

MathCastle

Statistics & Data

Complete course booklet • 11 topics

Contents

1	Data and Displays	3
1.1	Categorical and Quantitative	3
1.2	Histograms and Shape	3
1.3	Why It Works	4
2	Center and Spread	4
2.1	Measures of Center	5
2.2	Measures of Spread	5
2.3	Why It Works	6
3	Percentiles and Box Plots	6
3.1	Quartiles and the Five-Number Summary	6
3.2	Percentiles	7
4	The Normal Distribution	7
4.1	Z-Scores	8
4.2	The Empirical Rule and the Table	8
4.3	Why It Works	9
5	Two Variables	9
5.1	Scatter Plots and Correlation	9
5.2	The Least-Squares Line	10
6	Sampling and Experiments	11
6.1	Samples and Bias	11
6.2	Experiments versus Observation	12
7	Random Variables	13
7.1	Discrete Random Variables	13
7.2	Combining Random Variables	13

7.3	Why It Works	14
8	Binomial and Geometric	14
8.1	The Binomial Distribution	14
8.2	The Geometric Distribution	15
9	Sampling Distributions	15
9.1	The Sampling Distribution of a Mean	16
9.2	The Sampling Distribution of a Proportion	16
9.3	Why It Works	17
10	Confidence Intervals	17
10.1	Intervals for a Proportion	17
10.2	Intervals for a Mean	18
11	Hypothesis Tests	18
11.1	The Logic of a Test	18
11.2	Tests for Means and Two Samples	19
11.3	Why It Works	20

1 Data and Displays

Kinds of data and the pictures that fit them

1.1 Categorical and Quantitative

Two kinds of data

A variable is **categorical** when it sorts things into groups (eye color, favorite sport) and **quantitative** when it measures an amount (height, test score). The kind of data decides the kind of picture: bar charts and pie charts for categories; dot plots, histograms and box plots for numbers.

The distinction matters because arithmetic only makes sense for quantitative data. The "average zip code" of a class is a number you can compute and it means nothing. Before summarizing anything, ask: would adding two of these values together make sense? If not, the data is categorical, and counts and percentages are the only honest summaries.

Choosing a display

A survey records each student's pet type and the number of hours of sleep. Pet type is categorical: a bar chart of counts. Hours of sleep is quantitative: a dot plot or histogram, which shows the shape of the distribution.

Where it goes wrong

Treating a number that is really a label as quantitative — jersey numbers, ID numbers, zip codes — or treating a rating like "1 to 5 stars" as if the gaps between stars were equal amounts.

Try it: classify

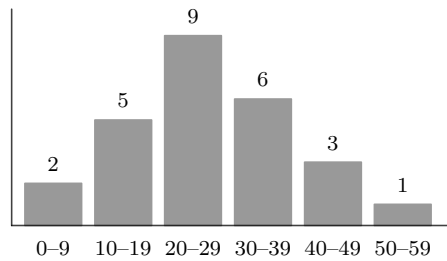
Classify each as categorical or quantitative: (a) blood type, (b) number of siblings, (c) commute time, (d) car color.

Answer. (a) categorical, (b) quantitative (a count), (c) quantitative, (d) categorical.

1.2 Histograms and Shape

Reading a histogram

A histogram groups values into equal-width **bins** and draws a bar whose height is the count in each bin. Look for the **shape**: symmetric, skewed right (a long tail to the right), or skewed left; one peak or two; any gaps or outliers.



Shape is the first thing to report because it decides which summaries are meaningful later. A skewed distribution — incomes, house prices, waiting times — has a mean pulled toward the tail and a median that stays put, so the two disagree, and the direction of the disagreement tells you the direction of the skew.

Skew points to the tail

Skewed **right** means the long tail is on the right (a few large values). For right-skewed data, $\text{mean} > \text{median}$. For left-skewed data, $\text{mean} < \text{median}$.

Bin widths change the picture

The same 40 values shown with bins of width 5 look jagged; with width 20, two bars hide everything. Choose bins so that the shape is visible without being noise — usually 5 to 15 bins.

Try it: read the shape

A histogram of household incomes has most bars on the left and a long thin tail to the right. Which is larger, the mean or the median?

Answer. Right-skewed: the tail of very high incomes pulls the **mean** above the median.

1.3 Why It Works

The mean is the balance point of the data: it is the unique value m for which the deviations $x_i - m$ sum to zero. A single far-out value contributes a huge deviation, so the balance point must slide toward it to compensate. The median, defined by position alone, does not feel how far away the extreme value is — only that it is on one side. That is the whole reason the two disagree exactly when a distribution is skewed.

2 Center and Spread

Mean, median, variance and standard deviation

2.1 Measures of Center

Mean and median

The **mean** $\bar{x} = \frac{\sum x_i}{n}$ is the fair-share average. The **median** is the middle value of the sorted list (the average of the two middle values when n is even). The **mode** is the most frequent value.

Use the mean when the distribution is roughly symmetric and you want every value to count; use the median when the data is skewed or has outliers and you want a typical value. Newspapers report median house prices for exactly this reason: one mansion should not change what "typical" means.

Compute both

Data: 4, 8, 15, 16, 23, 42. Mean = $108/6 = 18$. Median = average of 15 and 16 = 15.5. The mean is larger because 42 pulls it up.

Try it: which summary?

Nine employees earn (in thousands) 38, 40, 41, 42, 45, 47, 50, 52, 300. Find the mean and median, and say which describes a typical salary.

Answer. Total 655, mean ≈ 72.8 ; median (5th value) = 45. The **median** — the mean is inflated by the one 300.

2.2 Measures of Spread

Range, IQR, standard deviation

The **range** is max minus min. The **interquartile range** $IQR = Q_3 - Q_1$ is the spread of the middle half. The **standard deviation** $s = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n - 1}}$ is the typical distance of a value from the mean.

Standard deviation is built from squared deviations, so it is in the same units as the data once the square root is taken, and it weights far-away values heavily. That makes it the natural partner of the mean: both are sensitive to outliers. The IQR is the partner of the median: both depend only on positions in the sorted list.

Compute s

Data: 2, 4, 4, 4, 5, 5, 7, 9; mean 5. Squared deviations: 9, 1, 1, 1, 0, 0, 4, 16, sum 32. Sample variance $32/7 \approx 4.57$, $s \approx 2.14$. (The population version divides by 8: $\sigma = 2$.)

Where it goes wrong

Forgetting to square the deviations (they would sum to zero), or dividing by n when the data is a sample — the $n - 1$ corrects for the fact that the sample mean sits closer to the sample than the true mean does.

Shifting and scaling

Adding c to every value shifts the mean by c and leaves the spread unchanged. Multiplying every value by k multiplies the mean by k and the standard deviation by $|k|$.

Try it: rescale

Temperatures in Celsius have mean 20 and standard deviation 5. Convert to Fahrenheit, $F = 1.8C + 32$: what are the new mean and standard deviation?

Answer. Mean $1.8 \cdot 20 + 32 = \mathbf{68}$; standard deviation $1.8 \cdot 5 = \mathbf{9}$ (the $+32$ does not spread anything).

2.3 Why It Works

The deviations from the sample mean are not free: they must sum to zero, so only $n - 1$ of them carry information — once $n - 1$ are known, the last is forced. Dividing by $n - 1$ instead of n makes the sample variance an *unbiased* estimate of the population variance: averaged over all possible samples, it hits the true value exactly. Dividing by n would systematically underestimate, because the sample mean is always at least as close to the sample points as the true mean is.

3 Percentiles and Box Plots

The five-number summary and outliers

3.1 Quartiles and the Five-Number Summary

Five numbers

The **five-number summary** is minimum, Q_1 , median, Q_3 , maximum. Q_1 is the median of the lower half and Q_3 the median of the upper half. A **box plot** draws a box from Q_1 to Q_3 with a line at the median and whiskers out to the extremes (or to the last non-outlier).

A box plot throws away most of the data on purpose. What survives is exactly what you need to compare groups side by side: where the middle is, how wide the middle half is, and whether the tails are lopsided. Two box plots on the same axis answer "which group is higher, and which is more consistent?" at a glance.

Build one

Data (sorted): 3, 5, 7, 8, 10, 12, 14, 18, 30. Median 10. Lower half 3, 5, 7, 8: $Q_1 = 6$. Upper half 12, 14, 18, 30: $Q_3 = 16$. Summary: 3, 6, 10, 16, 30. IQR = 10.

The 1.5 IQR rule

A value is an **outlier** if it lies below $Q_1 - 1.5 \text{ IQR}$ or above $Q_3 + 1.5 \text{ IQR}$. Outliers are plotted as separate dots beyond the whiskers.

Try it: outlier?

For the data above ($Q_1 = 6$, $Q_3 = 16$), is 30 an outlier?

Answer. Upper fence $16 + 1.5 \cdot 10 = 31$. Since $30 < 31$, it is **not** an outlier (barely).

3.2 Percentiles

The p th percentile

The p th percentile is the value below which $p\%$ of the data falls. The median is the 50th percentile; Q_1 and Q_3 are the 25th and 75th. A score at the 90th percentile beat 90% of the scores — it says nothing about the score itself.

Percentile of a value

In a class of 40, a student has the 34th highest score; 6 scored higher. She is above 33 of 40: about the 83rd percentile.

Where it goes wrong

Reading "90th percentile" as "scored 90%". Percentiles rank; percents measure.

Try it: from the summary

A box plot shows $Q_1 = 52$, median 60, $Q_3 = 71$. What percent of the data lies between 52 and 71? Above 71?

Answer. Between Q_1 and Q_3 : **50%**. Above Q_3 : **25%**.

4 The Normal Distribution

Z-scores, the empirical rule and the standard normal table

4.1 Z-Scores

Standardizing

A **z-score** says how many standard deviations a value sits from the mean: $z = \frac{x - \mu}{\sigma}$. It lets you compare values from different scales — an SAT score and an ACT score — on one ruler.

A z-score of 2 is unusual in any roughly bell-shaped data set, whatever the units, because the standard deviation has already absorbed the scale. Negative z means below the mean. Converting back is the same formula solved for x : $x = \mu + z\sigma$.

Compare two scores

SAT: mean 1050, SD 200, score 1350: $z = 1.5$. ACT: mean 21, SD 5, score 29: $z = 1.6$. The ACT score is (slightly) the stronger performance.

Try it: back to raw

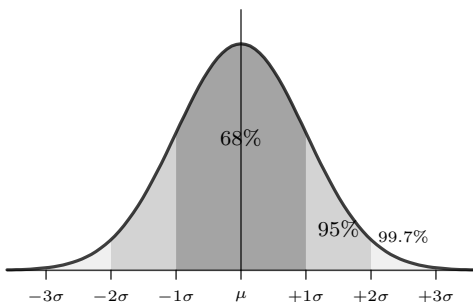
Heights have mean 170 cm and SD 8 cm. What height has $z = -1.25$?

Answer. $170 + (-1.25)(8) = 160$ cm.

4.2 The Empirical Rule and the Table

68–95–99.7

For a **normal** (bell-shaped) distribution, about 68% of values lie within 1 SD of the mean, 95% within 2, and 99.7% within 3. For other cutoffs, the standard normal table (or a calculator) gives the area to the left of any z .



The rule is a rounding of table values: $P(-1 < Z < 1) = 0.6827$, $P(-2 < Z < 2) = 0.9545$, $P(-3 < Z < 3) = 0.9973$. Because the curve is symmetric, every "between" question is a subtraction of two left-tail areas, and every "above" question is one minus a left-tail area.

Above a cutoff

IQ scores: mean 100, SD 15. Fraction above 130: $z = 2$, and $P(Z > 2) = \frac{1 - 0.95}{2} = 0.025$, about 2.5%.

Between two values

Fraction between 85 and 115: z from -1 to 1 : 68%.

Where it goes wrong

Using the table area (which is always the area to the LEFT) as if it were the area to the right, or applying the empirical rule to data that is clearly skewed.

Try it: percentile from z

A value has $z = 1$. Roughly what percentile is it?

Answer. Left of $z = 1$: $50\% + 34\% = 84\text{th}$ percentile.

4.3 Why It Works

Add up many small independent effects — measurement errors, genetic contributions, rounding — and the total tends toward a normal distribution no matter what the individual pieces look like. This is the central limit theorem, and it is why heights, test scores and averages of anything are bell-shaped while single coin flips are not. The density $\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2}$ is the unique shape that is preserved under this adding-up, which is why the same table serves every normal distribution once you standardize.

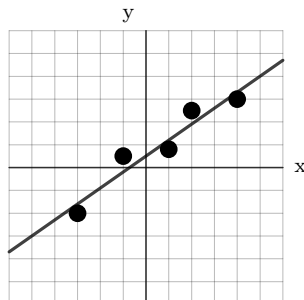
5 Two Variables

Scatter plots, correlation and the least-squares line

5.1 Scatter Plots and Correlation

Correlation r

A scatter plot shows pairs (x, y) . The **correlation coefficient** r measures the strength and direction of a LINEAR relationship: $-1 \leq r \leq 1$, with r near ± 1 for tight lines and near 0 for no linear pattern. r has no units and does not change if either variable is rescaled.



Correlation is the average product of the two z-scores: $r = \frac{1}{n-1} \sum z_x z_y$. When x is above its mean, is y usually above its mean too? If so, the products are mostly positive and r is positive. That formula also explains what r misses: a perfect U-shaped relationship has $r \approx 0$, because above-average x pairs with above-average y on one side and below-average y on the other.

Correlation is not causation

A strong r says the variables move together. It does not say one causes the other: a lurking variable may drive both (ice-cream sales and drownings both rise in summer).

Estimate r

Points tightly along a line rising to the right: $r \approx 0.9$. A loose cloud falling to the right: $r \approx -0.4$. A circle of points: $r \approx 0$.

Try it: what happens to r ?

Heights (cm) and weights (kg) have $r = 0.6$. If heights are converted to inches, what is r ?

Answer. Unchanged: **0.6**. Rescaling does not change z-scores.

5.2 The Least-Squares Line

Regression line

The **least-squares line** $\hat{y} = a + bx$ minimizes the sum of squared vertical residuals $y - \hat{y}$. Its slope is $b = r \frac{s_y}{s_x}$ and it always passes through $(\bar{x}, \bar{y})$, so $a = \bar{y} - b\bar{x}$. The slope means: for each unit increase in x , $\hat{y}$ changes by b on average.

The line predicts, it does not explain. Its slope is r times the ratio of spreads, which says something surprising: for $r < 1$, the predicted y is always fewer standard deviations from its mean than x is from its mean. Tall parents have children who are tall, but on average less tall than the parents — "regression toward the mean", which gave the whole method its name.

Build the line

$\bar{x} = 10$, $s_x = 2$, $\bar{y} = 50$, $s_y = 8$, $r = 0.75$. Slope $b = 0.75 \cdot 8/2 = 3$. Intercept $a = 50 - 3 \cdot 10 = 20$.
Line: $\hat{y} = 20 + 3x$. Predicted y at $x = 12$: 56.

r^2 and residuals

r^2 is the fraction of the variation in y explained by the line. A **residual** is actual minus predicted; a curved pattern in the residual plot means a line was the wrong model.

Where it goes wrong

Extrapolating far outside the range of the data, and reading $r^2 = 0.64$ as "64% of the points are on the line".

Try it: residual

Using $\hat{y} = 20 + 3x$, a data point is $(12, 60)$. Find the residual.

Answer. Predicted 56; residual $60 - 56 = 4$ (the point is above the line).

6 Sampling and Experiments

Where data comes from and what can go wrong

6.1 Samples and Bias

Population, sample, bias

A **population** is everyone you want to describe; a **sample** is the part you actually measure. A **simple random sample** gives every group of n individuals the same chance of being chosen. **Bias** is a systematic tilt: voluntary response, convenience sampling, undercoverage, non-response, and leading questions all produce it.

Randomness is not sloppiness; it is the only known defense against the biases you cannot think of. A sample chosen by a person — even a careful one — reflects that person's reach and habits. A sample chosen by a random number generator reflects nothing but the population. Larger samples reduce random error; they do nothing at all about bias.

Spot the bias

A website asks visitors to rate the site. Only people who feel strongly respond: voluntary response bias. Surveying shoppers at a mall about online shopping: undercoverage of people

who never visit malls.

Stratified and cluster sampling

A **stratified** sample takes a random sample within each group (stratum) so every group is represented. A **cluster** sample randomly picks whole groups (classrooms, blocks) and measures everyone in them.

Try it: name the design

A school randomly picks 5 homerooms and surveys every student in them. Stratified or cluster?

Answer. Cluster: whole groups are chosen at random.

6.2 Experiments versus Observation

Cause needs an experiment

An **observational study** records what happens; an **experiment** assigns treatments. Only an experiment with **random assignment** can establish cause, because randomization balances every lurking variable across the groups, seen and unseen. A **control group**, **blinding** and **replication** make the comparison fair.

The logic is the same as sampling, turned inward: random assignment makes the treatment groups alike in everything except the treatment, so any difference too large to be chance must come from the treatment. Without randomization, the people who chose the treatment differ from those who did not, and the difference could be them rather than it.

Design it

To test whether a study app raises scores: randomly split 200 volunteers into two groups of 100; one uses the app, one does not; compare mean scores. Blocking by prior grade reduces noise.

Where it goes wrong

Concluding cause from an observational study ("students who use the app score higher") — the app users may simply be the more motivated students.

Try it: what is the confound?

Cities with more police have more crime. Does hiring police cause crime?

Answer. No — a lurking variable: **city size** (and crime rate) drives both police hiring and crime counts.

7 Random Variables

Expected value, variance and combining variables

7.1 Discrete Random Variables

Expected value and variance

A **random variable** X assigns a number to each outcome; its **probability distribution** lists each value with its probability (summing to 1). $E[X] = \mu = \sum x p(x)$ is the long-run average; $\text{Var}(X) = \sum (x - \mu)^2 p(x)$ and $\sigma = \sqrt{\text{Var}(X)}$.

Expected value is a weighted mean with probabilities as weights, so everything you know about means transfers: it is linear, it can be a value X never takes, and it is what the average of many repetitions settles toward. Variance measures how far from that average a single trial typically lands.

A raffle

A \$2 ticket wins \$100 with probability 0.01 and nothing otherwise. Net winnings X : 98 w.p. 0.01, -2 w.p. 0.99. $E[X] = 0.98 - 1.98 = -1$: you lose a dollar per ticket on average.

Try it: expected value

Roll a fair die; you win the number shown in dollars. Find $E[X]$ and $\text{Var}(X)$.

Answer. $E[X] = 3.5$. $E[X^2] = 91/6$, so $\text{Var} = 91/6 - 12.25 = \mathbf{35/12} \approx 2.92$.

7.2 Combining Random Variables

Linear combinations

For any random variables, $E[aX + bY] = aE[X] + bE[Y]$. For **independent** X and Y , $\text{Var}(aX + bY) = a^2\text{Var}(X) + b^2\text{Var}(Y)$ — variances add even when you subtract, because b^2 is positive.

Standard deviations never add; variances do, and only for independent variables. The difference of two independent measurements is MORE variable than either one, not less, which is why comparing two groups needs bigger samples than measuring one.

Total weight

Boxes have mean 10 kg, SD 1 kg, independently. Five boxes: mean 50, variance 5, SD

$\sqrt{5} \approx 2.24$ — not 5.

Where it goes wrong

Writing $\sigma_{X-Y} = \sigma_X - \sigma_Y$. It is $\sqrt{\sigma_X^2 + \sigma_Y^2}$ for independent variables.

Try it: a difference

X has SD 3 and Y has SD 4, independent. Find the SD of $X - Y$.

Answer. $\sqrt{9 + 16} = 5$.

7.3 Why It Works

For independent X, Y with means μ_X, μ_Y : $\text{Var}(X + Y) = E[(X - \mu_X + Y - \mu_Y)^2] = \text{Var}(X) + \text{Var}(Y) + 2E[(X - \mu_X)(Y - \mu_Y)]$. Independence makes the cross term the product of two expectations, each zero. Without independence the cross term is $2\text{Cov}(X, Y)$, which is where correlation enters — and why the formula fails for, say, a value and itself: $\text{Var}(X + X) = 4\text{Var}(X)$, not $2\text{Var}(X)$.

8 Binomial and Geometric

Counting successes and waiting for the first one

8.1 The Binomial Distribution

Binomial setting

Fixed number n of independent trials, each a success with the same probability p . The number of successes X is **binomial**: $P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}$, with mean np and standard deviation $\sqrt{np(1-p)}$.

The formula is a counting argument: any particular sequence with k successes has probability $p^k(1-p)^{n-k}$, and there are $\binom{n}{k}$ such sequences. The mean np is linearity of expectation over n indicator variables; the variance $np(1-p)$ is n copies of a single trial's variance $p(1-p)$, added because the trials are independent.

Exactly and at least

A free-throw shooter makes 70%. In 5 shots, $P(\text{exactly } 3) = \binom{5}{3} (0.7)^3 (0.3)^2 = 10 \cdot 0.343 \cdot 0.09 = 0.3087$. $P(\text{at least } 4) = \binom{5}{4} (0.7)^4 (0.3) + (0.7)^5 = 0.36015 + 0.16807 = 0.52822$.

Normal approximation

When $np \geq 10$ and $n(1 - p) \geq 10$, the binomial is close to normal with the same mean and SD, so z-scores answer "at least k " questions without summing.

Try it: mean and SD

A fair coin is tossed 100 times. Find the mean and standard deviation of the number of heads, and the approximate probability of 60 or more heads.

Answer. Mean 50, SD 5. $z = 2$: about **2.5%**.

8.2 The Geometric Distribution

Waiting for the first success

Repeat independent trials with success probability p until the first success. The trial number X is **geometric**: $P(X = k) = (1 - p)^{k-1}p$, $E[X] = \frac{1}{p}$, and $P(X > k) = (1 - p)^k$.

The expected wait $1/p$ is the fact everyone should carry around: a 1-in-20 event takes 20 tries on average. The memoryless property is the surprising one — having failed 10 times does not make success "due"; the remaining wait is still $1/p$ on average, because the trials are independent.

Rolling a six

Expected number of rolls to get a six: 6. Probability the first six comes on roll 3: $(5/6)^2(1/6) \approx 0.116$. Probability it takes more than 3 rolls: $(5/6)^3 \approx 0.579$.

Where it goes wrong

Confusing binomial ("how many successes in n trials") with geometric ("how many trials until the first success"), and the gambler's fallacy that a long run of failures raises the chance of success.

Try it: geometric

A 20% free-sample offer: what is the expected number of customers until the first taker, and the probability the first taker is the 2nd customer?

Answer. Expected $1/0.2 = 5$. $P(X = 2) = 0.8 \cdot 0.2 = 0.16$.

9 Sampling Distributions

How a statistic varies from sample to sample

9.1 The Sampling Distribution of a Mean

Means of samples

Take a random sample of size n from a population with mean μ and standard deviation σ . The sample mean $\bar{x}$ is itself random: its mean is μ and its standard deviation is $\frac{\sigma}{\sqrt{n}}$. By the **central limit theorem**, for large n (about 30 or more) $\bar{x}$ is approximately normal whatever the population looks like.

The $\sqrt{n}$ is the most consequential square root in statistics. Averaging four measurements halves the noise; averaging a hundred cuts it by ten. To halve the error again you must quadruple the sample — which is why precision is expensive and why polls of a thousand people are about as good as it gets for the price.

Probability for a mean

Bottles hold mean 500 mL, SD 10 mL. For a sample of 25, $\bar{x}$ has SD $10/5 = 2$. $P(\bar{x} < 496) = P(Z < -2) \approx 0.023$.

Try it: sample size

Population SD is 12. How large a sample makes the SD of $\bar{x}$ equal to 2?

Answer. $12/\sqrt{n} = 2$, so $\sqrt{n} = 6$, $n = \mathbf{36}$.

9.2 The Sampling Distribution of a Proportion

Proportions

A sample proportion $\hat{p}$ of successes has mean p and standard deviation $\sqrt{\frac{p(1-p)}{n}}$, approximately normal when $np \geq 10$ and $n(1-p) \geq 10$. It is the binomial count divided by n .

A poll

If 52% of voters support a measure and $n = 400$: SD of $\hat{p}$ is $\sqrt{0.52 \cdot 0.48/400} \approx 0.025$. A poll showing 47% would be 2 SDs low — unlikely if the true value is 52%.

Where it goes wrong

Using σ instead of $\sigma/\sqrt{n}$ for a sample mean, or applying the normal approximation to a proportion when $np < 10$.

Try it: proportion

True proportion $p = 0.3$, sample size $n = 210$. Find the SD of $\hat{p}$.

Answer. $\sqrt{0.3 \cdot 0.7 / 210} = \sqrt{0.001} \approx \mathbf{0.0316}$.

9.3 Why It Works

The sample mean is $\bar{x} = \frac{X_1 + \cdots + X_n}{n}$. Variances of independent terms add: $\text{Var}(X_1 + \cdots + X_n) = n\sigma^2$. Dividing a variable by n divides its variance by n^2 : $\text{Var}(\bar{x}) = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}$. Taking the square root gives $\sigma/\sqrt{n}$. The central limit theorem is a deeper fact — the SHAPE becomes normal — but the size of the spread is just this arithmetic.

10 Confidence Intervals

Estimating with a margin of error

10.1 Intervals for a Proportion

Estimate $\pm$ margin

A **confidence interval** is estimate $\pm z^* \cdot \text{SE}$. For a proportion: $\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1 - \hat{p})}{n}}$, with $z^* = 1.96$ for 95% confidence (1.645 for 90%, 2.576 for 99%). "95% confident" means the method captures the true value in 95% of all samples.

The interval is a statement about the procedure, not about this one answer. The true proportion is a fixed number; it is either in the interval or not. What is 95% is the fraction of intervals, over many repetitions of the sampling, that would contain it. Widening the interval (higher confidence) or enlarging the sample (smaller SE) are the only two ways to buy more certainty.

A poll interval

$\hat{p} = 0.52$, $n = 400$: $\text{SE} = \sqrt{0.52 \cdot 0.48 / 400} \approx 0.025$. 95% interval: $0.52 \pm 1.96 \cdot 0.025 = 0.52 \pm 0.049$, i.e. (0.471, 0.569).

Sample size for a margin

To get margin of error m at 95%: $n = \left(\frac{1.96}{m}\right)^2 \hat{p}(1 - \hat{p})$; with no guess, use $\hat{p} = 0.5$ (the worst case): $n \approx \frac{0.96}{m^2}$.

Try it: size

How many people must be polled for a $\pm 3\%$ margin at 95% confidence (no prior guess)?

Answer. $n = (1.96/0.03)^2 \cdot 0.25 \approx 1067.1$, so **1068**.

10.2 Intervals for a Mean

The t interval

For a mean with unknown σ : $\bar{x} \pm t^* \frac{s}{\sqrt{n}}$, where t^* comes from the t distribution with $n - 1$ degrees of freedom. For large n , $t^* \approx z^*$; for small n it is larger, paying for the extra uncertainty in s .

A small sample

$n = 16$, $\bar{x} = 30$, $s = 4$, 95%: $t_{15}^* = 2.131$. Interval $30 \pm 2.131 \cdot 4/4 = 30 \pm 2.13$, i.e. (27.87, 32.13).

Where it goes wrong

Saying "there is a 95% probability the true mean is in this interval" (the interval is what is random), or using z^* with a tiny sample.

Try it: interpret

A 95% interval for mean commute time is (28, 34) minutes. Which is correct: (a) 95% of commutes take 28–34 minutes; (b) we are 95% confident the MEAN commute is between 28 and 34?

Answer. (b). The interval is about the mean, not individual commutes.

11 Hypothesis Tests

Is the effect real, or could chance explain it?

11.1 The Logic of a Test

Null, alternative, p -value

A test starts from the **null hypothesis** H_0 (nothing is going on: $p = 0.5$, $\mu = 100$) and an **alternative** H_a . Compute the test statistic $z = \frac{\text{estimate} - \text{null value}}{\text{SE}}$. The **p-value** is the

probability, IF H_0 were true, of a result at least this extreme. Small p (below α , usually 0.05) is evidence against H_0 .

A test is a proof by contradiction with probability in place of certainty. Assume nothing is going on; ask how surprising the data would then be; if very surprising, doubt the assumption. What a p -value does NOT measure is the probability that H_0 is true — it conditions on H_0 , it does not evaluate it. And a large p -value never proves H_0 ; it only says the data are compatible with it.

A coin

A coin lands heads 60 times in 100 flips. $H_0: p = 0.5$. $SE = \sqrt{0.25/100} = 0.05$; $z = (0.6 - 0.5)/0.05 = 2$. Two-sided p -value $= 2P(Z > 2) \approx 0.046$. At $\alpha = 0.05$, reject H_0 : the coin looks biased.

Errors

A **Type I** error rejects a true H_0 (probability α). A **Type II** error fails to reject a false H_0 . **Power** is the probability of correctly rejecting; it grows with sample size and effect size.

Try it: one-sided

For the same coin, test $H_a: p > 0.5$. What is the p -value?

Answer. One tail: $P(Z > 2) \approx 0.023$.

11.2 Tests for Means and Two Samples

The t test

For a mean: $t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}$ with $n - 1$ degrees of freedom. To compare two independent groups:
 $t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}$. Paired data (before/after on the same subjects) uses a one-sample t test on the differences.

Two groups

Group A: $\bar{x} = 78$, $s = 8$, $n = 32$. Group B: $\bar{x} = 74$, $s = 6$, $n = 32$. $SE = \sqrt{64/32 + 36/32} = \sqrt{3.125} \approx 1.77$; $t \approx 2.26$. With about 58 df, the two-sided p -value is about 0.03: a real difference at the 5% level.

Where it goes wrong

Treating "statistically significant" as "large" or "important" — with a huge sample, a trivial difference has a tiny p -value. Report the effect size alongside the test.

Try it: statistic

A sample of $n = 25$ has $\bar{x} = 52$, $s = 10$; test $H_0: \mu = 50$. Compute t .

Answer. $t = (52 - 50)/(10/5) = 1$ — not significant (the p -value is about 0.33).

11.3 Why It Works

If H_0 is true, the test statistic follows a known distribution, so we can compute exactly how often results as extreme as ours would occur by chance alone. When that frequency is 2%, one of two things is true: either H_0 holds and we witnessed a 1-in-50 fluke, or H_0 is false. The rarer the fluke would have to be, the less comfortable the first explanation becomes. Fixing α in advance keeps us honest: it is a promise about how often we are willing to be fooled by chance, made before seeing the data.