical variable (germinates or not), one sample, a claim about a proportion: a one-sample z test for p.
2. State: H_0: $p = 0.90$; H_a: $p < 0.90$, where p is the proportion of all the company's carrot seeds that germinate. Use $\alpha = 0.05$.
3. Plan: Random: a random sample of seeds. 10%: 400 is far less than 10% of all the company's seeds. Large Counts: $np_0 = 400 \times 0.90 = 360$ and $n(1 - p_0) = 400 \times 0.10 = 40$, both at least 10.
4. Do: $\hat{p} = \dfrac{348}{400} = 0.87$; standard deviation $\sqrt{\dfrac{0.90 \times 0.10}{400}} = \sqrt{0.000225} = 0.015$.
5. $z = \dfrac{0.87 - 0.90}{0.015} = -2.00$; P-value $= P(Z \le -2.00) = 0.0228$.
6. Conclude: because $0.0228 < 0.05$, we reject H_0. There is convincing evidence that fewer than 90% of the company's carrot seeds germinate.
7. Practical importance: the sample rate, 87%, is 3 percentage points below the advertised 90%, about 3 more failed seeds per 100. The advertisement overstates the rate, which is worth reporting honestly, but a gardener can make up the difference by sowing a little more thickly.

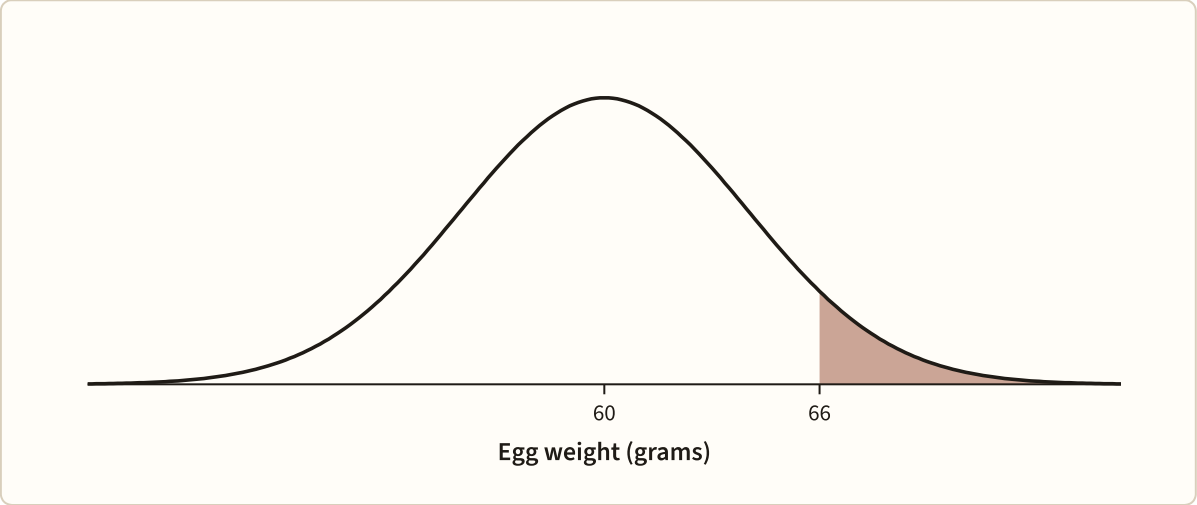
Answer: $z = -2.00$, P-value $= 0.0228$. Reject H_0: there is convincing evidence that fewer than 90% of the seeds germinate. The shortfall (87% in the sample) is real but small in practice.

PRACTICE

1. EASY A rancher records two variables for each calf born this spring: (a) its breed and (b) its birth weight in pounds. Classify each variable as categorical or quantitative.

2. EASY The snowfall in five winter storms was 4, 7, 7, 9 and 13 inches. Find the mean and the median.

3. EASY Egg weights from a flock are approximately normal with mean 60 grams and standard deviation 4 grams. (a) Find the z-score of a 66-gram egg. (b) What proportion of eggs weigh more than 66 grams (4 decimal places)?



4. EASY A radio station invites listeners to call in and say whether they would attend a second night of the town's summer concert series. Of 212 callers, 78% say yes. Explain why this percent is probably not a good estimate for the whole town.

5. MEDIUM A gardener's least-squares line for tomato yield $\hat{y}$ (pounds per plant) on weekly water x (gallons) is $\hat{y} = 2.4 + 0.8x$. (a) Predict the yield of a plant given 10 gallons per week. (b) A plant given 10 gallons per week yielded 11.0 pounds. Find its residual.

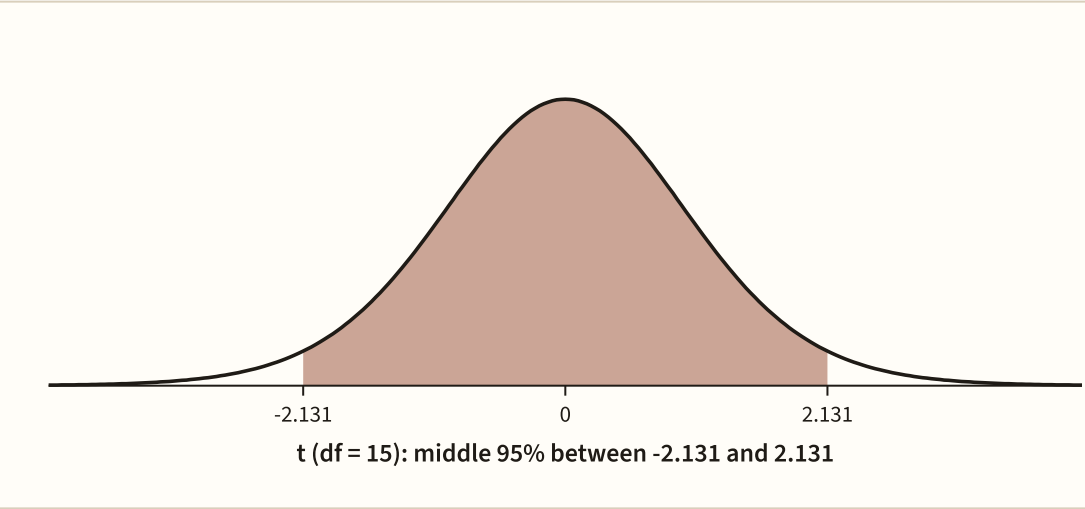
6. MEDIUM Each pepper seed in a packet germinates with probability 0.85, independently of the others. Ruth plants 3 seeds in a pot. What is the probability that at least one germinates (4 decimal places)?

7. MEDIUM Grace makes 80% of her free throws, and her shots are independent. In 5 shots, what is the probability that she makes exactly 4 (4 decimal places)? The graph shows the whole distribution.

8. MEDIUM In a random sample of 400 households in a county, 120 grow a vegetable garden. Construct a 95% confidence interval for the proportion of all households in the county that grow a vegetable garden (endpoints to 3 decimal places), and interpret it.

9. MEDIUM A golf ball maker tests a new ball on 50,000 drives by a robot and reports that it travels farther than the old ball, with P-value = 0.001. The mean difference was 1.2 yards. Should a golfer rush to buy the new ball? Explain.

10. HARD A sawmill cuts boards meant to be 96 inches long. A random sample of 16 boards has mean length $\bar{x} = 96.20$ inches and standard deviation $s = 0.40$ inch, and a dotplot of the lengths is roughly symmetric with no outliers. (a) Construct a 95% confidence interval for the mean length of all boards (endpoints to 2 decimal places). (b) Do these data give convincing evidence at the $\alpha = 0.05$ level that the mean length differs from 96 inches?



11. HARD A vegetable grower tests 20 different 'miracle' garden products, each against an untreated control at the $\alpha = 0.05$ level. Suppose none of the products really works and the tests are independent. (a) How many tests would you expect to come out significant? (b) What is the probability that at least one does (4 decimal places)? (c) Why would it be dishonest to advertise only the products that came out significant?

12. HARD Name the inference procedure that fits each study, and give a reason. (a) Is the proportion of cracked eggs in a random sample of 300 from a farm's shipments below the 5% the farm claims? (b) Twelve fields are each split in half, with one fertilizer on each half chosen at random; does the mean yield differ? (c) Does the distribution of favorite season differ among random samples of children, adults and seniors? (d) Do lambs raised on two pastures (independent random samples) differ in mean weight?

13. CHALLENGE Two players' hits and at-bats for the two halves of a season are in the table. Henry has the higher batting average in each half (0.400 versus 0.350, and 0.250 versus 0.200). (a) Find Henry's batting average for the whole season (3 decimal places). (b) Find Samuel's (3 decimal places). (c) Explain how Henry can be better in each half but worse for the season.

Hits and at-bats

Player	First half (hits / at-bats)	Second half (hits / at-bats)
Henry	4 / 10	25 / 100
Samuel	35 / 100	2 / 10

COMMON MISTAKES

- Treating 'statistically significant' as 'important'. Always look at the size of the effect, ideally with a confidence interval, and ask whether it matters in context.
- Trusting a large sample that was badly collected. Size reduces random variation but cannot remove bias from voluntary response, undercoverage or slanted questions.
- Concluding cause from observational data. Ask whether there was random assignment; if not, look for lurking variables.
- Reporting only the results you like. Say how many tests you ran and report the disappointing results along with the exciting ones.

ANSWERS TO PRACTICE

1. (a) Categorical. (b) Quantitative.
2. Mean $= \dfrac{4+7+7+9+13}{5} = \dfrac{40}{5} = 8.0$ inches; median $= 7$ inches.
3. (a) $z = \dfrac{66-60}{4} = 1.50$. (b) $1 - 0.9332 = 0.0668$.

4. This is a voluntary response sample: only people who chose to call are counted, and people who care most about the concerts (likely fans) are the ones who call. The 78% probably overstates the town's interest; a random sample of residents would be needed.
5. (a) $\hat{y} = 2.4 + 0.8(10) = 10.4$ pounds. (b) Residual $= 11.0 - 10.4 = 0.6$ pound (the plant did 0.6 pound better than predicted).
6. $P(\text{at least one}) = 1 - P(\text{none}) = 1 - 0.15^3 = 1 - 0.003375 = 0.9966$.
7. $P(X = 4) = \binom{5}{4}(0.8)^4(0.2)^1 = 5 \times 0.4096 \times 0.2 = 0.4096$.
8. $\hat{p} = 0.30$; $0.30 \pm 1.960\sqrt{\dfrac{0.30 \times 0.70}{400}} = 0.30 \pm 1.960(0.02291) = 0.30 \pm 0.045$, giving $(0.255, 0.345)$. (Conditions: random sample; 400 is less than 10% of households; 120 and 280 are both at least 10.) We are 95% confident that between 25.5% and 34.5% of all households in the county grow a vegetable garden.
9. Not necessarily. The tiny P-value shows that a difference this large would be very unlikely if the new ball were really no better than the old one, but with 50,000 drives even a very small difference is statistically significant. An extra 1.2 yards on drives of over 200 yards has almost no practical importance for an ordinary golfer.
10. (a) $df = 15$, $t^* = 2.131$; $96.20 \pm 2.131 \times \dfrac{0.40}{\sqrt{16}} = 96.20 \pm 2.131(0.10) = 96.20 \pm 0.2131$, giving $(95.99, 96.41)$ inches. (b) No: 96 lies inside the 95% interval, so a two-sided test at $\alpha = 0.05$ would fail to reject H_0: $\mu = 96$. There is not convincing evidence that the mean length differs from 96 inches.
11. (a) $20 \times 0.05 = 1$. (b) $1 - 0.95^{20} = 1 - 0.3585 = 0.6415$. (c) With 20 tests, at least one false alarm is more likely than not, so a single significant result is just what chance would produce. Reporting it without saying that 20 tests were run hides the truth from buyers; an honest report gives all 20 results.
12. (a) One-sample z test for a proportion: one sample, one yes-or-no variable, a claim about p. (b) Paired t test: each field gives two linked measurements, so analyze the 12 differences. (c) Chi-square test for homogeneity: one categorical variable with several categories, compared across three separate samples. (d) Two-sample t test (or interval): two independent samples and a quantitative variable.
13. (a) Henry: $\dfrac{4 + 25}{10 + 100} = \dfrac{29}{110} = 0.264$. (b) Samuel: $\dfrac{35 + 2}{100 + 10} = \dfrac{37}{110} = 0.336$. (c) This is Simpson's paradox. Most of Henry's at-bats came in the second half, when both players hit poorly, while most of Samuel's came in the first half, when both hit well. The half of the season is a lurking variable; combining the halves hides it and reverses the comparison.

Unit 13 test

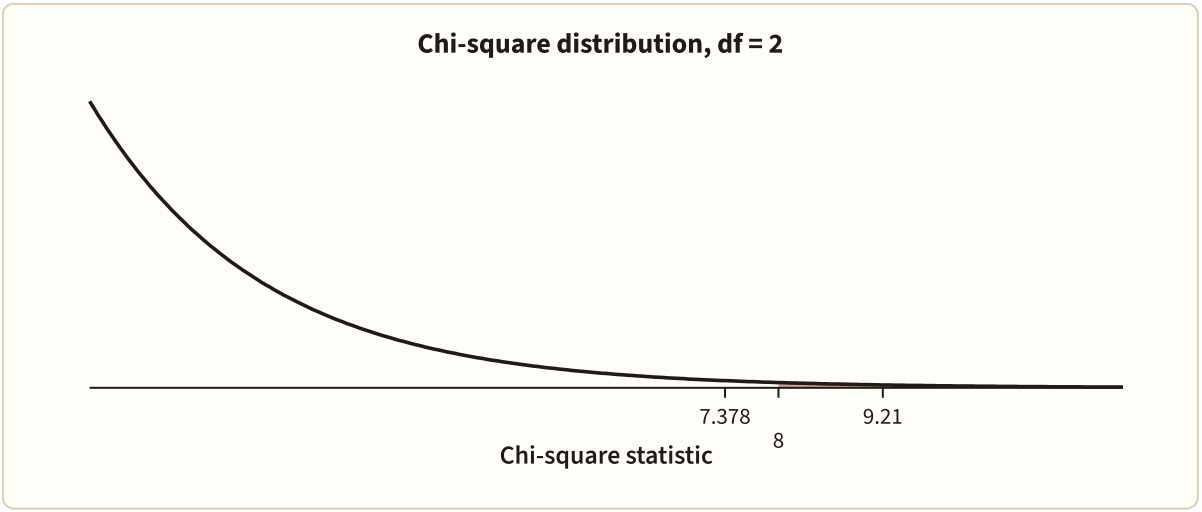
You have 55 minutes. You may use a scientific or graphing calculator and the chi-square table at the back of the course. Show your work: for every significance test, write the hypotheses in symbols or words, check the conditions, show the expected counts and the chi-square calculation, and state a conclusion in context. Unless a problem says otherwise, round χ^2 to 2 decimal places and P-values to 4 decimal places; if you use the table, give bounds for the P-value instead. Total: 100 points.

1. **5 PTS** A nursery says its mixed tulip bulbs are 50% red, 30% yellow and 20% white. A random sample of 150 bulbs is planted. Find the expected count of each color if the nursery's claim is true.
2. **5 PTS** Find the degrees of freedom for (a) a goodness-of-fit test of whether calves on a large ranch are born equally often on each of the 7 days of the week, and (b) a test for independence on a two-way table with 4 rows and 3 columns.
3. **6 PTS** A church's potluck committee believes the dishes brought to its potlucks are 25% salads, 50% main dishes and 25% desserts. A random sample of 100 dishes from the year's potlucks gives the counts in the table. Find the chi-square statistic (2 decimal places).

Random sample of 100 potluck dishes

Dish	Salad	Main dish	Dessert	Total
Observed	30	45	25	100

4. **6 PTS** A chi-square test with $df = 2$ gives $\chi^2 = 8.00$. (a) Find the P-value (4 decimal places); it is the shaded area in the graph. (b) What is the decision at the $\alpha = 0.05$ level? (c) At the $\alpha = 0.01$ level?



5. **6 PTS** Name the chi-square test (homogeneity or independence) that fits each study, and explain your choice. (a) An astronomy club takes a random sample of 150 members of a national amateur astronomy society and records each member's type of telescope and favorite kind of object to observe. (b) A farmer randomly assigns 90 piglets to three feeds, 30 piglets each, and records whether each piglet's weight gain is low, medium or high.

6. **6 PTS** In a two-way table of 300 randomly chosen students at a large school, classified by grade and favorite sport, the row Grade 11 has total 120 and the column Soccer has total 75. (a) Find the expected count for 11th graders whose favorite sport is soccer. (b) Explain in context why this is the count expected if grade and favorite sport are independent.

7. **12 PTS** An orchard says its apples are 10% small, 20% medium, 40% large, 20% extra large and 10% jumbo. A buyer grades a random sample of 200 apples from the season's harvest (table). Carry out a test at the $\alpha = 0.05$ level of whether the size distribution differs from the orchard's claim.

Random sample of 200 apples by size grade

Size	Small	Medium	Large	Extra large	Jumbo	Total
Claimed proportion	0.1	0.2	0.4	0.2	0.1	1
Observed count	14	36	92	42	16	200

8. **12 PTS** A random sample of 300 visitors to a county fair was asked whether they live on a farm or in town and which of three events they enjoyed most (table). Do these data give convincing evidence at the $\alpha = 0.05$ level of an association between where visitors live and their favorite event? If so, describe the association using the largest components.

Random sample of 300 fair visitors

Home	Livestock show	Tractor pull	Baking contest	Total
On a farm	52	38	30	120
In town	48	62	70	180
Total	100	100	100	300

9. **8 PTS** A random sample of hikers at a state park was classified by age group (under 40, 40 and over) and which of four trails they chose. The computer output is shown. (a) Explain why $df = 3$. (b) What is the decision at the $\alpha = 0.05$ level? (c) At the $\alpha = 0.01$ level? (d) State the conclusion in context at the $\alpha = 0.05$ level.

Computer output: chi-square test

Chi-Sq	DF	P-Value
9.00	3	0.0293

10. **8 PTS** A county extension agent says the fruit trees in local backyard orchards are 40% apple, 30% peach, 20% pear, 6% cherry and 4% plum. Joseph plans to identify a random sample of 60 backyard fruit trees and test the claim. (a) Find the five expected counts. (b) Is the Large Counts condition met? (c) Suggest one way to fix the problem.

11. **10 PTS** A feed store's website says: 'Dairy farmers who switched to our premium feed report 12% more milk per cow! Based on 40 farmers who answered our online survey.' Identify at least three problems with this claim, and describe a better study.

12. **8 PTS** A gardener tests 10 different fertilizers, each against an untreated control at the $\alpha = 0.05$ level. Suppose none of the fertilizers really helps and the tests are independent. (a) How many tests would you expect to come out significant? (b) What is the probability that at least one does (4 decimal places)? (c) What should an honest report of the study include?

13. **8 PTS** (a) Explain why the chi-square statistic divides each squared difference $(O - E)^2$ by the expected count E. (b) Explain why the P-value for a chi-square test is always the area to the right of the observed χ^2.

1. Red $150 \times 0.50 = 75$, yellow $150 \times 0.30 = 45$, white $150 \times 0.20 = 30$.
2. (a) $7 - 1 = 6$. (b) $(4 - 1)(3 - 1) = 6$.
3. Expected: 25, 50, 25. $\chi^2 = \dfrac{(30 - 25)^2}{25} + \dfrac{(45 - 50)^2}{50} + \dfrac{(25 - 25)^2}{25} = 1.00 + 0.50 + 0 = 1.50$.
4. (a) $P(\chi^2 \geq 8.00) = 0.0183$ (table: between 7.378 and 9.210, so $0.01 < P < 0.025$). (b) $0.0183 < 0.05$: reject H_0. (c) $0.0183 > 0.01$: fail to reject H_0.
5. (a) Independence: one random sample, with two categorical variables recorded on each member. (b) Homogeneity: three groups formed by random assignment, comparing the distribution of one variable (weight-gain category) across the feeds.
6. (a) $\dfrac{120 \times 75}{300} = 30$. (b) Overall, $\dfrac{75}{300} = 25\%$ of the students prefer soccer. If grade and favorite sport are independent, 11th graders should prefer soccer at the same 25% rate, and 25% of 120 is 30.
7. State: H_0: the proportions are 0.10, 0.20, 0.40, 0.20, 0.10 as claimed; H_a: at least one proportion differs; $\alpha = 0.05$. Plan: goodness-of-fit test; random sample; 200 is less than 10% of the harvest; expected counts 20, 40, 80, 40, 20, all at least 5. Do: $\chi^2 = \dfrac{(-6)^2}{20} + \dfrac{(-4)^2}{40} + \dfrac{12^2}{80} + \dfrac{2^2}{40} + \dfrac{(-4)^2}{20} = 1.80 + 0.40 + 1.80 + 0.10 + 0.80 = 4.90$, $df = 4$, P-value $= 0.2977$ (table: P > 0.10). Conclude: $0.2977 > 0.05$, so fail to reject H_0. There is not convincing evidence that the size distribution of the orchard's apples differs from its claim.
8. State: H_0: no association between home (farm or town) and favorite event among fair visitors; H_a: there is an association; $\alpha = 0.05$. Plan: test for independence; one random sample; 300 is less than 10% of all visitors; expected counts $\dfrac{120 \times 100}{300} = 40$ in each farm cell and $\dfrac{180 \times 100}{300} = 60$ in each town cell, all at least 5. Do: $\chi^2 = 3.600 + 0.100 + 2.500 + 2.400 + 0.067 + 1.667 = 10.33$, $df = 2$, P-value $= 0.0057$ (table: between 9.210 and 10.597). Conclude: $0.0057 < 0.05$, so reject H_0. There is convincing evidence of an association between where visitors live and their favorite event. Largest components: farm visitors chose the livestock show more often than expected (52 versus 40, component 3.60) and the baking contest less often (30 versus 40, component 2.50); town visitors did the reverse.
9. (a) The table has 2 rows and 4 columns, so $df = (2 - 1)(4 - 1) = 3$. (b) $0.0293 < 0.05$: reject H_0. (c) $0.0293 > 0.01$: fail to reject H_0. (d) At $\alpha = 0.05$ there is convincing evidence of an association between age group and choice of trail among hikers at this park.
10. (a) $60 \times 0.40 = 24$, $60 \times 0.30 = 18$, $60 \times 0.20 = 12$, $60 \times 0.06 = 3.6$, $60 \times 0.04 = 2.4$. (b) No: two expected counts are below 5. (c) Combine cherry and plum into one category (expected count 6), or sample at least 125 trees so that the smallest expected count is $125 \times 0.04 = 5$.

11. (1) Voluntary response: only farmers who chose to answer are counted, and satisfied customers are more likely to respond. (2) No comparison group: milk yields change with season, weather, herd age and pasture, so 12% more than last year may have nothing to do with the feed. (3) Self-reported, not measured: the farmers' figures may be estimates. (4) Observational, so no cause can be shown; farmers who buy premium feed may also manage their herds differently (a lurking variable). (5) No measure of chance variation is given. Better: randomly assign cows (or herds, or matched pairs of similar cows) to the premium feed or the usual feed, measure milk yields over the same period, and compare with a two-sample or paired t procedure.
12. (a) $10 \times 0.05 = 0.5$. (b) $1 - 0.95^{10} = 1 - 0.5987 = 0.4013$. (c) It should say that 10 fertilizers were tested and report every result, not only the significant ones, so readers can see that one significant result could easily be a false alarm.
13. (a) A difference only means something compared with what was expected: missing by 10 when 20 were expected is a big surprise, but missing by 10 when 500 were expected is tiny. Dividing by E measures each miss relative to the size of its expected count. (b) Every kind of disagreement between observed and expected counts, in any direction, makes χ^2 larger, while a good fit makes it small. So only large values are evidence against H_0, and the P-value is the chance of a χ^2 at least as large as the one observed.

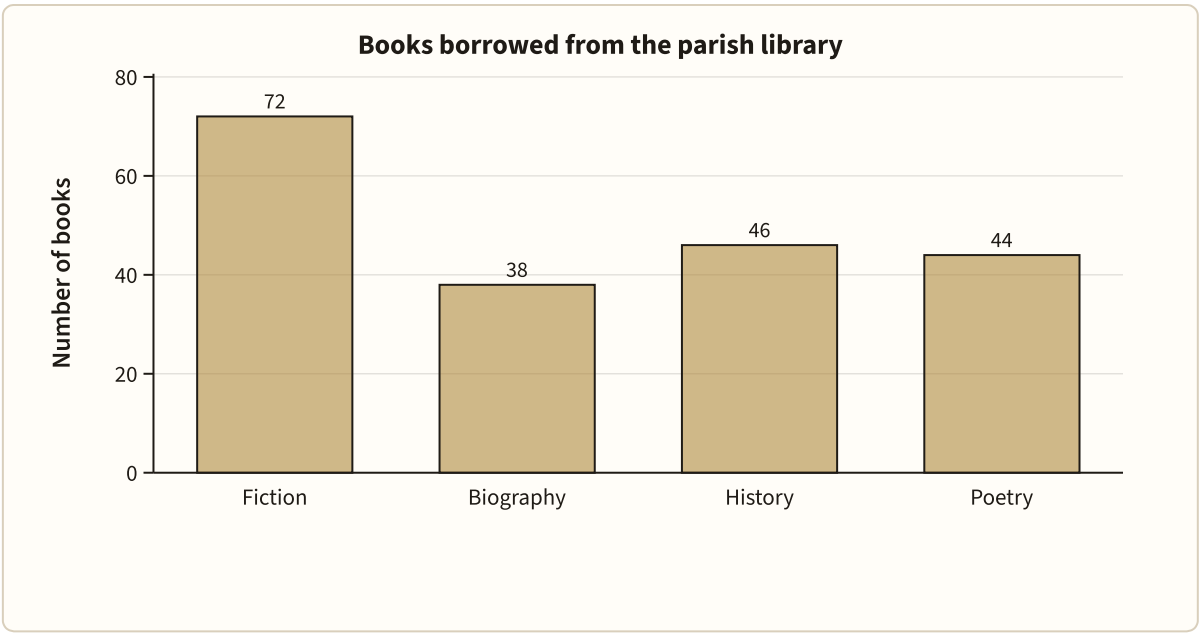
Midterm and final exams

Midterm Exam

Covers Units 1–6. Time: 90 minutes. You have 90 minutes. You may use a scientific calculator and the normal table at the back of the course. Show your work for every problem: write the formula, the numbers you substitute and the result. Interpret answers in a sentence in the context of the problem, with units. Keep full precision until the last step; unless a problem says otherwise, round probabilities to 4 decimal places and z-scores to 2 decimal places. Answers that differ from the key only in the last decimal place because of the table are correct. There are 25 problems worth 4 points each (100 points).

1. **4 PTS** For each hen in her flock, Ruth records the hen's breed, the number of eggs the hen laid in one week, and the average weight of the hen's eggs in grams. (a) Is breed a categorical or a quantitative variable? (b) Is the number of eggs laid in a week discrete or continuous? (c) Is the average egg weight discrete or continuous?

2. **4 PTS** The bar graph shows the types of 200 books borrowed from a parish library in one month. (a) What percent of the books borrowed were history books? (b) Explain why a pie chart would also be a sensible display of these data.

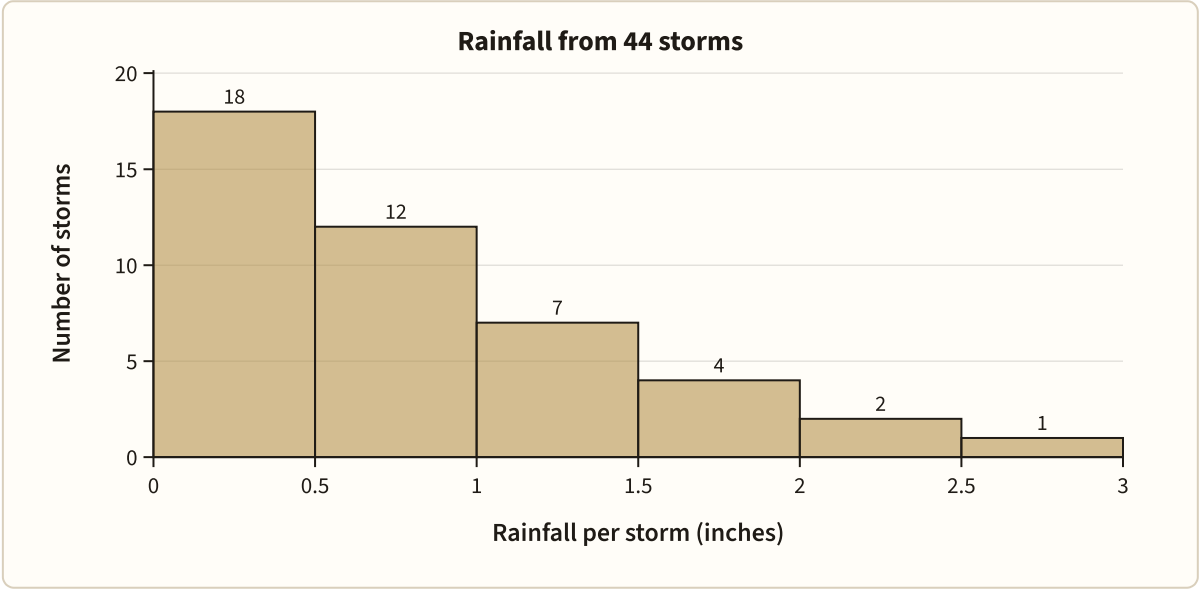


3. **4 PTS** The two-way table shows the age group and ribbon result of 200 exhibitors at a county fair. (a) What percent of junior exhibitors won a blue ribbon? (b) What percent of senior exhibitors won a blue ribbon?

County fair exhibitors

Age group	Blue ribbon	Red ribbon	No ribbon	Total
Junior	18	30	32	80
Senior	36	42	42	120
Total	54	72	74	200

4. **4 PTS** The histogram shows the rainfall, in inches, from each of 44 storms recorded at a farm. (a) Describe the shape of the distribution (symmetric, skewed left or skewed right). (b) How many storms produced at least 1.5 inches of rain?



5. **4 PTS** A farmer sold 120 bushels of apples at \$18 per bushel and 80 bushels at \$23 per bushel. What was the average price per bushel for all 200 bushels? Explain why the answer is not \$20.50.

6. **4 PTS** The stemplot shows the weights, in pounds, of 15 pumpkins from a family garden. Find the five-number summary.

Weights of 15 pumpkins

0	8
1	1 2 4 5 5 7 8
2	0 1 3 4 6
3	1
4	4

Key: 1 | 4 means 14 pounds

7. **4 PTS** Use the pumpkin weights in the stemplot of Problem 6. (a) Find the upper fence of the 1.5 × IQR rule. (b) Is the 44-pound pumpkin an outlier?

Weights of 15 pumpkins

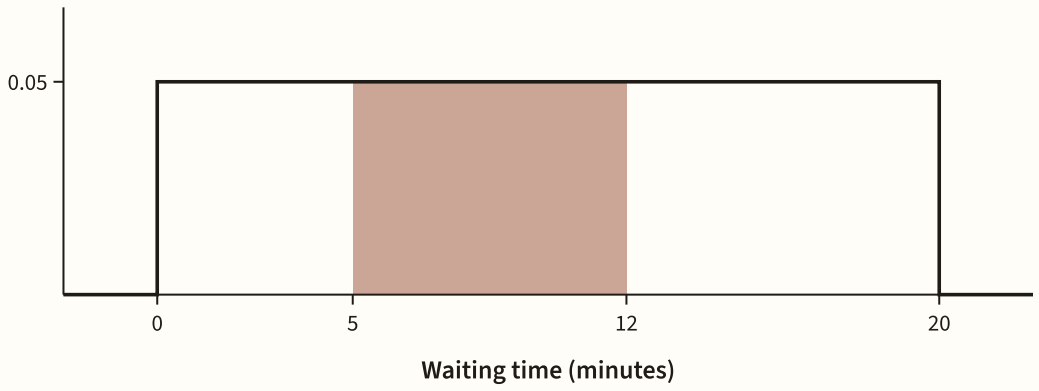
0	8
1	1 2 4 5 5 7 8
2	0 1 3 4 6
3	1
4	4

Key: 1 | 4 means 14 pounds

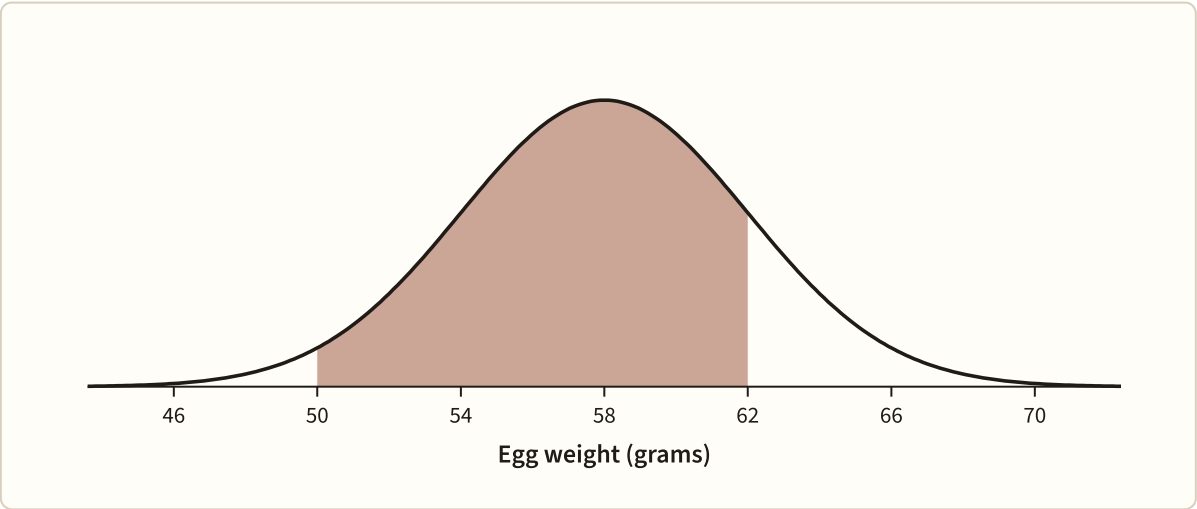
8. **4 PTS** A baker weighs five loaves of rye bread, in ounces: 22, 24, 25, 26, 28. Find the mean and the standard deviation of these weights. Round the standard deviation to two decimal places.

9. **4 PTS** A corn stalk in Field A is 96 inches tall; the corn in Field A has mean height 84 inches and standard deviation 8 inches. A sunflower in the garden is 116 inches tall; the sunflowers have mean height 96 inches and standard deviation 10 inches. (a) Find the z-score of the corn stalk. (b) Find the z-score of the sunflower. Which plant is taller relative to its own kind?

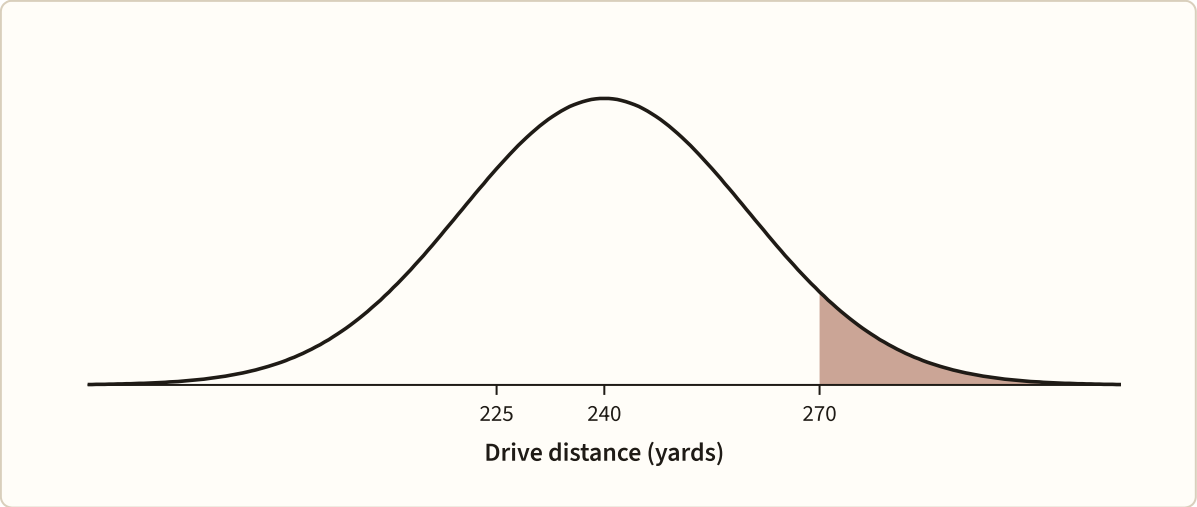
10. **4 PTS** A ferry leaves every 20 minutes. If you arrive at a random time, your waiting time is uniformly distributed between 0 and 20 minutes, as shown. (a) What is the probability that you wait between 5 and 12 minutes? (b) What is the median waiting time?



11. **4 PTS** The weights of eggs from a large flock are approximately normal with mean 58 grams and standard deviation 4 grams. Use the 68–95–99.7 rule to find the percent of eggs that weigh between 50 and 62 grams.

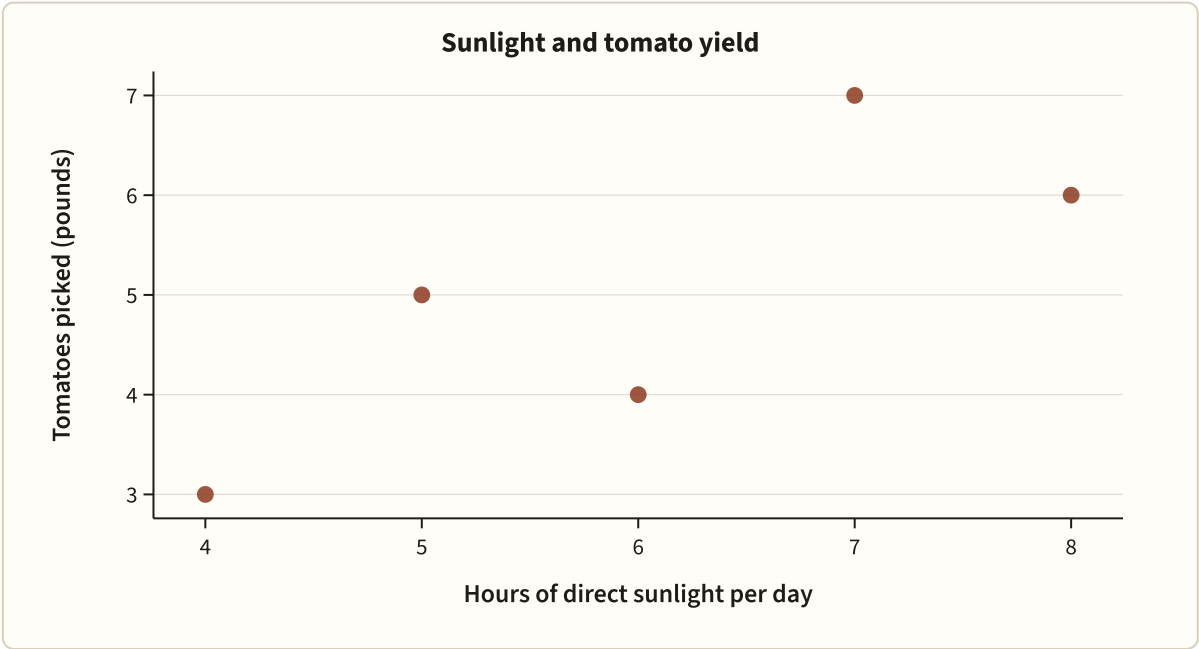


12. **4 PTS** Henry's golf drives are approximately normal with mean 240 yards and standard deviation 20 yards. (a) What proportion of his drives travel more than 270 yards? (b) What proportion travel between 225 and 270 yards?



13. **4 PTS** The weights of watermelons from a farm are approximately normal with mean 18 pounds and standard deviation 2.5 pounds. The heaviest 10% are sold as "premium". What is the least weight a premium watermelon can have? Round to one decimal place.

14. **4 PTS** Grace records the hours of direct sunlight per day and the pounds of tomatoes picked from five plants: (4, 3), (5, 5), (6, 4), (7, 7), (8, 6). Compute the correlation r by hand. Round to two decimal places.



15. **4 PTS** For 30 hay fields in one county, the rainfall during the growing season had mean 20 inches and standard deviation 4 inches, and the hay yield had mean 3.0 tons per acre and standard deviation 0.5 tons per acre. The correlation is r = 0.8. Find the equation of the least-squares regression line for predicting yield from rainfall.

16. **4 PTS** Use the regression line from Problem 15. (a) A field that received 18 inches of rain yielded 3.1 tons per acre. Find its residual. (b) What percent of the variation in hay yield is accounted for by the linear relationship with rainfall?

17. **4 PTS** Over one summer, the number of ice cream cones sold each day at a farm stand and the number of visitors treated for sunburn at the nearby park have a strong positive correlation. Does eating ice cream cause sunburn? Explain, and name a lurking variable.

18. **4 PTS** To estimate the average number of hours county residents spend reading each week, a librarian asks the first 50 people who walk into the county library on a Saturday. Explain why this sampling method is biased, and say whether the estimate is likely too high or too low.

19. **4 PTS** A church directory lists 400 families in alphabetical order. The secretary chooses a random number from 1 to 20, starts with that family, and then takes every 20th family after it. (a) What kind of sample is this? (b) How many families are in the sample?

20. **4 PTS** Samuel has 30 potted bean plants and wants to know whether a new organic fertilizer increases the weight of beans a plant produces. Describe a completely randomized design for his experiment, and explain why random assignment is important.

21. 4 PTS The table classifies the 120 members of a church youth group. One member is chosen at random. What is the probability that the member sings in the choir or plays an instrument?

Youth group members

	Plays an instrument	Does not play	Total
Sings in choir	18	22	40
Does not sing in choir	30	50	80
Total	48	72	120

22. 4 PTS Use the youth group table. (a) Find the probability that a randomly chosen member plays an instrument, given that the member sings in the choir. (b) Are the events "sings in the choir" and "plays an instrument" independent or not independent?

Youth group members

	Plays an instrument	Does not play	Total
Sings in choir	18	22	40
Does not sing in choir	30	50	80
Total	48	72	120

23. 4 PTS Samuel makes 70% of his free throws, and his shots are independent. He shoots 3 free throws. What is the probability that he misses at least one?

24. 4 PTS A seed company tests each lot of seed before selling it. 90% of lots are good. The test passes 95% of good lots, but it also passes 20% of bad lots. (a) What is the probability that a randomly chosen lot passes the test? (b) Given that a lot passes, what is the probability that it is good? Round to 4 decimal places.

25. 4 PTS A homeschool co-op has 5 juniors and 7 seniors who volunteer for a 3-person planning committee. (a) How many different committees of 3 can be chosen from the 12 volunteers? (b) If the committee is chosen at random, what is the probability that all three members are seniors? Round to 4 decimal places.

MIDTERM EXAM: ANSWER KEY

1. (a) Categorical: breed places each hen in a group. (b) Discrete: it is a count (0, 1, 2, ...). (c) Continuous: a weight can take any value in an interval.
2. (a) $\frac{46}{200} = 0.23$, so 23% were history books. (b) Each of the 200 books falls in exactly one of the four types, so the four counts are parts of one whole (they add to 200), which is exactly what a pie chart shows.
3. (a) $\frac{18}{80} = 0.225$, so 22.5% of juniors won blue. (b) $\frac{36}{120} = 0.30$, so 30% of seniors won blue. A higher percent of seniors than juniors won blue ribbons, so these data show an association between age group and ribbon result.
4. (a) Skewed right: most storms brought less than 1 inch, with a long tail toward the few storms that brought 2 to 3 inches. (b) $4 + 2 + 1 = 7$ storms.
5. Weighted mean $= \frac{120(18) + 80(23)}{200} = \frac{2160 + 1840}{200} = \frac{4000}{200} = \$20.00$ per bushel. It is not $\frac{18 + 23}{2} = \$20.50$ because more bushels were sold at the lower price, so the lower price must count more.
6. Median = 8th value = 18. Lower half (first 7 values): Q1 = 14. Upper half (last 7 values): Q3 = 24. Five-number summary: min 8, Q1 14, median 18, Q3 24, max 44 pounds.
7. (a) IQR $= 24 - 14 = 10$; upper fence $= 24 + 1.5(10) = 39$ pounds. (b) Yes: 44 pounds is above 39 pounds, so it is an outlier.
8. $\bar{x} = \frac{125}{5} = 25$ ounces. Deviations: $-3, -1, 0, 1, 3$; squares: $9, 1, 0, 1, 9$, sum 20. $s^2 = \frac{20}{4} = 5$, so $s = \sqrt{5} \approx 2.24$ ounces. A typical loaf is about 2.24 ounces from the mean of 25 ounces.
9. (a) $z = \frac{96 - 84}{8} = 1.50$. (b) $z = \frac{116 - 96}{10} = 2.00$. The sunflower is 2 standard deviations above its mean, more than the corn stalk's 1.5, so the sunflower is taller relative to its own kind.
10. (a) The density height is $\frac{1}{20}$; the area is $(12 - 5) \times \frac{1}{20} = 0.35$. (b) The median splits the area in half: 10 minutes.
11. 50 grams is 2 standard deviations below the mean and 62 grams is 1 standard deviation above. Percent $= 47.5\% + 34\% = 81.5\%$. About 81.5% of the eggs weigh between 50 and 62 grams.
12. (a) $z = \frac{270 - 240}{20} = 1.50$; area to the right $= 1 - 0.9332 = 0.0668$. (b) $z = \frac{225 - 240}{20} = -0.75$; area $= 0.9332 - 0.2266 = 0.7066$. About 70.7% of his drives travel between 225 and 270 yards.
13. The area to the left is 0.90, so $z = 1.28$ (technology: 1.2816). $x = 18 + 1.2816(2.5) \approx 21.2$ pounds. A watermelon must weigh at least about 21.2 pounds to be premium.

14. $\bar{x} = 6$, $\bar{y} = 5$. Deviations in x: $-2, -1, 0, 1, 2$; in y: $-2, 0, -1, 2, 1$. Products: $4, 0, 0, 2, 2$, sum 8. $s_x = s_y = \sqrt{\dfrac{10}{4}} = \sqrt{2.5}$. $r = \dfrac{1}{4} \cdot \dfrac{8}{2.5} = 0.80$. There is a fairly strong positive linear relationship between sunlight and tomato yield.

15. $b = r\dfrac{s_y}{s_x} = 0.8 \times \dfrac{0.5}{4} = 0.1$; $a = \bar{y} - b\bar{x} = 3.0 - 0.1(20) = 1.0$. So $\hat{y} = 1.0 + 0.1x$: for each additional inch of rain, the predicted hay yield increases by 0.1 ton per acre.

16. (a) $\hat{y} = 1.0 + 0.1(18) = 2.8$; residual $= 3.1 - 2.8 = 0.3$ ton per acre (the field yielded 0.3 ton per acre more than predicted). (b) $r^2 = 0.8^2 = 0.64$, so 64%.

17. No. The data are observational, and correlation does not show cause. A lurking variable such as hot, sunny weather (daily temperature or hours of sunshine) increases both ice cream sales and sunburns, so the two rise together without one causing the other.

18. This is a convenience sample: people who visit the library on a Saturday are more likely to be frequent readers than county residents in general. The method systematically favors readers, so the estimate is likely too high.

19. (a) A systematic random sample. (b) $\dfrac{400}{20} = 20$ families.

20. Label the plants 1 to 30. Use a random number generator (or draw numbers from a hat) to choose 15 plants for the new fertilizer; the other 15 get no fertilizer (the control group). Keep everything else the same (pot size, soil, water, sunlight), and at the end compare the mean weight of beans per plant in the two groups. Random assignment makes the two groups alike on average in everything except the fertilizer, so a large difference in yield can be attributed to the fertilizer rather than to differences among the plants.

21. $P(\text{choir or instrument}) = \dfrac{40}{120} + \dfrac{48}{120} - \dfrac{18}{120} = \dfrac{70}{120} \approx 0.5833$. The 18 members in both groups are subtracted so they are not counted twice.

22. (a) $P(\text{instrument} \mid \text{choir}) = \dfrac{18}{40} = 0.45$. (b) $P(\text{instrument}) = \dfrac{48}{120} = 0.40$. Since $0.45 \neq 0.40$, knowing a member sings in the choir changes the probability, so the events are not independent.

23. $P(\text{at least one miss}) = 1 - P(\text{all three made}) = 1 - 0.7^3 = 1 - 0.343 = 0.657$.

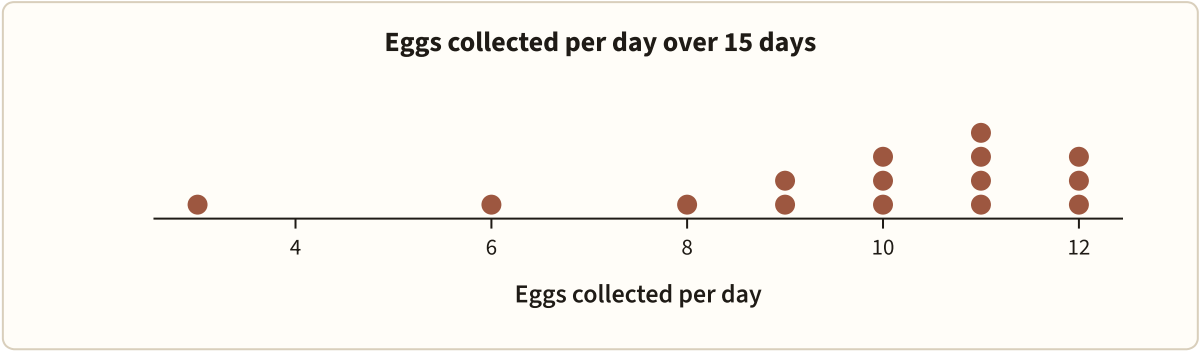
24. (a) Tree: $P(\text{good and pass}) = 0.9 \times 0.95 = 0.855$; $P(\text{bad and pass}) = 0.1 \times 0.2 = 0.02$; $P(\text{pass}) = 0.855 + 0.02 = 0.875$. (b) $P(\text{good} \mid \text{pass}) = \dfrac{0.855}{0.875} \approx 0.9771$.

25. (a) $\binom{12}{3} = \dfrac{12 \times 11 \times 10}{3 \times 2 \times 1} = 220$. (b) $\binom{7}{3} = 35$ all-senior committees, so $P = \dfrac{35}{220} \approx 0.1591$.

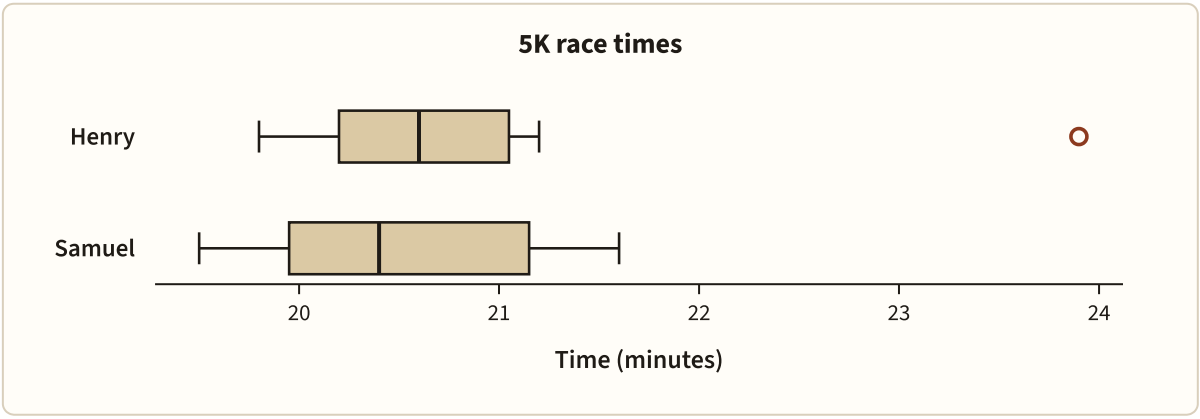
Final Exam

Covers Units 1–13. Time: 2 hours. You have 2 hours. You may use a scientific calculator and the normal, t and chi-square tables at the back of the course. Show your work for every problem: write the formula, the numbers you substitute and the result. Interpret answers in a sentence in the context of the problem, with units. For confidence intervals and significance tests, show the four steps (State, Plan, Do, Conclude) briefly and check the conditions. Keep full precision until the last step; unless a problem says otherwise, round probabilities and P-values to 4 decimal places and z, t and chi-square statistics to 2 decimal places. Answers that differ from the key only in the last decimal place because of the tables are correct. There are 30 problems worth 100 points in all; the points for each problem are shown.

1. **3 PTS** The dotplot shows the number of eggs the Miller family collected from their 12 hens on each of 15 days. (a) Find the median number of eggs per day. (b) Describe the shape of the distribution (symmetric, skewed left or skewed right).

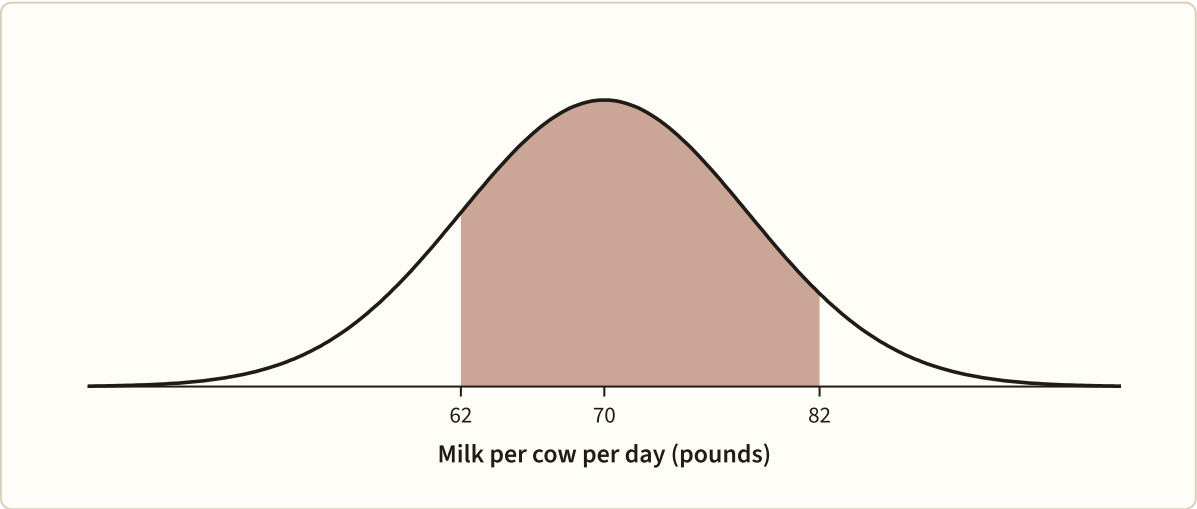


2. **3 PTS** The boxplots compare Henry's and Samuel's times, in minutes, in nine 5K races each. Henry's times were 19.8, 20.1, 20.3, 20.4, 20.6, 20.7, 20.9, 21.2, 23.9. (a) Find the IQR of Henry's times. (b) Is Henry's time of 23.9 minutes an outlier by the 1.5 × IQR rule?



3. **3 PTS** A sawmill's boards have mean length 96.2 inches and standard deviation 0.3 inch. Every board is trimmed by 2 inches, and then the lengths are converted to centimeters (1 inch = 2.54 cm). Find the mean and standard deviation of the trimmed lengths in centimeters.

4. 3 PTS The daily milk yield of the cows in a large dairy herd is approximately normal with mean 70 pounds and standard deviation 8 pounds. What proportion of the cows give between 62 and 82 pounds of milk per day?



5. 3 PTS Challenge. A sawmill cuts boards whose lengths are approximately normal with mean 96.0 inches. The buyer will accept the boards only if no more than 5% are shorter than 95.5 inches. What is the largest standard deviation the mill's saw can have? Round to three decimal places.

6. 3 PTS An orchard keeper regressed the yield of apple trees (bushels per tree) on the age of the tree (years) for trees 4 to 12 years old. The computer output is shown. (a) Predict the yield of an 8-year-old tree. (b) Find the correlation r.

Regression of apple yield on tree age

Predictor	Coef
Constant	2.4
Age (years)	1.15
S	1.8
R-sq	81.0%

7. 3 PTS A survey of library patrons asked: "Wouldn't you agree that our wonderful library should stay open later on weekday evenings?" Explain what is wrong with this question, and rewrite it fairly.

8. 3 PTS An agricultural college randomly assigned the 60 calves on its own farm to two feeds, 30 calves each. The calves on Feed B gained significantly more weight than those on Feed A. (a) Can we conclude that Feed B caused the greater weight gain for these calves? (b) Can we generalize the result to all calves in the country? Answer yes or no and explain.

9. 3 PTS The forecast gives a 0.3 probability of rain on Saturday and a 0.4 probability of rain on Sunday. Assume the two days are independent. What is the probability that it rains on at least one of the two days?

10. 3 PTS Of 10 pies entered at a church bake sale contest, 4 are apple. A judge chooses 3 of the pies at random to taste first. What is the probability that all three are apple? Round to 4 decimal places.

11. 3 PTS A farmer must choose whether to plant a field in corn or in hay. The table shows the profit for each crop in good weather (probability 0.7) and in poor weather (probability 0.3). (a) Find the expected profit for corn. (b) Find the expected profit for hay.

Profit per field by weather

Weather	Probability	Corn profit (dollars)	Hay profit (dollars)
Good	0.7	12000	9000
Poor	0.3	2000	6000

12. 3 PTS Empty honey jars have weights with mean 8.0 ounces and standard deviation 0.3 ounce. The honey put in each jar has weight with mean 16.0 ounces and standard deviation 0.4 ounce, independent of the jar. Find the mean and standard deviation of the weight of a filled jar.

13. 4 PTS Challenge. A carpenter cuts shelf boards whose lengths are normal with mean 48.0 inches and standard deviation 0.06 inch. The spaces between the uprights of his bookcases are normal with mean 48.1 inches and standard deviation 0.08 inch, independent of the boards. A board fits if it is shorter than the space. What is the probability that a randomly chosen board fits a randomly chosen space?

14. 3 PTS Each pepper seed from a certain packet germinates with probability 0.8, independently. Grace plants 6 seeds. What is the probability that exactly 5 germinate? Round to 4 decimal places.

15. 3 PTS Ruth makes 75% of her free throws, independently. In a season she shoots 48 free throws. Let X be the number she makes. Find the mean and standard deviation of X.

16. 3 PTS In a board game, a player must roll a 6 on a fair die to start. Let X be the roll on which Henry gets his first 6. (a) Find $P(X = 4)$. Round to 4 decimal places. (b) Find the mean of X.

17. 3 PTS In a large apple harvest, 20% of the apples have blemishes. An inspector takes a random sample of 100 apples. (a) Find the standard deviation of the sample proportion $\hat{p}$ of blemished apples. (b) Find the probability that more than 26% of the sample are blemished.

18. 3 PTS The weaning weights of lambs on a large ranch are normal with mean 64 pounds and standard deviation 8 pounds. A random sample of 16 lambs is weighed. (a) Find the standard deviation of the sample mean $\bar{x}$. (b) Find the probability that the sample mean is more than 68.5 pounds.

19. **4 PTS** A hatchery incubated a random sample of 400 eggs from its large supply, and 340 of them hatched. Construct and interpret a 95% confidence interval for the proportion of all the hatchery's eggs that hatch. Round the endpoints to three decimal places.

20. **4 PTS** A rancher weighs a random sample of 16 newborn calves from a large herd. The mean is 82.4 pounds and the standard deviation is 6.0 pounds; a dotplot of the weights shows no strong skewness and no outliers. (a) Construct a 95% confidence interval for the mean birth weight of all calves in the herd. Round the endpoints to one decimal place. (b) Explain what "95% confident" means here.

21. **3 PTS** A golf coach wants to estimate the mean drive distance of the members of a large golf club within 2 yards with 95% confidence. From past records she uses a planning value of $\sigma = 12$ yards. How many drives (one per golfer) must she measure?

22. **3 PTS** A beekeeper tested H_0: $\mu = 40$ pounds against H_a: $\mu > 40$ pounds, where μ is the mean honey harvest per hive with a new hive design. The P-value was 0.03. Explain what this P-value means in context.

23. **4 PTS** A pecan grower's records show that 80% of his nuts are free of insect damage. After a wet summer, he suspects that the proportion is lower. In a random sample of 400 nuts from this year's large harvest, 304 are free of damage. Carry out a test at the $\alpha = 0.05$ level.

24. **4 PTS** Jars of honey are labeled 16 ounces. An inspector weighs the honey in a random sample of 25 jars from a large shipment: $\bar{x} = 15.92$ ounces and $s = 0.20$ ounce, and a graph of the data shows no strong skewness or outliers. Is there convincing evidence that the mean amount of honey in all the jars is less than 16 ounces? Use $\alpha = 0.05$, and give the t statistic, the degrees of freedom and bounds for the P-value from the t table.

25. **3 PTS** Use the honey test of Problem 24. (a) Describe a Type I error in context. (b) If H_0 is true, what is the probability of making a Type I error? (c) Describe a Type II error in context.

26. **4 PTS** A gardener tests two brands of tomato seed. In a random sample of 200 seeds of Brand A, 180 germinate; in a random sample of 200 seeds of Brand B, 160 germinate. Is there convincing evidence that the germination rates of the two brands differ? Use $\alpha = 0.05$.

27. **4 PTS** Two varieties of wheat were planted on randomly assigned plots. The table gives the summary statistics. Construct a 95% confidence interval for the difference in mean yield (Variety A minus Variety B), using the conservative degrees of freedom. Round the endpoints to two decimal places. Assume the conditions are met.

Summary statistics for two wheat varieties

Variety	Number of plots	Mean yield (bushels per acre)	Standard deviation
A	10	52	4
B	12	47	6

28. **4 PTS** Challenge. Six garden beds were each split in half; a coin flip decided which half got Fertilizer A and which got Fertilizer B. The table gives the tomato yields in pounds. Is there convincing evidence that Fertilizer A gives a higher mean yield? Use $\alpha = 0.05$, and give the t statistic, the degrees of freedom and bounds for the P-value from the t table. Assume the differences show no strong skewness or outliers.

Tomato yields (pounds) in six split garden beds

Bed	1	2	3	4	5	6
Fertilizer A	12	15	10	14	17	13
Fertilizer B	11	12	11	12	13	10

29. **4 PTS** A wildflower seed mix is advertised as 50% poppy, 25% cornflower and 25% daisy seeds. Grace plants the mix and identifies a random sample of 200 seedlings; the counts are in the table. Do the data give convincing evidence that the mix differs from the advertised percents? Use $\alpha = 0.05$.

Wildflower seedlings

Flower	Advertised percent	Observed count
Poppy	50%	112
Cornflower	25%	38
Daisy	25%	50
Total	100%	200

30. **4 PTS** An inspector graded a random sample of 100 apples from each of three orchards; the results are in the table. Is there convincing evidence that the distribution of grades differs among the three orchards? Use $\alpha = 0.05$.

Apple grades from three orchards

Grade	Orchard A	Orchard B	Orchard C	Total
Grade 1	70	60	50	180
Grade 2	30	40	50	120
Total	100	100	100	300

FINAL EXAM: ANSWER KEY

1. (a) The 8th of the 15 ordered values is 10, so the median is 10 eggs per day. (b) Skewed left: most days gave 9 to 12 eggs, with a tail toward a few low days (3 and 6 eggs).
2. (a) Median $= 20.6$; Q1 = median of 19.8, 20.1, 20.3, 20.4 $= 20.2$; Q3 = median of 20.7, 20.9, 21.2, 23.9 $= 21.05$; IQR $= 21.05 - 20.2 = 0.85$ minute. (b) Upper fence $= 21.05 + 1.5(0.85) = 22.325$ minutes; 23.9 is above it, so yes, it is an outlier.
3. Trimming subtracts 2 from the mean and leaves the spread alone: mean 94.2 inches, standard deviation 0.3 inch. Multiplying by 2.54 multiplies both: mean $= 94.2 \times 2.54 = 239.268$ cm, standard deviation $= 0.3 \times 2.54 = 0.762$ cm.
4. $z = \dfrac{62 - 70}{8} = -1.00$ and $z = \dfrac{82 - 70}{8} = 1.50$. Area $= 0.9332 - 0.1587 = 0.7745$. About 77.45% of the cows give between 62 and 82 pounds per day.
5. The 5th percentile must be at 95.5 inches: $z = -1.645$. From $95.5 = 96.0 - 1.645\sigma$, $\sigma = \dfrac{0.5}{1.645} \approx 0.304$ inch. The saw's standard deviation must be at most about 0.304 inch.
6. (a) $\hat{y} = 2.4 + 1.15(8) = 11.6$ bushels. (b) $r = \sqrt{0.81} = 0.9$, positive because the slope is positive. For each additional year of age, the predicted yield increases by about 1.15 bushels per tree.
7. The wording leads the respondent toward "yes" ("Wouldn't you agree", "wonderful"), which causes response bias: the percent who say yes will likely be too high. A fair version: "Should the library's weekday closing time be changed? (Keep it at 6 p.m. / Change it to 8 p.m.)"
8. (a) Yes: the feeds were randomly assigned, so the groups were alike apart from the feed, and the significant difference can be attributed to the feed. (b) No: the calves were not a random sample from all calves in the country; they were only the calves on this farm, so the result applies to calves like these.
9. $P(\text{no rain either day}) = 0.7 \times 0.6 = 0.42$, so $P(\text{at least one rainy day}) = 1 - 0.42 = 0.58$.
10. $P = \dfrac{\binom{4}{3}}{\binom{10}{3}} = \dfrac{4}{120} \approx 0.0333$.
11. (a) $E = 0.7(\$12{,}000) + 0.3(\$2{,}000) = \$8{,}400 + \$600 = \$9{,}000$. (b) $E = 0.7(\$9{,}000) + 0.3(\$6{,}000) = \$6{,}300 + \$1{,}800 = \$8{,}100$. Corn earns more on average in the long run, but hay is less risky in a poor year.
12. Means add: $\mu = 8.0 + 16.0 = 24.0$ ounces. Variances add for independent variables: $\sigma = \sqrt{0.3^2 + 0.4^2} = \sqrt{0.09 + 0.16} = \sqrt{0.25} = 0.5$ ounce (not $0.3 + 0.4 = 0.7$).
13. Let $D = \text{space} - \text{board}$. $\mu_D = 48.1 - 48.0 = 0.1$ inch; $\sigma_D = \sqrt{0.06^2 + 0.08^2} = \sqrt{0.01} = 0.1$ inch; D is normal. $P(D > 0)$: $z = \dfrac{0 - 0.1}{0.1} = -1.00$; area to the right $= 1 - 0.1587 = 0.8413$. About 84% of the boards fit.
14. $P(X = 5) = \binom{6}{5}(0.8)^5(0.2)^1 = 6 \times 0.32768 \times 0.2 = 0.393216 \approx 0.3932$.

15. $\mu = np = 48(0.75) = 36$ free throws; $\sigma = \sqrt{np(1-p)} = \sqrt{48(0.75)(0.25)} = \sqrt{9} = 3$ free throws.

16. (a) $P(X = 4) = \left(\frac{5}{6}\right)^3 \left(\frac{1}{6}\right) = \frac{125}{1296} \approx 0.0965$. (b) $\mu = \frac{1}{p} = 6$ rolls: on average Henry needs 6 rolls to start.

17. (a) $\sigma_{\hat{p}} = \sqrt{\frac{0.2(0.8)}{100}} = \sqrt{0.0016} = 0.04$. (b) Large Counts: $100(0.2) = 20$ and $100(0.8) = 80$ are at least 10, so $\hat{p}$ is approximately normal. $z = \frac{0.26 - 0.20}{0.04} = 1.50$; area to the right $= 1 - 0.9332 = 0.0668$.

18. (a) $\sigma_{\bar{x}} = \frac{8}{\sqrt{16}} = 2$ pounds. (b) The population is normal, so $\bar{x}$ is normal. $z = \frac{68.5 - 64}{2} = 2.25$; area to the right $= 1 - 0.9878 = 0.0122$.

19. State: estimate p, the proportion of all the hatchery's eggs that hatch, with 95% confidence. Plan: one-proportion z interval; random sample; 400 is less than 10% of the large supply; 340 and 60 are both at least 10. Do: $\hat{p} = \frac{340}{400} = 0.85$; $0.85 \pm 1.960\sqrt{\frac{0.85(0.15)}{400}} = 0.85 \pm 0.035$, giving $(0.815, 0.885)$. Conclude: we are 95% confident that between 81.5% and 88.5% of all the hatchery's eggs hatch.

20. (a) One-sample t interval, df $= 15$, $t^* = 2.131$. $82.4 \pm 2.131 \times \frac{6.0}{\sqrt{16}} = 82.4 \pm 3.20$, giving $(79.2, 85.6)$ pounds. We are 95% confident that the mean birth weight of all calves in the herd is between 79.2 and 85.6 pounds. (b) If we took many random samples of 16 calves and built an interval from each in this way, about 95% of those intervals would capture the true mean birth weight. It does not mean there is a 95% probability that the true mean is in this particular interval.

21. $n = \left(\frac{z^* \sigma}{ME}\right)^2 = \left(\frac{1.960 \times 12}{2}\right)^2 = 11.76^2 = 138.2976$. Round up: $n = 139$ golfers.

22. If the true mean harvest with the new hive design were 40 pounds per hive, there would be only a 0.03 probability of getting a sample mean as large as the one observed (or larger) by chance alone. It is not the probability that H_0 is true.

23. State: H_0: $p = 0.80$, H_a: $p < 0.80$, where p is the proportion of this year's nuts free of damage; $\alpha = 0.05$. Plan: one-proportion z test; random sample; 400 is less than 10% of the harvest; $400(0.80) = 320$ and $400(0.20) = 80$ are at least 10. Do: $\hat{p} = \frac{304}{400} = 0.76$; $z = \frac{0.76 - 0.80}{\sqrt{\frac{0.80(0.20)}{400}}} = \frac{-0.04}{0.02} = -2.00$; P-value $= 0.0228$. Conclude: since $0.0228 < 0.05$, reject H_0. There is convincing evidence that less than 80% of this year's nuts are free of damage.

24. State: H_0: $\mu = 16$, H_a: $\mu < 16$, where μ is the mean amount of honey in all jars in the shipment; $\alpha = 0.05$. Plan: one-sample t test; random sample; 25 is less than 10% of the shipment; no strong skewness or outliers. Do: $t = \dfrac{15.92 - 16}{0.20/\sqrt{25}} = \dfrac{-0.08}{0.04} = -2.00$, $\text{df} = 24$; 2.00 lies between 1.711 and 2.064, so $0.025 < \text{P-value} < 0.05$ (technology: about 0.03). Conclude: since the P-value is less than 0.05, reject H_0. There is convincing evidence that the mean amount of honey in the jars is less than 16 ounces.

25. (a) Concluding that the mean amount of honey is less than 16 ounces when it really is 16 ounces; the packer might needlessly adjust or stop the filling machine. (b) $\alpha = 0.05$. (c) Failing to find convincing evidence that the mean is less than 16 ounces when it really is less; customers would keep receiving underfilled jars.

26. State: H_0: $p_A = p_B$, H_a: $p_A \neq p_B$, where p_A and p_B are the proportions of all seeds of each brand that germinate; $\alpha = 0.05$. Plan: two-proportion z test; independent random samples, each less than 10% of the brand's seeds; pooled $\hat{p}_c = \dfrac{340}{400} = 0.85$, and the expected counts $200(0.85) = 170$ and $200(0.15) = 30$ are at least 10. Do: $\hat{p}_A = 0.90$, $\hat{p}_B = 0.80$; $z = \dfrac{0.90 - 0.80}{\sqrt{0.85(0.15)\left(\dfrac{1}{200} + \dfrac{1}{200}\right)}} = \dfrac{0.10}{0.0357} \approx 2.80$; P-value $= 2(0.00255) \approx 0.0051$. Conclude: since $0.0051 < 0.05$, reject H_0. There is convincing evidence that the germination rates of the two brands differ; Brand A appears higher.

27. Two-sample t interval; df = smaller of 9 and 11 $= 9$, $t^* = 2.262$. $SE = \sqrt{\dfrac{4.0^2}{10} + \dfrac{6.0^2}{12}} = \sqrt{1.6 + 3.0} = \sqrt{4.6} \approx 2.1448$. $(52.0 - 47.0) \pm 2.262(2.1448) = 5.0 \pm 4.85$, giving $(0.15, 9.85)$ bushels per acre. We are 95% confident that Variety A's mean yield is between 0.15 and 9.85 bushels per acre higher than Variety B's.

28. State: H_0: $\mu_d = 0$, H_a: $\mu_d > 0$, where μ_d is the true mean difference in yield (A minus B); $\alpha = 0.05$. Plan: paired t test on the differences (the data are paired by bed); random assignment within each bed; no strong skewness or outliers. Do: differences $1, 3, -1, 2, 4, 3$; $\bar{x}_d = 2$, $s_d = \sqrt{\dfrac{16}{5}} \approx 1.789$; $t = \dfrac{2}{1.789/\sqrt{6}} \approx 2.74$, $\text{df} = 5$; 2.74 lies between 2.571 and 2.757, so $0.02 < \text{P-value} < 0.025$. Conclude: since the P-value is less than 0.05, reject H_0. There is convincing evidence that Fertilizer A gives a higher mean tomato yield than Fertilizer B.

29. State: H_0: the proportions are 0.50, 0.25, 0.25; H_a: at least one proportion differs; $\alpha = 0.05$. Plan: chi-square goodness-of-fit test; random sample; expected counts 100, 50, 50 are all at least 5. Do: $\chi^2 = \dfrac{(112 - 100)^2}{100} + \dfrac{(38 - 50)^2}{50} + \dfrac{(50 - 50)^2}{50} = 1.44 + 2.88 + 0 = 4.32$, $\text{df} = 2$, P-value ≈ 0.1153. Conclude: since $0.1153 > 0.05$, fail to reject H_0. There is not convincing evidence that the seed mix differs from the advertised percents.

30. State: H_0: the distribution of grades is the same for all three orchards; H_a: the distributions are not all the same; $\alpha = 0.05$. Plan: chi-square test for homogeneity; independent random samples; expected counts $\frac{180 \times 100}{300} = 60$ (Grade 1) and $\frac{120 \times 100}{300} = 40$ (Grade 2) are all at least 5. Do: $\chi^2 = \frac{10^2}{60} + 0 + \frac{10^2}{60} + \frac{10^2}{40} + 0 + \frac{10^2}{40} \approx 8.33$, $\text{df} = (2-1)(3-1) = 2$, P-value ≈ 0.0155. Conclude: since $0.0155 < 0.05$, reject H_0. There is convincing evidence that the distribution of apple grades differs among the three orchards; Orchard A has the most Grade 1 apples and Orchard C the fewest.

APPENDIX

Statistical tables

Computed with software for this course. Answers from these tables may differ from technology answers in the last decimal place.

Table A. Standard normal probabilities

Each entry is the area under the standard normal curve to the LEFT of z. Row: z to one decimal; column: the second decimal.

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−3.4	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0003	0.0002
−3.3	0.0005	0.0005	0.0005	0.0004	0.0004	0.0004	0.0004	0.0004	0.0004	0.0003
−3.2	0.0007	0.0007	0.0006	0.0006	0.0006	0.0006	0.0006	0.0005	0.0005	0.0005
−3.1	0.0010	0.0009	0.0009	0.0009	0.0008	0.0008	0.0008	0.0008	0.0007	0.0007
−3.0	0.0013	0.0013	0.0013	0.0012	0.0012	0.0011	0.0011	0.0011	0.0010	0.0010
−2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
−2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
−2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
−2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
−2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
−2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
−2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
−2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
−2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
−2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
−1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
−1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
−1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
−1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
−1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
−1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
−1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
−1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
−1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
−1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
−0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611
−0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
−0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
−0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
−0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
−0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
−0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
−0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
−0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
−0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9177
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964

z	.00	.01	.02	.03	.04	.05	.06	.07	.08	.09
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986
3.0	0.9987	0.9987	0.9987	0.9988	0.9988	0.9989	0.9989	0.9989	0.9990	0.9990
3.1	0.9990	0.9991	0.9991	0.9991	0.9992	0.9992	0.9992	0.9992	0.9993	0.9993
3.2	0.9993	0.9993	0.9994	0.9994	0.9994	0.9994	0.9994	0.9995	0.9995	0.9995
3.3	0.9995	0.9995	0.9995	0.9996	0.9996	0.9996	0.9996	0.9996	0.9996	0.9997
3.4	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9997	0.9998

Table B. t distribution critical values

Each entry is the value t* with area p to its right. The bottom rows give z* and the confidence level C for a two-sided interval.

df	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.000	1.376	1.963	3.078	6.314	12.706	15.895	31.821	63.657	127.321	318.309	636.619
2	0.816	1.061	1.386	1.886	2.920	4.303	4.849	6.965	9.925	14.089	22.327	31.599
3	0.765	0.978	1.250	1.638	2.353	3.182	3.482	4.541	5.841	7.453	10.215	12.924
4	0.741	0.941	1.190	1.533	2.132	2.776	2.999	3.747	4.604	5.598	7.173	8.610
5	0.727	0.920	1.156	1.476	2.015	2.571	2.757	3.365	4.032	4.773	5.893	6.869
6	0.718	0.906	1.134	1.440	1.943	2.447	2.612	3.143	3.707	4.317	5.208	5.959
7	0.711	0.896	1.119	1.415	1.895	2.365	2.517	2.998	3.499	4.029	4.785	5.408
8	0.706	0.889	1.108	1.397	1.860	2.306	2.449	2.896	3.355	3.833	4.501	5.041
9	0.703	0.883	1.100	1.383	1.833	2.262	2.398	2.821	3.250	3.690	4.297	4.781
10	0.700	0.879	1.093	1.372	1.812	2.228	2.359	2.764	3.169	3.581	4.144	4.587
11	0.697	0.876	1.088	1.363	1.796	2.201	2.328	2.718	3.106	3.497	4.025	4.437
12	0.695	0.873	1.083	1.356	1.782	2.179	2.303	2.681	3.055	3.428	3.930	4.318
13	0.694	0.870	1.079	1.350	1.771	2.160	2.282	2.650	3.012	3.372	3.852	4.221
14	0.692	0.868	1.076	1.345	1.761	2.145	2.264	2.624	2.977	3.326	3.787	4.140
15	0.691	0.866	1.074	1.341	1.753	2.131	2.249	2.602	2.947	3.286	3.733	4.073
16	0.690	0.865	1.071	1.337	1.746	2.120	2.235	2.583	2.921	3.252	3.686	4.015
17	0.689	0.863	1.069	1.333	1.740	2.110	2.224	2.567	2.898	3.222	3.646	3.965
18	0.688	0.862	1.067	1.330	1.734	2.101	2.214	2.552	2.878	3.197	3.610	3.922
19	0.688	0.861	1.066	1.328	1.729	2.093	2.205	2.539	2.861	3.174	3.579	3.883
20	0.687	0.860	1.064	1.325	1.725	2.086	2.197	2.528	2.845	3.153	3.552	3.850
21	0.686	0.859	1.063	1.323	1.721	2.080	2.189	2.518	2.831	3.135	3.527	3.819
22	0.686	0.858	1.061	1.321	1.717	2.074	2.183	2.508	2.819	3.119	3.505	3.792
23	0.685	0.858	1.060	1.319	1.714	2.069	2.177	2.500	2.807	3.104	3.485	3.768

df	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
24	0.685	0.857	1.059	1.318	1.711	2.064	2.172	2.492	2.797	3.091	3.467	3.745
25	0.684	0.856	1.058	1.316	1.708	2.060	2.167	2.485	2.787	3.078	3.450	3.725
26	0.684	0.856	1.058	1.315	1.706	2.056	2.162	2.479	2.779	3.067	3.435	3.707
27	0.684	0.855	1.057	1.314	1.703	2.052	2.158	2.473	2.771	3.057	3.421	3.690
28	0.683	0.855	1.056	1.313	1.701	2.048	2.154	2.467	2.763	3.047	3.408	3.674
29	0.683	0.854	1.055	1.311	1.699	2.045	2.150	2.462	2.756	3.038	3.396	3.659
30	0.683	0.854	1.055	1.310	1.697	2.042	2.147	2.457	2.750	3.030	3.385	3.646
40	0.681	0.851	1.050	1.303	1.684	2.021	2.123	2.423	2.704	2.971	3.307	3.551
50	0.679	0.849	1.047	1.299	1.676	2.009	2.109	2.403	2.678	2.937	3.261	3.496
60	0.679	0.848	1.045	1.296	1.671	2.000	2.099	2.390	2.660	2.915	3.232	3.460
80	0.678	0.846	1.043	1.292	1.664	1.990	2.088	2.374	2.639	2.887	3.195	3.416
100	0.677	0.845	1.042	1.290	1.660	1.984	2.081	2.364	2.626	2.871	3.174	3.390
z*	0.674	0.842	1.036	1.282	1.645	1.960	2.054	2.326	2.576	2.807	3.090	3.291
C	50%	60%	70%	80%	90%	95%	96%	98%	99%	99.5%	99.8%	99.9%

Table C. Chi-square critical values

Each entry is the value χ^2 with area p to its right.

df	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
1	1.32	1.64	2.07	2.71	3.84	5.02	5.41	6.63	7.88	9.14	10.83	12.12
2	2.77	3.22	3.79	4.61	5.99	7.38	7.82	9.21	10.60	11.98	13.82	15.20
3	4.11	4.64	5.32	6.25	7.81	9.35	9.84	11.34	12.84	14.32	16.27	17.73
4	5.39	5.99	6.74	7.78	9.49	11.14	11.67	13.28	14.86	16.42	18.47	20.00
5	6.63	7.29	8.12	9.24	11.07	12.83	13.39	15.09	16.75	18.39	20.52	22.11
6	7.84	8.56	9.45	10.64	12.59	14.45	15.03	16.81	18.55	20.25	22.46	24.10
7	9.04	9.80	10.75	12.02	14.07	16.01	16.62	18.48	20.28	22.04	24.32	26.02
8	10.22	11.03	12.03	13.36	15.51	17.53	18.17	20.09	21.95	23.77	26.12	27.87
9	11.39	12.24	13.29	14.68	16.92	19.02	19.68	21.67	23.59	25.46	27.88	29.67
10	12.55	13.44	14.53	15.99	18.31	20.48	21.16	23.21	25.19	27.11	29.59	31.42
11	13.70	14.63	15.77	17.28	19.68	21.92	22.62	24.72	26.76	28.73	31.26	33.14
12	14.85	15.81	16.99	18.55	21.03	23.34	24.05	26.22	28.30	30.32	32.91	34.82
13	15.98	16.98	18.20	19.81	22.36	24.74	25.47	27.69	29.82	31.88	34.53	36.48
14	17.12	18.15	19.41	21.06	23.68	26.12	26.87	29.14	31.32	33.43	36.12	38.11
15	18.25	19.31	20.60	22.31	25.00	27.49	28.26	30.58	32.80	34.95	37.70	39.72
16	19.37	20.47	21.79	23.54	26.30	28.85	29.63	32.00	34.27	36.46	39.25	41.31
17	20.49	21.61	22.98	24.77	27.59	30.19	31.00	33.41	35.72	37.95	40.79	42.88
18	21.60	22.76	24.16	25.99	28.87	31.53	32.35	34.81	37.16	39.42	42.31	44.43

df	0.25	0.2	0.15	0.1	0.05	0.025	0.02	0.01	0.005	0.0025	0.001	0.0005
19	22.72	23.90	25.33	27.20	30.14	32.85	33.69	36.19	38.58	40.88	43.82	45.97
20	23.83	25.04	26.50	28.41	31.41	34.17	35.02	37.57	40.00	42.34	45.31	47.50
21	24.93	26.17	27.66	29.62	32.67	35.48	36.34	38.93	41.40	43.78	46.80	49.01
22	26.04	27.30	28.82	30.81	33.92	36.78	37.66	40.29	42.80	45.20	48.27	50.51
23	27.14	28.43	29.98	32.01	35.17	38.08	38.97	41.64	44.18	46.62	49.73	52.00
24	28.24	29.55	31.13	33.20	36.42	39.36	40.27	42.98	45.56	48.03	51.18	53.48
25	29.34	30.68	32.28	34.38	37.65	40.65	41.57	44.31	46.93	49.44	52.62	54.95
26	30.43	31.79	33.43	35.56	38.89	41.92	42.86	45.64	48.29	50.83	54.05	56.41
27	31.53	32.91	34.57	36.74	40.11	43.19	44.14	46.96	49.64	52.22	55.48	57.86
28	32.62	34.03	35.71	37.92	41.34	44.46	45.42	48.28	50.99	53.59	56.89	59.30
29	33.71	35.14	36.85	39.09	42.56	45.72	46.69	49.59	52.34	54.97	58.30	60.73
30	34.80	36.25	37.99	40.26	43.77	46.98	47.96	50.89	53.67	56.33	59.70	62.16
40	45.62	47.27	49.24	51.81	55.76	59.34	60.44	63.69	66.77	69.70	73.40	76.09
50	56.33	58.16	60.35	63.17	67.50	71.42	72.61	76.15	79.49	82.66	86.66	89.56
60	66.98	68.97	71.34	74.40	79.08	83.30	84.58	88.38	91.95	95.34	99.61	102.69
80	88.13	90.41	93.11	96.58	101.88	106.63	108.07	112.33	116.32	120.10	124.84	128.26
100	109.14	111.67	114.66	118.50	124.34	129.56	131.14	135.81	140.17	144.29	149.45	153.17