STATISTICS 1 · 2 · COLLEGE

Grades 10–12 · College | 6 pages

Data, displays, center & spread, normal curve, probability, distributions, CLT, intervals, tests, power, regression, z and t tables.

1 Data | 2 Displays | 3 Center & spread | 4 Normal | 5 Probability | 6 Random variables | 7 Sampling dist. | 8 Intervals | 9 Tests | 10 Errors & power | 11 Regression | 12 Tests & tables

1 DATA & SAMPLING

who you ask decides what you can say

Population every individual you want to describe
Sample the part you actually measure
Parameter a number about the **population**: μ, σ, p
Statistic a number from the **sample**: $\bar{x}$, s, $\hat{p}$
Variable a trait recorded for each individual

MEMORY TRICK
P goes with P, S goes with S
Parameter ↔ **P**opulation (Greek μ, σ); **S**tatistic ↔ **S**ample ($\bar{x}$, s).

Types of variables

Type	Kind	Example
Categorical	nominal	eye color
	ordinal	shirt size
Quantitative	discrete (counts)	siblings
	continuous	height

Levels of measurement

Level	Adds	Example
Nominal	labels	jersey color
Ordinal	order	finishing place
Interval	equal gaps, no true 0	°F, year
Ratio	true zero; ratios work	weight, time

MEMORY TRICK **NOIR**
French for black: **N**ominal, **O**rdinal, **I**nterval, **R**atio, each adding one power.

Sampling methods

Method	How (school of 1,200)
SRS	every set of 60 names equally likely
Stratified	random 15 from **each** grade
Cluster	2 random homerooms, **everyone** in them
Systematic	random start, then every 20th name
Convenience ✗	whoever is in the cafeteria
Voluntary ✗	whoever answers a web poll

MEMORY TRICK
Strata: some from all · Clusters: all from some
Stratified: **some** from **every** group. Cluster: **all** of **some** groups.

⚠ Sources of bias

- **Undercoverage**: some can never be chosen
- **Nonresponse**: chosen people do not answer
- **Response bias**: wording steers answers
- A bigger biased sample is still biased.

Observational study vs experiment

Observational: watch only → association. **Experiment**: assign treatments → **cause and effect**. A **confounding** variable tied to both x and y blurs their link; random assignment spreads it out.

Principles of experimental design

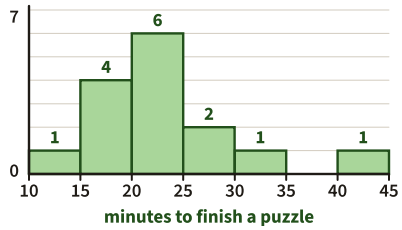
- **Control**: a control or placebo group
- **Randomize**: chance assigns treatments
- **Replicate**: enough subjects per group
- **Block** similar subjects; **double-blind** when possible

Design it WORKED

40 tomato plants: random 20 get A, 20 get B; compare yields

2 DISPLAYING DATA

graph first, then compute

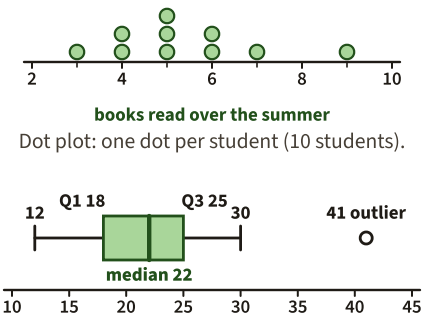


Histogram: bars touch; a class holds its left end (20 ≤ x < 25).

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

key: 2 | 4 = 24 minutes

Stem-and-leaf keeps every value.



Dot plot: one dot per student (10 students).

Box plot: whiskers stop at the last non-outlier.

Pick the display

Data	Use
Categorical	bar chart (gaps), pie chart
Quantitative	dot, stem, histogram, box
Two categorical	two-way table
Two quantitative	scatterplot (§11)



MEMORY TRICK **The tail tells the skew**
Skew is named for the **long tail**, not the peak, and the mean is pulled toward the tail.

MEMORY TRICK **SOCS**
Describe a distribution by **S**hape, **O**utliers, **C**enter and **S**pread, in context.

Frequency language

- relative frequency = $\frac{\text{count}}{n}$: class 20–25 has 6/15 = **40%**
- cumulative: count at or below a value

3 CENTER & SPREAD

where the data sit and how far they wander

The 15 puzzle times (§2)

Measure	Rule	Puzzle times
Mean $\bar{x}$	$\frac{\Sigma x}{n}$	$\frac{334}{15} \approx$ **22.27**
Median	middle value (n odd) or mean of the middle two	8th value = **22**
Mode	most frequent	**22**
Range	max − min	41 − 12 = **29**
IQR	Q3 − Q1	25 − 18 = **7**
SD s	see formula	≈ **6.93**

Sample standard deviation

Population: $\sigma = \sqrt{\frac{\Sigma(x-\mu)^2}{N}}$. Variance = SD^2.

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

Mean from a frequency table WORKED

pets: 1, 2, 3 with frequencies 4, 5, 1

$\bar{x} = \frac{1\cdot4 + 2\cdot5 + 3\cdot1}{10} = \frac{17}{10} = \mathbf{1.7}$ *Σxf ÷ Σf*

SD by hand: 2, 4, 4, 5, 10 WORKED

$\bar{x} = \frac{25}{5} = 5$; deviations −3, −1, −1, 0, 5

squares 9 + 1 + 1 + 0 + 25 = 36

sum = 0 check ✓ on deviations

$s^2 = \frac{36}{4} = 9 \rightarrow \mathbf{s = 3}$ *n − 1 = 4*

MEMORY TRICK

Divide by n − 1 for a sample

$\bar{x}$ used up one degree of freedom. Whole population known? Divide by N.

Quartiles & five-number summary

- Q1, Q3 = medians of the lower and upper halves (median left out when n is odd)
- Puzzle times: **12, 18, 22, 25, 41**
- p-th percentile: p% of values at or below it

1.5 × IQR outlier rule WORKED

IQR = 25 − 18 = 7; 1.5 × 7 = 10.5

fences 18 − 10.5 = **7.5** and 25 + 10.5 = **35.5**

41 > 35.5 → **41 is an outlier**

Resistant or not?

- **Resistant**: median, IQR. **Not**: mean, SD, range.
- Drop the 41: mean 22.27 → **20.93**; median 22 → **21.5**
- Skewed → median & IQR; symmetric → mean & SD

MEMORY TRICK

Median stands firm, mean gets pulled

An extreme value drags the mean; the median barely moves.

Shifting and scaling data

- Add c to every value: center **+ c**, spread unchanged.
- Multiply by k: center **× k**, spread **× |k|**.
- **EX** °C to °F (× 1.8, + 32): mean 20 °C → **68 °F**; SD 5 °C → **9 °F**

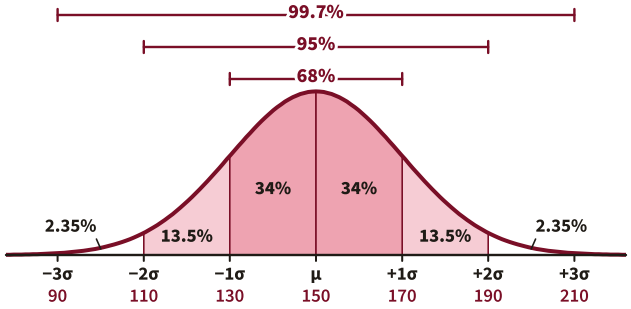
Weighted mean WORKED

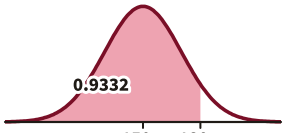
tests 60% (avg 85), homework 40% (avg 95)

0.6 · 85 + 0.4 · 95 = 51 + 38 = **89**

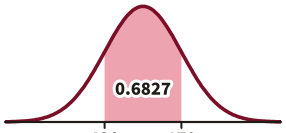
4 Z-SCORES & THE NORMAL CURVE

bell-shaped, symmetric, set by μ and σ





P(X < 180) = Φ(1.5)



P(130 < X < 170)

z-score

$x = \mu + z\sigma$. Sample: $z = \frac{x - \bar{x}}{s}$.

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

MEMORY TRICK

z counts σ-steps from home

z = 2: two SDs above the mean. Unitless, so z compares apples to oranges.

Compare with z WORKED

Quiz A: 85 (μ 75, σ 5) → z = **2.0**

Quiz B: 90 (μ 80, σ 8) → z = **1.25**

A is the stronger score *relative standing*

Empirical rule (68-95-99.7)

- **EX** apples μ = 150 g, σ = 20 g: 95% weigh **110–190 g**; 2.5% weigh more than 190 g.

Normal calculations (apples) WORKED

$P(X < 180)$: $z = \frac{180 - 150}{20} = 1.5 \rightarrow \mathbf{0.9332}$ *table*

$P(X > 180) = 1 - 0.9332 = \mathbf{0.0668}$ *complement*

$P(130 < X < 170) = \Phi(1) - \Phi(-1) = \mathbf{0.6827}$

90th percentile: z = 1.282 → 150 + 1.282 · 20 ≈ **175.6 g**

inverse

Reading Φ(z) (area to the left)

- right of z: 1 − Φ(z) · between: Φ(b) − Φ(a)
- Φ(−z) = 1 − Φ(z); table in §12

Chebyshev's rule (any shape)

At least $1 - \frac{1}{k^2}$ of values lie within k SDs. k = 2: at least **75%**; k = 3: at least **88.9%**.

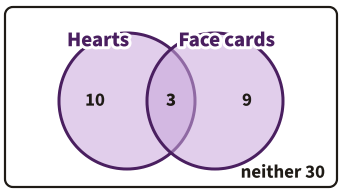
Normal approximation to the binomial

If np ≥ 10 and n(1 − p) ≥ 10, use μ = np, σ = √(np(1 − p)).

5 PROBABILITY & COUNTING

0 = impossible · 1 = certain

Sample space all possible outcomes
Event a set of outcomes
P(A) $\frac{\text{favorable}}{\text{total}}$ when outcomes are equally likely
Complement A^c: A does not happen



52 cards: 13 hearts, 12 face cards, 3 both. P(heart or face) = $\frac{13+12-3}{52} = \frac{22}{52} \approx$ **0.423**.

Probability rules

Rule	Formula
Complement	$P(A^c) = 1 - P(A)$
Addition	$P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$
Disjoint	$P(A \text{ and } B) = 0$ (dependent)
Multiplication	$P(A \text{ and } B) = P(A) \cdot P(B \mid A)$
Independent	$P(B \mid A) = P(B)$
Conditional	$P(B \mid A) = \frac{P(A \text{ and } B)}{P(A)}$

MEMORY TRICK
OR adds, AND multiplies
Or: add, minus the overlap. **And**: multiply (use P(B | A) if dependent).

At least one WORKED

P(at least one 6 in 4 rolls) = $1 - (\frac{5}{6})^4$
complement
$= 1 - 0.482 =$ **0.518**

Two-way table: 200 library visitors

	Fiction	Nonfiction	Total
Adult	70	50	120
Teen	60	20	80
Total	130	70	200

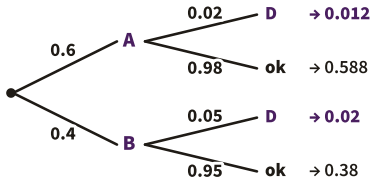
Conditional from the table WORKED

P(teen | fiction) = $\frac{60}{130} \approx$ **0.462**
row ÷ column total
P(teen) = $\frac{80}{200} = 0.40 \neq 0.462 \rightarrow$ **not independent**

Counting

- **Multiply choices**: 3 shirts × 4 pants = **12** outfits; 4-digit PIN: 10^4 = **10,000**
- n! = n(n − 1)⋯1; 0! = 1; 5! = 120
- **Permutations** (order matters): ${}_nP_r = \frac{n!}{(n-r)!}$; 1st–2nd–3rd among 8 runners: ${}_8P_3$ = **336**
- **Combinations** (order ignored): ${}_nC_r = \frac{n!}{r!(n-r)!}$; 3-person committee from 8: ${}_8C_3$ = **56**

MEMORY TRICK
Permutation = Position · Combination = Committee
Swapping two picks makes a new result (gold vs silver)? **Permutation**. Not (a committee)? **Combination**.



Machine A makes 60% of bolts (2% defective); B makes 40% (5% defective).

Bayes' rule from the tree WORKED

P(D) = 0.6 · 0.02 + 0.4 · 0.05 = 0.012 + 0.020 = **0.032**
all D paths
P(A | D) = $\frac{0.012}{0.032}$ = **0.375** *A path ÷ all D*

MEMORY TRICK
Wanted path over all paths
P(A | D) = the A-then-D path ÷ every path ending in D.

6 DISCRETE RANDOM VARIABLES

a number decided by chance

Spinner game (P's sum to 1)

x (prize $)	0	2	5
P(x)	0.5	0.3	0.2

Mean & SD of X WORKED

$\mu = \Sigma x \cdot P(x) = 0 + 0.6 + 1.0 =$ **$1.60**
expected value
$\sigma^2 = \Sigma(x-\mu)^2 P(x) = 1.28 + 0.048 + 2.312 = 3.64$
$\sigma = \sqrt{3.64} \approx$ **$1.91**

Combining random variables

- $E(X \pm Y) = E(X) \pm E(Y)$; $E(a + bX) = a + b\mu$
- independent: $Var(X \pm Y) = Var(X) + Var(Y)$
- $SD(a + bX) = |b|\sigma$

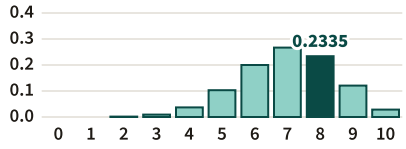
MEMORY TRICK **Variances always add**
Independent X, Y: Var(X − Y) = Var X + Var Y. Add **variances**, never SDs.

MEMORY TRICK **BINS**
Binary outcomes, **I**ndependent trials, fixed **N**umber, **S**ame p.

Binomial

$\mu = np$, $\sigma = \sqrt{np(1-p)}$

$$P(X = k) = {}_nC_k\, p^k (1-p)^{n-k}$$



makes in 10 free throws, p = 0.7

Free throws, p = 0.7, n = 10 WORKED

$P(8) = {}_{10}C_8(0.7)^8(0.3)^2 \approx$ **0.2335**
$P(X \geq 9) = P(9) + P(10) \approx$ **0.1493**
$\mu = 10 \cdot 0.7 =$ **7**; $\sigma = \sqrt{2.1} \approx$ **1.45**

Geometric (trials until the first success)

$P(X = k) = (1-p)^{k-1}p$; $\mu = \frac{1}{p}$

EX first 6 on the 3rd roll: $(\frac{5}{6})^2(\frac{1}{6}) = \frac{25}{216} \approx$ **0.116**; mean wait **6** rolls

Poisson (count in a fixed interval)

$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$; mean = variance = λ

EX λ = 3 online orders per hour: $P(2) = \frac{e^{-3} \cdot 9}{2} \approx$ **0.224**; P(0) ≈ **0.0498**

Continuous random variables

Probability = **area** under a density curve, so P(X = a) = 0.

EX uniform on 0 to 10: $P(2 < X < 5) = \frac{3}{10} =$ **0.3**

7 SAMPLING DISTRIBUTIONS & THE CLT

statistics vary from sample to sample, predictably

Sampling distribution the distribution of a statistic over all samples of size n

Standard error (SE) the SD of a sampling distribution

Unbiased the statistic is centered on the parameter

Two sampling distributions

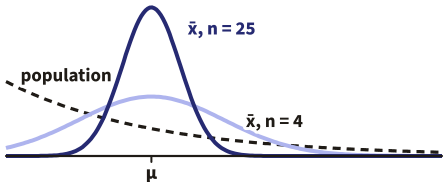
Statistic	Center	SE	Normal when
$\bar{x}$	μ	$\frac{\sigma}{\sqrt{n}}$	population normal, or $n \geq 30$ (CLT)
$\hat{p}$	p	$\sqrt{(p(1-p)/n)}$	$np \geq 10$ and $n(1-p) \geq 10$

10% condition

When sampling without replacement, keep n ≤ 10% of the population so the SE formulas hold.

⚠ **SD or SE?**

SD describes individuals; **SE** describes a statistic. SE = SD/√n is always smaller.



Skewed population (dashed). Means of samples pile up around μ, narrower and more normal as n grows.

Central Limit Theorem

For a large enough random sample (rule of thumb **n ≥ 30**), $\bar{x}$ is approximately normal with mean μ and SE σ/√n, **whatever the population shape**.

MEMORY TRICK

Thirty makes it normal

n ≥ 30 lets the CLT take over for means; strongly skewed populations may need more.

Difference of two means

$\bar{x}_1 - \bar{x}_2$ centers on $\mu_1 - \mu_2$ with SE $\sqrt{(\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2})}$; independent samples.

MEMORY TRICK **√n lives downstairs**

SE = σ/√n: **4 times** the sample halves the spread. n = 25 → SE = 20/5 = **4**; n = 100 → SE = 20/10 = **2**.

Mean of 25 apples WORKED

$\mu = 150$, $\sigma = 20 \rightarrow SE = \frac{20}{\sqrt{25}} = 4$

$P(\bar{x} > 156)$: $z = \frac{156 - 150}{4} = 1.5 \rightarrow 1 - 0.9332 =$ **0.0668**

p̂ with p = 0.3, n = 100 WORKED

$np = 30$, $n(1-p) = 70$: both ≥ 10 ✓

$SE = \sqrt{(\frac{0.3 \cdot 0.7}{100})} \approx 0.0458$

$P(\hat{p} > 0.39)$: $z = \frac{0.09}{0.0458} \approx 1.96 \rightarrow$ **0.025**

Bias vs variability

- **Bias**: aim is off (fix the method).
- **Variability**: shots scatter (fix with a bigger sample).

8 CONFIDENCE INTERVALS

estimate ± margin of error

Every interval

Margin of error ME = critical value × SE.

estimate ± (critical value) (SE)

Common intervals

Parameter	Interval
μ (σ known)	$\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$
μ	$\bar{x} \pm t^* \frac{s}{\sqrt{n}}$, df = n − 1
p	$\hat{p} \pm z^* \sqrt{(\frac{\hat{p}(1-\hat{p})}{n})}$
$\mu_1 - \mu_2$	$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{(\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2})}$
$p_1 - p_2$	$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{(\frac{\hat{p}_1\hat{q}_1}{n_1} + \frac{\hat{p}_2\hat{q}_2}{n_2})}$
μ_d (paired)	$\bar{d} \pm t^* \frac{s_d}{\sqrt{n}}$, df = n − 1

Sample size for a proportion WORKED

ME 0.03 at 95%, no guess for p → use 0.5

$n = (\frac{1.96}{0.03})^2 (0.5)(0.5) = 1{,}067.1 \rightarrow$ **1,068**

always round up

t interval for a mean WORKED

$n = 16$, $\bar{x} = 52.0$, $s = 8 \rightarrow SE = \frac{8}{4} = 2$

95%, df 15: $t^* = 2.131 \rightarrow ME = 4.26$ *t table*

(47.74, 56.26)

Interval for a proportion WORKED

240 of 400 like a new recipe: $\hat{p} = 0.600$

$SE = \sqrt{(\frac{0.6 \cdot 0.4}{400})} \approx 0.0245$; $ME = 1.96 \cdot 0.0245 \approx 0.048$

95% CI **(0.552, 0.648)**

Two-sample t interval WORKED

A: $\bar{x} = 2.4$ kg, $s = 0.3$; B: $\bar{x} = 2.1$, $s = 0.4$; $n = 10$ each

$SE = \sqrt{(\frac{0.09}{10} + \frac{0.16}{10})} \approx 0.158$; df ≈ 16.7 → $t^* \approx 2.11$ (course convention: df = 9, $t^* = 2.262$)

software df

0.3 ± 0.33 → **(−0.03, 0.63)**: 0 is inside



20 simulated 95% intervals: 19 capture μ; one (red, dashed) misses.

What 95% confidence means

If we repeated the sampling many times, about 95% of the intervals built this way would capture the true parameter. The 95% describes the **method**, not one interval.

MEMORY TRICK

More confidence, wider net; more data, tighter net

Raise the confidence level → bigger critical value → wider interval. Quadruple n → half the margin of error.

9 HYPOTHESIS TESTS

is the evidence strong enough to doubt H_0?

The pieces

- **H_0** (null): no effect, uses =. **H_a** (alternative): <, > or ≠.
- **α**: significance level, often 0.05.
- **p-value**: chance of a result at least this extreme **if H_0 is true**.
- $p \le \alpha \rightarrow$ **reject H_0**. $p > \alpha \rightarrow$ **fail to reject H_0** (never "accept").

MEMORY TRICK

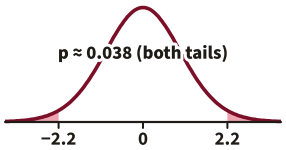
If p is low, the null must go

Traditional: small p = data rare if H_0 were true, so reject H_0.

MEMORY TRICK

Claim · Check · Compute · Conclude

Hypotheses · conditions · statistic and p-value · conclusion **in context**.



Test statistics

Test	Statistic	df	Test	Statistic	df
1 mean, σ known	$z = \frac{\bar{x} - \mu_0}{\sigma/\sqrt{n}}$	—	2 proportions	$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{(\hat{p}\hat{q}(1/n_1 + 1/n_2))}}$, pooled $\hat{p}$	—
1 mean	$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$	n − 1	χ^2	$\Sigma\frac{(O-E)^2}{E}$	k − 1; (r − 1)(c − 1)
1 proportion	$z = \frac{\hat{p} - p_0}{\sqrt{(p_0 q_0/n)}}$	—	ANOVA	$F = \frac{MSB}{MSW}$	k − 1, N − k
2 means	$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{(s_1^2/n_1 + s_2^2/n_2)}}$	software*	slope	$t = \frac{b}{SE_b}$	n − 2
paired	$t = \frac{\bar{d}}{s_d/\sqrt{n}}$	n − 1		* or $\min(n_1, n_2) - 1$ (safe)	

One-sample t (bottle filler) WORKED

H_0: μ = 500 mL vs ≠; n = 25, $\bar{x}$ = 497.8, s = 5

$t = \frac{497.8 - 500}{5/\sqrt{25}} =$ **−2.2**, df 24

p ≈ **0.038** < 0.05 → reject H_0: mean fill is off

χ^2 goodness of fit (fair die?) WORKED

60 rolls: 5, 8, 9, 8, 10, 20; expected 10 each

$\chi^2 = \frac{25 + 4 + 1 + 4 + 0 + 100}{10} =$ **13.4**, df 5

p ≈ **0.020** → evidence the die is unfair

ANOVA: 4,5,6 · 6,7,8 · 8,9,10

Source	SS	df	MS	F
Between	24	2	12	**12**
Within	6	6	1	
Total	30	8		

One-proportion z (fair coin?) WORKED

115 heads in 200 flips: $\hat{p}$ = 0.575; H_0: p = 0.5

$z = \frac{0.575 - 0.5}{\sqrt{(0.25/200)}} \approx$ **2.12**

p ≈ **0.034** (two-sided) → reject at α = 0.05

χ^2 independence (library table) WORKED

expected = $\frac{\text{row total} \times \text{column total}}{\text{grand total}}$: 78, 42, 52, 28

$\chi^2 = \frac{8^2}{78} + \frac{8^2}{42} + \frac{8^2}{52} + \frac{8^2}{28} \approx$ **5.86**, df 1

p ≈ **0.015** → age group and genre are associated

ANOVA overview

Tests H_0: all k means are equal. Large F (between-group spread ≫ within) → small p; here p ≈ **0.008**. A follow-up (e.g., Tukey) finds **which** means differ.

Paired t (after − before) WORKED

d: 3, 5, −1, 4, 2, 5 → $\bar{d} = 3$, $s_d \approx 2.28$

$t = \frac{3}{2.28/\sqrt{6}} \approx$ **3.22**, df 5

one-sided p ≈ **0.012** → scores rose

χ^2 conditions

Counts (not percents), random sample, every **expected** count ≥ 5.

Two-sample t (tomatoes) WORKED

$t = \frac{2.4 - 2.1}{0.158} \approx$ **1.90**, df ≈ 16.7 (course convention: df = 9, p ≈ 0.09)

p ≈ **0.075** → fail to reject (the §8 CI held 0)

10 ERRORS & POWER

every decision can be wrong

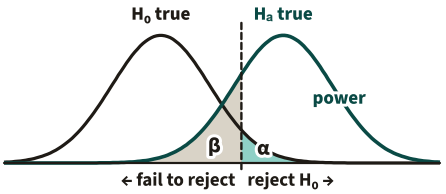
Decision vs truth

Decision	H_0 true	H_0 false
Reject H_0	**Type I error** (prob. α)	correct: **power** = 1 − β
Fail to reject	correct	**Type II error** (prob. β)

MEMORY TRICK

The boy who cried wolf

Aesop: villagers first believe a false alarm (**Type I**: reject a true "no wolf"), then ignore a real wolf (**Type II**).



Effect of 2.5 SE, one-sided α = 0.05: power ≈ 0.80.

Tests and intervals agree

A two-sided test at α = 0.05 rejects H_0: μ = μ_0 exactly when μ_0 lies **outside** the 95% CI.

Raise power by

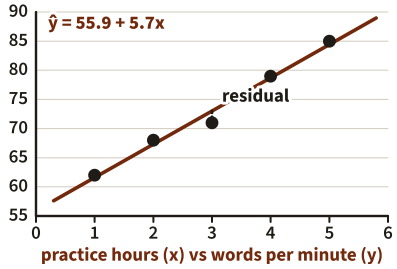
- a **larger sample** n
- a **larger α** (more Type I risk)
- a **bigger true effect**
- **less variability** (better measurement, blocking)

⚠ p-value traps

- p is **not** the chance H_0 is true.
- Significant ≠ important: huge n makes tiny effects significant.

11 CORRELATION & REGRESSION

two quantitative variables



MEMORY TRICK **DOFS**

Describe a scatterplot by **D**irection, **O**utliers, **F**orm (linear or curved), **S**trength, in context.

Correlation r

- $-1 \le r \le 1$; sign = direction, size = strength of a **linear** pattern.
- Unitless; swapping x and y keeps r.
- Not resistant: one outlier can change it a lot.

Least-squares line

$b = r\frac{s_y}{s_x}$, $a = \bar{y} - b\,\bar{x}$; the line passes through $(\bar{x}, \bar{y})$.

$$\hat{y} = a + bx$$

Typing practice (5 students) WORKED

$\bar{x} = 3$, $\bar{y} = 73$; $\Sigma(x-\bar{x})(y-\bar{y}) = 57$, $\Sigma(x-\bar{x})^2 = 10$

$b = \frac{57}{10} = \mathbf{5.7}$; $a = 73 - 5.7 \cdot 3 = \mathbf{55.9}$

$r \approx \mathbf{0.992}$, $r^2 \approx \mathbf{0.985}$

$x = 3$: residual $= 71 - 73 = \mathbf{-2}$ $\quad y - \hat{y}$

MEMORY TRICK **Residual = O − P**

Observed minus **P**redicted. Positive: the point sits above the line.

Read the output in context

- **Slope**: each extra practice hour adds about **5.7 wpm** to predicted speed.
- **$r^2 = 0.985$**: 98.5% of the variation in speed is explained by the line.
- Residual plot with no pattern → a line fits.

⚠ Regression cautions

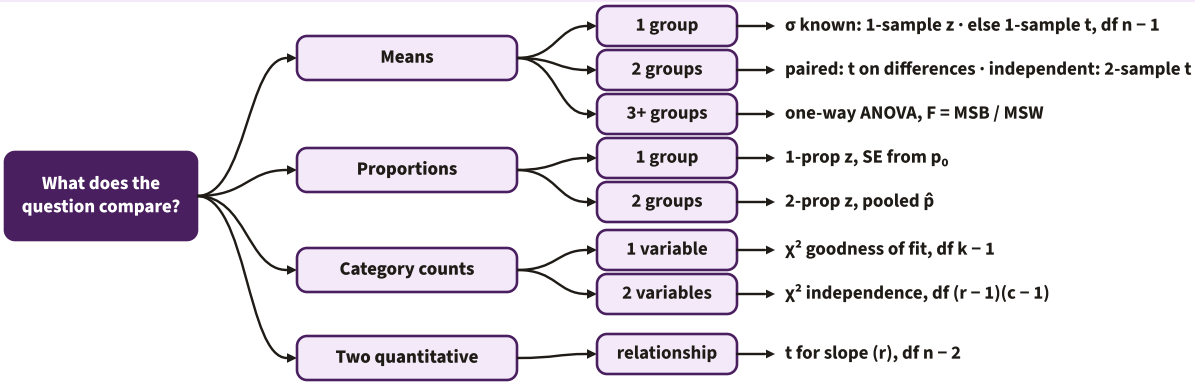
- **Correlation is not causation**: a lurking variable may drive both.
- **Extrapolation**: $x = 6$ gives $\hat{y} = 90.1$ but lies past the data; do not trust it.
- Check for curved patterns and influential points.

Test for the slope

H_0: $\beta = 0$ (no linear relationship). $t = \frac{b}{SE_b}$, $df = n - 2$.

12 WHICH TEST? · Z AND T TABLES

start at the left, follow the question

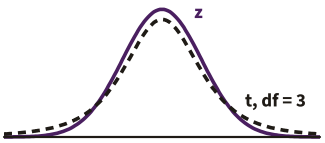


z table (computed): Φ(z) = area left of z

z	Φ(z)	z	Φ(z)	z	Φ(z)
0.0	0.5000	1.0	0.8413	2.0	0.9772
0.1	0.5398	1.1	0.8643	2.1	0.9821
0.2	0.5793	1.2	0.8849	2.2	0.9861
0.3	0.6179	1.3	0.9032	2.3	0.9893
0.4	0.6554	1.4	0.9192	2.4	0.9918
0.5	0.6915	1.5	0.9332	2.5	0.9938
0.6	0.7257	1.6	0.9452	2.6	0.9953
0.7	0.7580	1.7	0.9554	2.7	0.9965
0.8	0.7881	1.8	0.9641	2.8	0.9974
0.9	0.8159	1.9	0.9713	2.9	0.9981

t* table (computed): two-sided confidence; ∞ row = z*

df	90%	95%	98%	99%
1	6.314	12.706	31.821	63.657
2	2.920	4.303	6.965	9.925
3	2.353	3.182	4.541	5.841
4	2.132	2.776	3.747	4.604
5	2.015	2.571	3.365	4.032
6	1.943	2.447	3.143	3.707
7	1.895	2.365	2.998	3.499
8	1.860	2.306	2.896	3.355
9	1.833	2.262	2.821	3.250
10	1.812	2.228	2.764	3.169
12	1.782	2.179	2.681	3.055
15	1.753	2.131	2.602	2.947
20	1.725	2.086	2.528	2.845
30	1.697	2.042	2.457	2.750
60	1.671	2.000	2.390	2.660
∞ (z)	1.645	1.960	2.326	2.576



t curves are flatter with heavier tails; they approach z as df grows.

Conditions for every inference

- **Random**: a random sample or random assignment.
- **Independent**: $n \le 10\%$ of the population.
- **Normal/large**: $n \ge 30$, or a roughly symmetric sample; np, $n(1-p) \ge 10$; expected counts ≥ 5.

Interval or test?

- "How big is it?" → **confidence interval**.
- "Is there an effect?" → **test**.
- Paired data (same subject twice) → analyze the **differences**.

MEMORY TRICK

Paired is one sample in disguise

Subtract within each pair first; then it is a one-sample t on the differences.