Course Companion

AP Statistics

Every topic handout in one volume

全部专题讲义合集

exercises.lol

Contents 目录

Exploring One-Variable Data

AP Statistics

Introducing Statistics: What Can We Learn from Data?

Statistics 统计学 is the science of learning from **data** 数据 – numbers or labels collected from the real world. Data vary, so we describe patterns and account for the **variation** 变异 rather than expecting every value to match. A statistical question anticipates an answer based on data that vary.

Two distinctions run through the whole course. A **parameter** 参数 is a numerical summary of a whole **population**; a **statistic** 统计量 is a numerical summary of a **sample** - we use the statistic to estimate the parameter we cannot measure directly. And **descriptive statistics** 描述统计 only summarise the data set in hand, while **inferential statistics** 推断统计 use a sample to make and test claims about the larger population.

The Language of Variation: Variables

A **variable** 变量 is a characteristic that can differ between individuals. Two kinds:

- **Categorical** 分类 (qualitative): values are labels/groups (eye colour, brand).
- **Quantitative** 定量: values are numbers you can do arithmetic on (height, age). Quantitative variables are **discrete** (countable) or **continuous** (measured).

Choosing the right graph and summary depends on which kind you have.

Representing a Categorical Variable with Tables

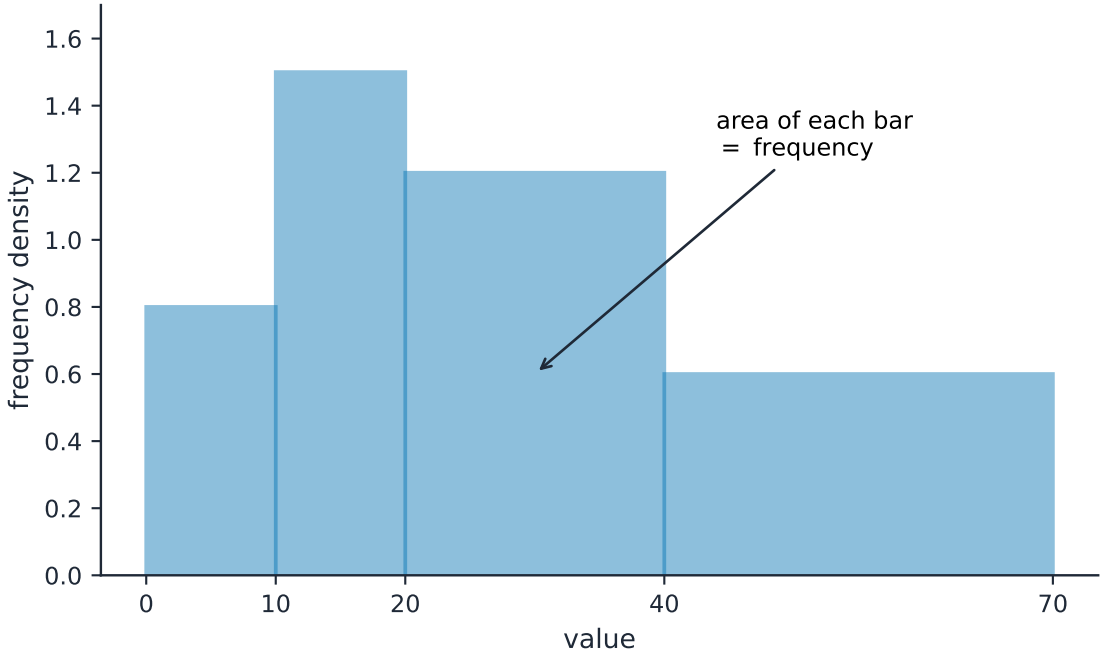
A **frequency table** 频数表 lists each category's **count** (frequency); a **relative frequency** 相对频率 table lists each category's **proportion** 比例 (count ÷ total). Relative frequencies let you compare groups of different sizes fairly.

Representing a Categorical Variable with Graphs

Bar charts 条形图 show the count or proportion of each category as separated bars; a **pie chart** shows each category's share of the whole. The bar heights (or slices) let you compare categories at a glance. Bars may be ordered by size or by a natural category order.

Representing a Quantitative Variable with Graphs

For numbers, use a **dotplot** 点图, **stem-and-leaf plot** 茎叶图, or **histogram** 直方图 (bars over value intervals called bins). These show the **distribution** 分布 – how the values spread out. A histogram's bin width changes the picture, so choose it to reveal the shape.



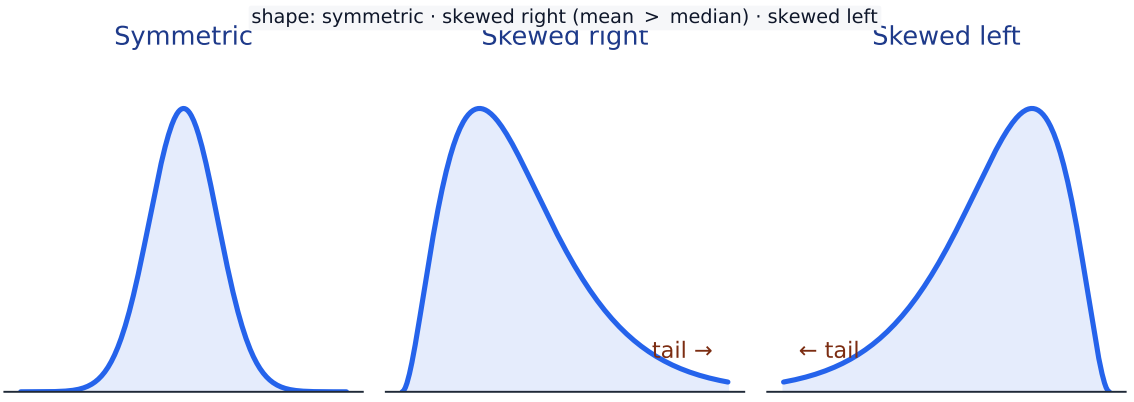
On a histogram with unequal class widths the bar area is the frequency

Describing the Distribution of a Quantitative Variable

Describe four things (remember **SOCS**):

- **Shape** 形状: symmetric, or **skewed** 偏斜 left/right (a long tail on that side), and how many peaks - one main peak is **unimodal** 单峰, two prominent peaks **bimodal** 双峰, and roughly equal bars **uniform** 均匀.
- **Outliers** 离群值: unusual values far from the rest.
- **Center**: a typical value (mean or median).
- **Spread**: how much the values vary (range, IQR, standard deviation).

Always describe shape/center/spread **in context**, with units.



The shape of a distribution: symmetric, skewed right (long right tail), or skewed left

Summary Statistics for a Quantitative Variable

- **Center:** the **mean** 均值 $\bar{x} = \dfrac{\sum x_i}{n}$ (average) and the **median** 中位数 (middle value). The median resists outliers; the mean is pulled toward a skew.
- **Spread:** the **range**, the **interquartile range** 四分位距 $\text{IQR} = Q_3 - Q_1$ (middle 50%), and the **standard deviation** 标准差 $s_x = \sqrt{\dfrac{\sum (x_i - \bar{x})^2}{n-1}}$ (typical distance from the mean; its square is the **variance** 方差).
- The **five-number summary** 五数概括: min, Q_1, median, Q_3, max.

Use **resistant** measures (median, IQR) for skewed data; mean and standard deviation for roughly symmetric data.

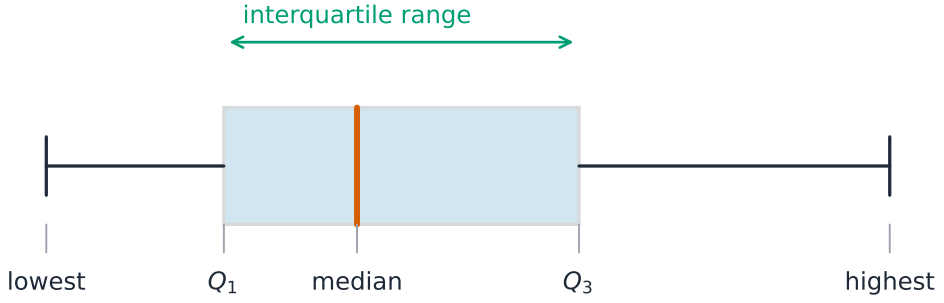
The **percentile** 百分位数 of a value is the percent of the data at or below it – so the median is the 50th percentile and Q_1 the 25th. A **cumulative relative frequency graph** 累积相对频率图 makes percentiles easy to read: for each value it plots the proportion of the data **at or below** it, rising from 0 to 1. Go up from a value to the curve and across to its percentile, or reverse the steps to find the value at a given percentile (the same reading works from a cumulative-frequency table).

Worked example. For the data $4, 8, 6, 10, 7$: the mean is $\bar{x} = \dfrac{4+8+6+10+7}{5} = \dfrac{35}{5} = 7$. Sorting to $4, 6, 7, 8, 10$, the median is the middle value, 7. The mean and median agree here because the data are roughly symmetric.

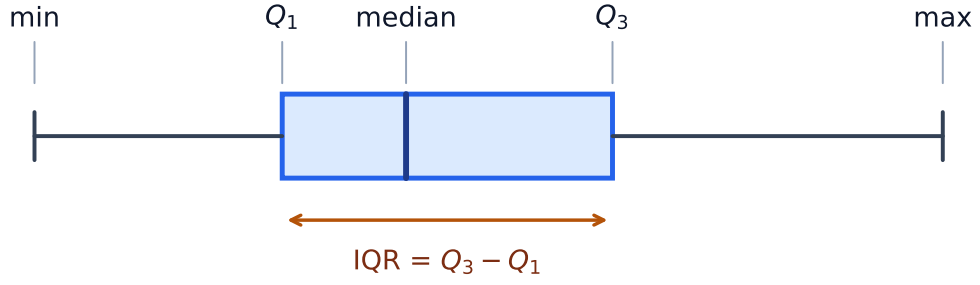
Graphical Representations of Summary Statistics

A **boxplot** 箱线图 draws the five-number summary: a box from Q_1 to Q_3 with the median inside, and whiskers to the most extreme non-outlier values. A point is an **outlier** if it lies more than $1.5 \times \text{IQR}$ beyond a quartile – a rule you may be asked to apply. Boxplots are ideal for comparing several groups side by side.

Worked example. A dataset has $Q_1 = 20$ and $Q_3 = 32$, so IQR $= 12$. The outlier fences are $Q_1 - 1.5(12) = 2$ and $Q_3 + 1.5(12) = 50$. Any value below 2 or above 50 is flagged as an outlier.



A box-and-whisker plot shows the quartiles and the range



five-number summary: min, Q_1, median, Q_3, max · IQR $= Q_3 - Q_1$

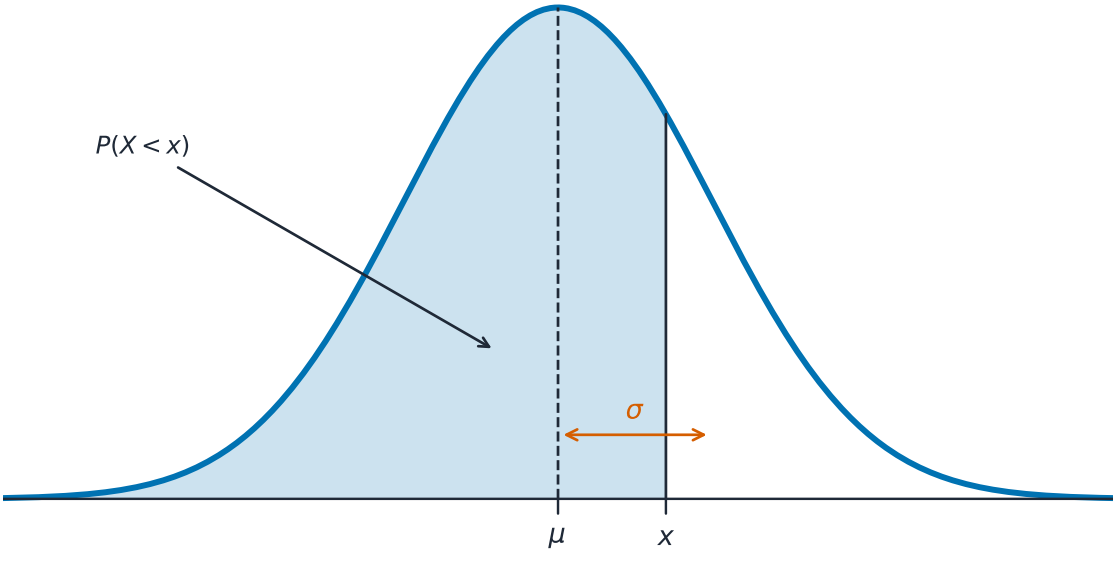
A boxplot draws the five-number summary; the box spans the IQR

Comparing Distributions of a Quantitative Variable

To compare two or more groups, compare **shape, center, and spread**, and mention outliers – always with comparative words ("Group A has a **higher** median than Group B") and in context. Do not just describe each group separately; make the comparison explicit.

The Normal Distribution

A **normal distribution** 正态分布 is a symmetric, bell-shaped model described by its mean μ and standard deviation σ. The **empirical rule** 经验法则 (68–95–99.7): about 68% of values lie within 1σ of the mean, 95% within 2σ, and 99.7% within 3σ.



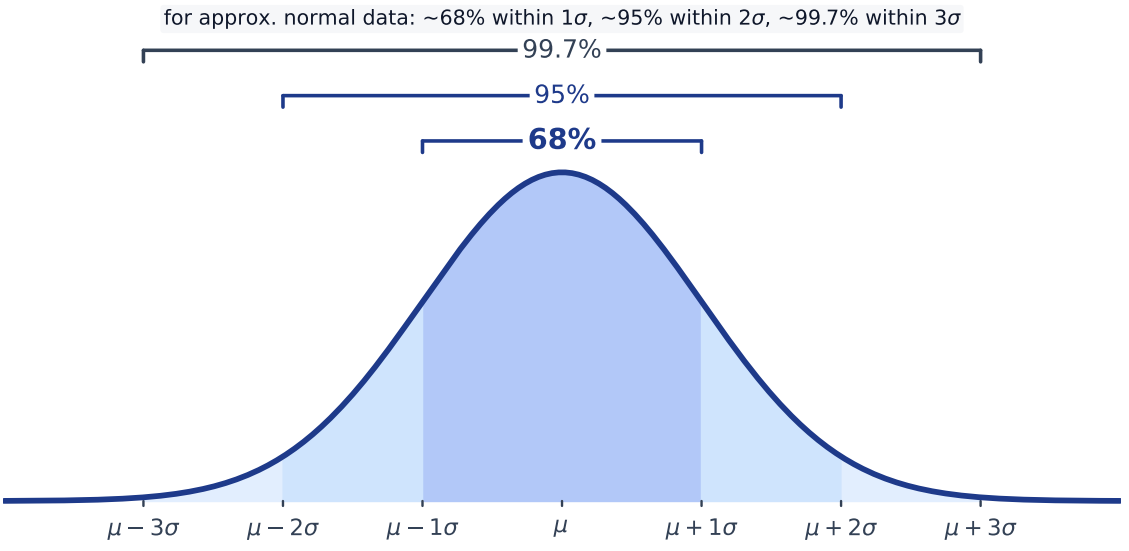
The normal curve: a probability is the area under it, centred on the mean

A z**-score** 标准分数 measures how many standard deviations a value is from the mean:

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

Convert to a z-score, then use the normal table or technology to find the **proportion** (area) below, above, or between values – and reverse the process to find a value from a given percentile.

Worked example. Test scores are normal with $\mu = 500$ and $\sigma = 100$. A score of 700 has $z = \dfrac{700 - 500}{100} = 2$. By the empirical rule, 95% of scores lie within 2σ, so 2.5% lie above 700 – meaning a 700 is at about the 97.5th percentile.



The normal curve and the 68-95-99.7 empirical rule

Exam tips

- Describe a distribution by **shape, center, spread, and outliers** (SOCS) — always in context.
- The **mean** is pulled by outliers; the **median** resists them, so prefer the median for skewed data.
- For a **normal** distribution use the 68–95–99.7 rule and z-scores $z = \frac{x-\mu}{\sigma}$.
- Compare distributions with side-by-side boxplots and comment on center, spread, and shape.
- Standard deviation measures a typical distance from the mean; the IQR pairs with the median.

Exploring Two-Variable Data

AP Statistics

Are Two Variables Related?

Two-variable data let us ask whether two characteristics are **associated** 关联 – whether knowing one tells you something about the other. An **explanatory variable** 解释变量 (the "input") may help predict a **response variable** 响应变量 (the "output"). Association is not the same as causation.

Two Categorical Variables

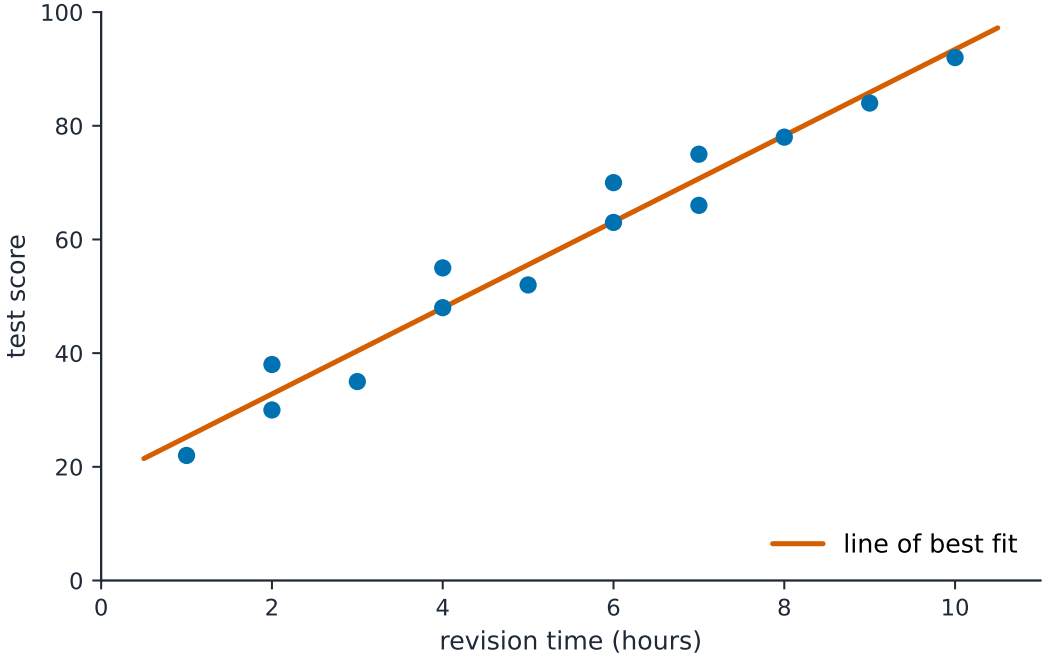
A **two-way table** 双向表 (contingency table) counts individuals by two categorical variables at once. A **marginal distribution** 边缘分布 is a row or column total written as a fraction of the grand total (the totals themselves are just counts). Comparing the inside cells shows whether the variables are related.

Comparing Groups with Conditional Distributions

A **conditional distribution** 条件分布 is the distribution of one variable *within* a fixed category of the other (found by dividing each cell by its row or column total). If the conditional distributions differ across groups, the two variables are **associated**; if they are the same, there is no association. **Segmented bar charts** 分段条形图 or mosaic plots display them.

Scatterplots for Two Quantitative Variables

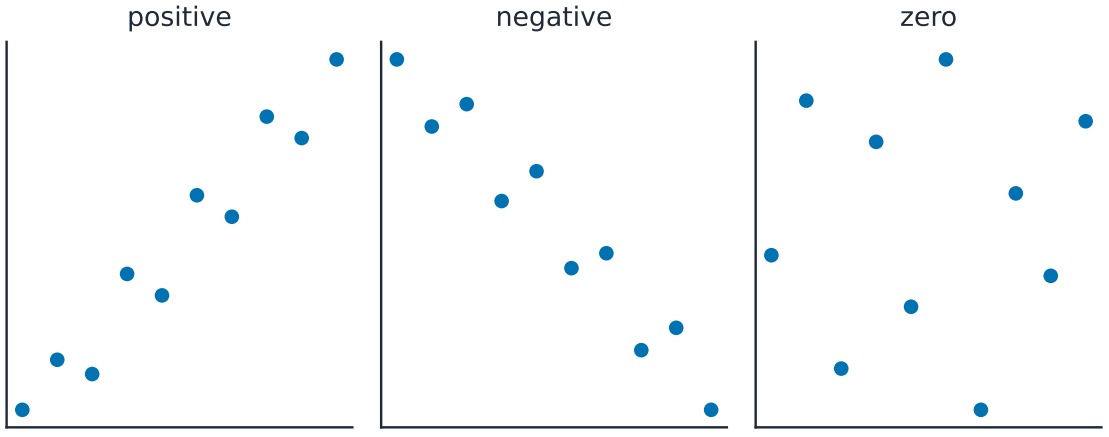
A **scatterplot** 散点图 plots each individual as a point, explanatory variable on the x-axis and response on the y-axis. Describe it with **DUFS**: **Direction** (positive/negative), **Unusual** features (outliers, clusters), **Form** (linear or curved), and **Strength** (how tightly the points follow the pattern) – always in context.



A line of best fit runs through the middle of the scattered points

Correlation

The **correlation coefficient** 相关系数 r measures the **strength and direction of a linear** relationship. It runs from -1 to 1: near ± 1 is strong linear, near 0 is weak linear. r has **no units** and does not change if you swap the variables. Warnings: r only measures *linear* strength, it is **not resistant** to outliers, and a strong r does **not** prove causation.



Positive correlation rises together; negative correlation moves in opposite directions

Linear Regression Models

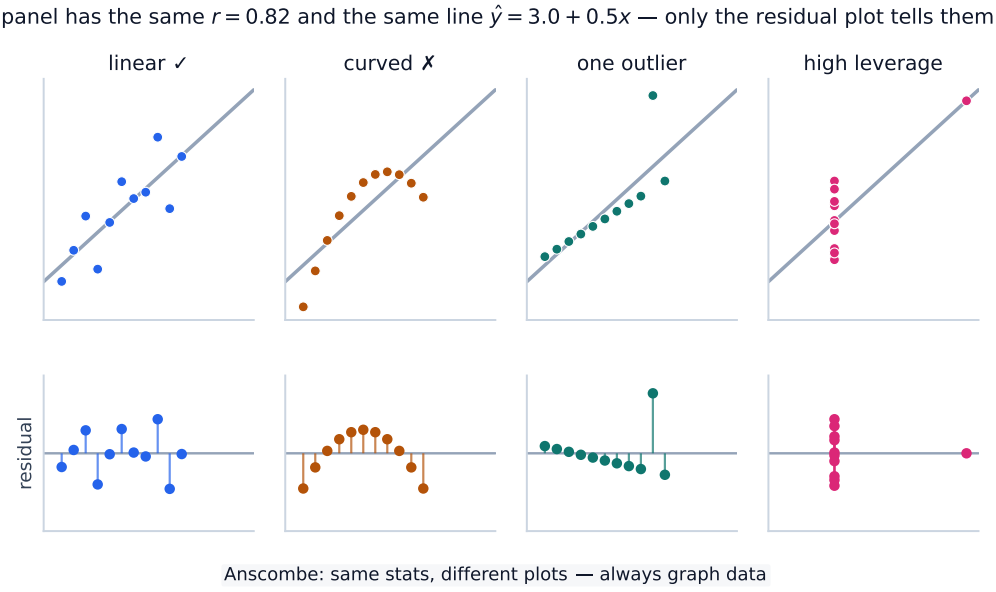
The **least-squares regression line** 最小二乘回归线 predicts the response: $\hat{y} = a + bx$, where $\hat{y}$ is the *predicted* response. The **slope** 斜率 b is the predicted change in y per one-unit increase in x; the **y-intercept** 截距 a is the predicted y when $x = 0$. Interpret both **in context and with units** – a graded skill. Avoid **extrapolation** 外推 (predicting far outside the data).

Worked example. A study of hours studied (x) and test score (y) gives $\hat{y} = 20 + 3x$. The slope means each extra hour of study is associated with a predicted 3-point increase. A student who studies 5 hours is predicted to score $\hat{y} = 20 + 3(5) = 35$.

Residuals

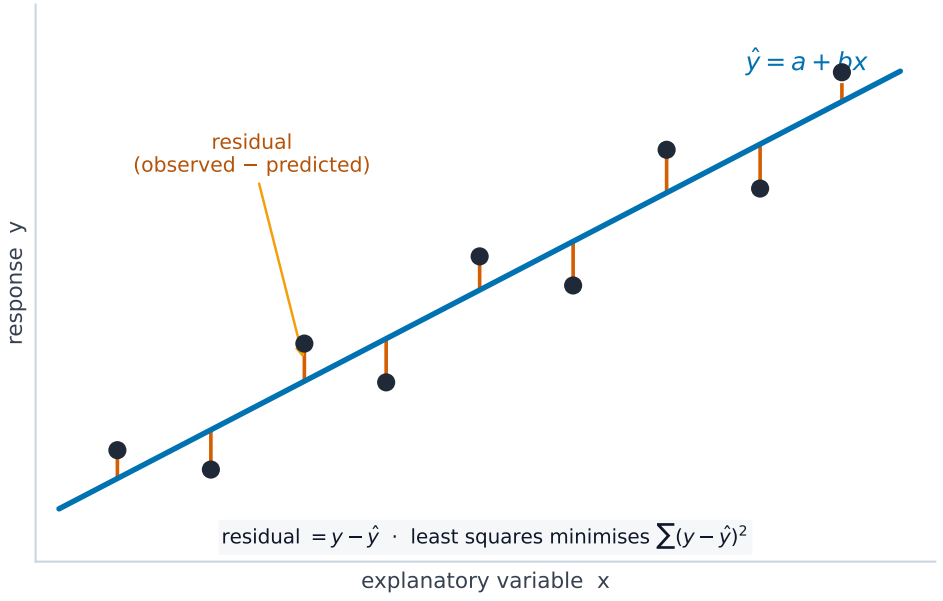
A **residual** 残差 is actual minus predicted, $y - \hat{y}$: how far a point sits above (+) or below (−) the line. A **residual plot** 残差图 graphs residuals against x. If it shows **no pattern** (random scatter), a linear model is appropriate; a curved or fanning pattern means the linear model is a poor fit.

Worked example. Continuing the study above, a student who studied 5 hours actually scored 40. The residual is $y - \hat{y} = 40 - 35 = +5$: the line **under-predicted** by 5 points, so this point sits above the line.



*A caution about r and the line: all four datasets have the same $r = 0.82$ and the same $\hat{y} = 3.0 + 0.5x$, yet only the first is genuinely linear. The scatterplots barely differ — the **residual plot** below each is what exposes the curve, the outlier, and the high-leverage point.*

Least-Squares Regression and Its Fit



The least-squares line minimizes the sum of squared residuals

The line minimizes the sum of squared residuals. Its fit is measured by:

- s, the standard deviation of the residuals – the typical prediction error, in the response's units.
- r^2, the **coefficient of determination** 决定系数 – the proportion of the variation in y that the linear model explains (a value between 0 and 1; multiply by 100 to state it as a percent). Report it in context: "$r^2 = 0.81$ means 81% of the variation in y is explained by the linear relationship with x."

Departures from Linearity

Some points strongly affect the line. A **high-leverage** 高杠杆 point has an extreme x-value; an **influential** 有影响的 point noticeably changes the slope or r when removed; an **outlier** here is a point with a large residual. When the pattern is curved, **transform** a variable (e.g. take a log) to straighten it, then fit a line to the transformed data.

Exam tips

- On a scatterplot describe **direction, form, strength, and outliers**; r ranges -1 to 1.
- **Correlation is not causation** — a lurking variable can drive both.
- Interpret the slope of the least-squares line in context ("per one unit of x, predicted y changes by b").
- Check a **residual plot**: no pattern means a line fits; a curve means it does not. Avoid extrapolation.
- r^2 is the fraction of variation in y explained by the model.

Collecting Data

AP Statistics

Can We Trust the Data We Collected?

A conclusion is only as good as the data behind it. **How** data are collected decides what you may conclude – whether you can generalize to a **population** 总体, and whether you can claim **cause and effect**. Poorly collected data can be worse than none.

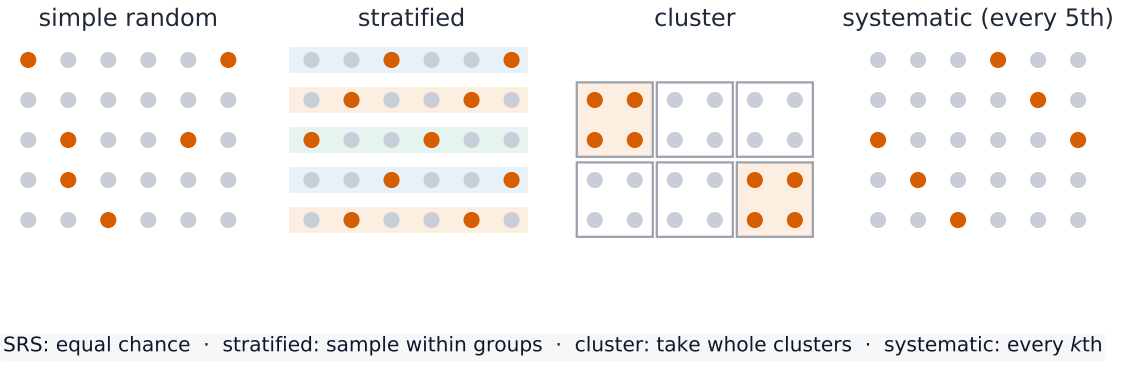
Observational Studies and Experiments

- In an **observational study** 观察性研究 you measure individuals without trying to influence them. It can show **association**, but not causation, because lurking variables may explain the link.
- In an **experiment** 实验 you deliberately impose a **treatment** 处理 and compare responses. A well-designed experiment **can** establish cause and effect.

Random Sampling

To learn about a population you take a **sample** 样本. **Random sampling** 随机抽样 protects against selection **bias** 偏差 and lets you generalize (it cannot fix undercoverage, nonresponse, or response bias — see below). Common designs:

- **Simple random sample (SRS)** 简单随机样本: every group of the chosen size is equally likely.
- **Stratified** 分层: split the population into similar strata, then sample within each.
- **Cluster** 整群: split into clusters, randomly choose whole clusters.
- **Systematic** 系统: pick every kth individual from a random start.



Four random sampling designs: who gets selected, and how

A **convenience sample** 方便样本 or **voluntary response** sample is **not** random and is biased.

Worked example. To survey a school, an administrator lists all students by grade and randomly selects 20 from each grade. This is a **stratified** sample – the grades are the strata – which guarantees every grade is represented, unlike an SRS that might by chance draw few from one grade.

Random outcomes: dice make each face equally likely under fair conditions

When Sampling Goes Wrong

Bias makes estimates systematically miss the truth:

- **Undercoverage** 覆盖不足: some groups are left out of the sampling frame.
- **Nonresponse** 无回应: selected people do not answer.
- **Response bias** 回应偏差: people answer inaccurately (bad wording, sensitive topics).

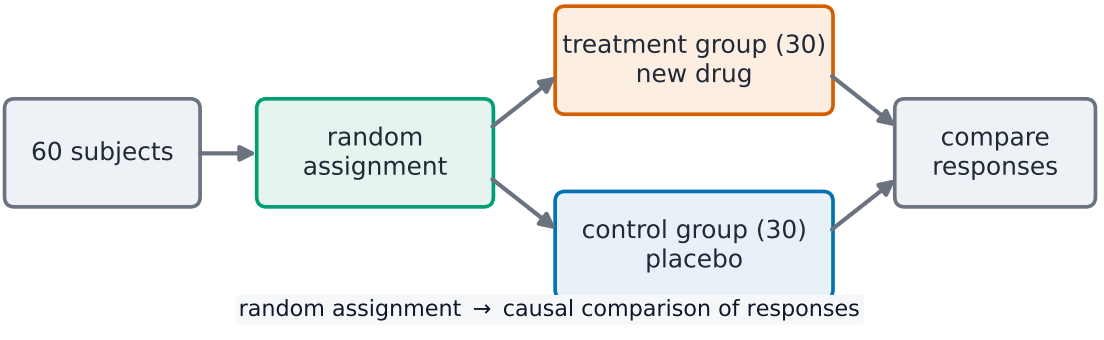
Bias is about a **consistent** error in one direction – increasing the sample size does not fix it.

Convenience samples miss the population: bias creeps in when selection is not random

Designing an Experiment

Good experiments follow three principles:

- **Comparison** with a **control group** 对照组 (often a **placebo** 安慰剂).
- **Random assignment** 随机分配 of subjects to treatments, to balance out other variables.
- **Replication** 重复: enough subjects per treatment to see a real effect.



A completely randomized experiment compares a treatment group with a control group

Confounding 混杂 occurs when another variable is tied to the treatment so their effects cannot be separated; random assignment guards against it. **Blinding** 盲法 hides who is getting which treatment to prevent expectation effects: in a **single-blind**

单盲 study only one side is kept unaware (usually the subjects, or only the people assessing the result), while in a **double-blind** 双盲 study **neither** the subjects nor the researchers who interact with them know, blocking both the placebo effect and biased assessment. **Blocking** 区组 groups similar subjects and randomizes within each block to reduce variability.

A clinical trial: random assignment separates treatment from control

Choosing the Right Design

Match the design to the goal: use a **completely randomized design** for uniform subjects; a **randomized block design** when a known variable (sex, age) affects the response; a **matched-pairs** design when each subject can serve as its own control. State how you would carry out the randomization.

What an Experiment Lets You Conclude

Two questions decide the scope of a conclusion:

- **Random assignment** used? Then a significant difference can be attributed to the treatment (**causation**) – for these subjects.
- **Random sampling** from a population? Then results **generalize** to that population.

Only an experiment with random assignment supports a cause-and-effect claim; only random sampling supports generalization. Say exactly which you have.

Worked example. Researchers randomly assign 100 **volunteers** to a new drug or a placebo, and the drug group improves significantly more. Because of the **random assignment**, the improvement can be attributed to the drug (causation) – but because the subjects were **not randomly sampled**, the conclusion applies only to these volunteers and does not automatically generalize to everyone.

Exam tips

- Distinguish an **observational study** (finds association) from an **experiment** (can show causation).
- Good sampling is **random** (SRS, stratified, cluster) — beware bias (voluntary response, undercoverage, nonresponse).
- Good experiments use **control, randomization, and replication**; blocking handles a known nuisance variable.
- Only a randomized experiment supports a cause-and-effect conclusion.
- Name the population, sample, and any confounding clearly.

Probability, Random Variables, and Probability Distributions

AP Statistics

Random and Non-Random Patterns

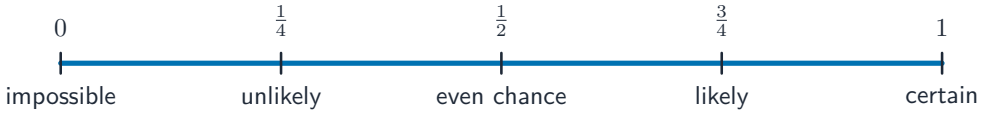
Something is **random** 随机 if individual outcomes are uncertain but a regular pattern emerges over **many** repetitions. Short-run results look erratic; long-run **relative frequencies** settle down. This long-run stability is what makes probability useful.

Estimating Probabilities Using Simulation

A **simulation** 模拟 imitates a chance process using random digits or technology. Steps: state the model, assign digits to outcomes, run many trials, and record the proportion of trials meeting the condition. The resulting proportion **estimates** the probability – more trials give a better estimate.

Introduction to Probability

The **probability** 概率 of an event is a number from 0 to 1 giving its long-run relative frequency. The **sample space** 样本空间 is the set of all outcomes. For an event A, the **complement** 补 rule: $P(A^c) = 1 - P(A)$. Probabilities of all outcomes sum to 1.



Probability runs from 0 (impossible) to 1 (certain)

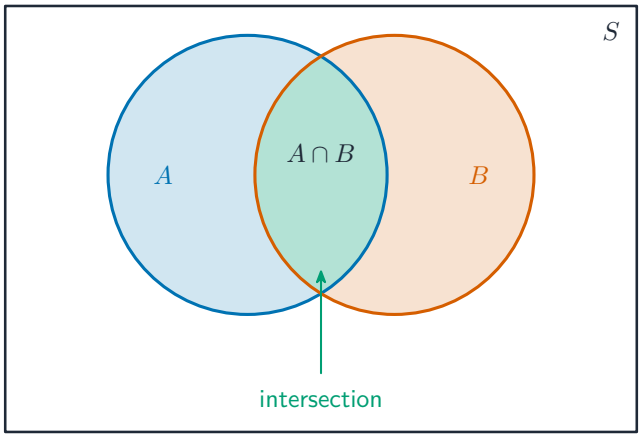
A deck of cards is a classic source of probability: 52 equally likely outcomes make the chances easy to count

Mutually Exclusive Events

Two events are **mutually exclusive** 互斥 (disjoint) if they cannot both happen. Then the **addition rule** simplifies:

$$P(A \text{ or } B) = P(A) + P(B) \quad \text{(if mutually exclusive)}.$$

In general, $P(A \text{ or } B) = P(A) + P(B) - P(A \text{ and } B)$ – subtract the overlap so it is not counted twice.



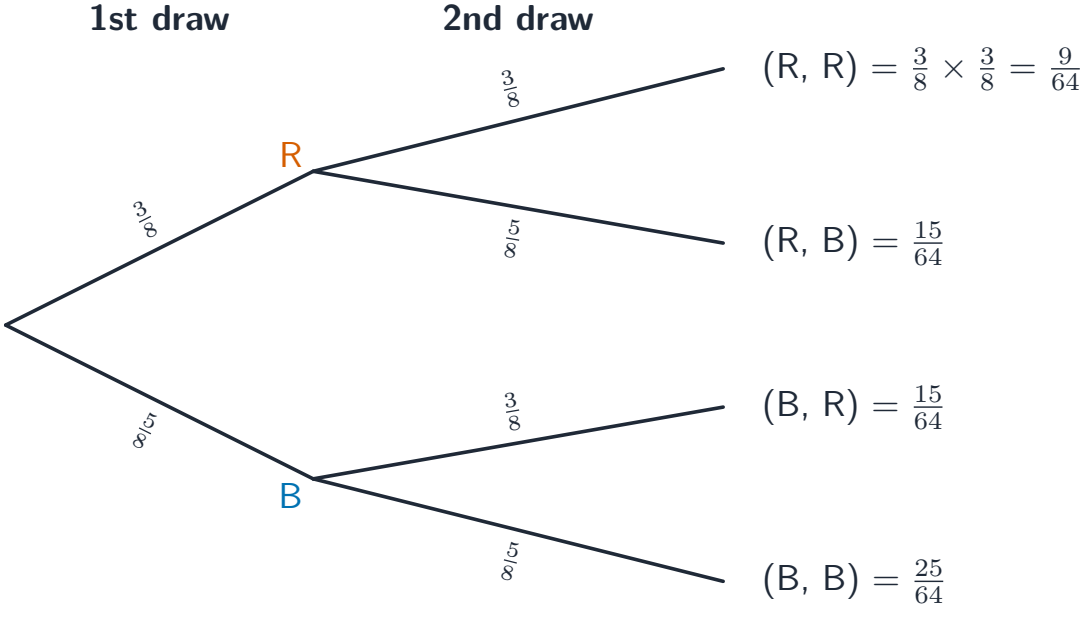
A Venn diagram: the overlap is the intersection of two events

Conditional Probability

The **conditional probability** 条件概率 of A given B is

$$P(A \mid B) = \frac{P(A \text{ and } B)}{P(B)}.$$

It is the chance of A once you know B happened. Two-way tables make these easy: restrict to the row/column for B, then find A's share.



On a tree diagram, multiply the probabilities along the branches

Independent Events and Unions of Events

Events are **independent** 独立 if knowing one does not change the other's probability: $P(A \mid B) = P(A)$. Then the **multiplication rule** simplifies:

$$P(A \text{ and } B) = P(A)\,P(B) \quad \text{(if independent)}.$$

Independent is not the same as mutually exclusive – mutually exclusive events with nonzero probability are actually **dependent** (if one happens, the other cannot).

die 1 \ die 2	1	2	3	4	5	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 six highlighted cells all give a total of 7

A sample space diagram lists every equally likely outcome

Random Variables and Probability Distributions

A **random variable** 随机变量 assigns a number to each outcome of a chance process. A **probability distribution** 概率分布 lists each possible value with its probability (they sum to 1). A distribution can be **discrete** (a table of values) or **continuous** (an area-under-a-curve model like the normal).

Mean and Standard Deviation of Random Variables

The **mean (expected value)** 期望值 of a discrete random variable is the probability-weighted average:

$$\mu_X = E(X) = \sum x_i \, P(x_i).$$

The **standard deviation** $\sigma_X = \sqrt{\sum (x_i - \mu_X)^2 \, P(x_i)}$ measures typical spread from the mean. The expected value is the long-run average outcome, not a value you expect on any single trial.

Worked example. A game pays \$5 with probability 0.2 and costs you \$1 (a -1 outcome) with probability 0.8. The expected value is

$$E(X) = 5(0.2) + (-1)(0.8) = 1 - 0.8 = \$0.20,$$

so over many plays you gain about 20 cents per play on average, even though no single play gives exactly that.

Combining Random Variables

When you add or subtract random variables, **means add**: $\mu_{X \pm Y} = \mu_X \pm \mu_Y$. If X and Y are **independent**, **variances add** (even when subtracting):

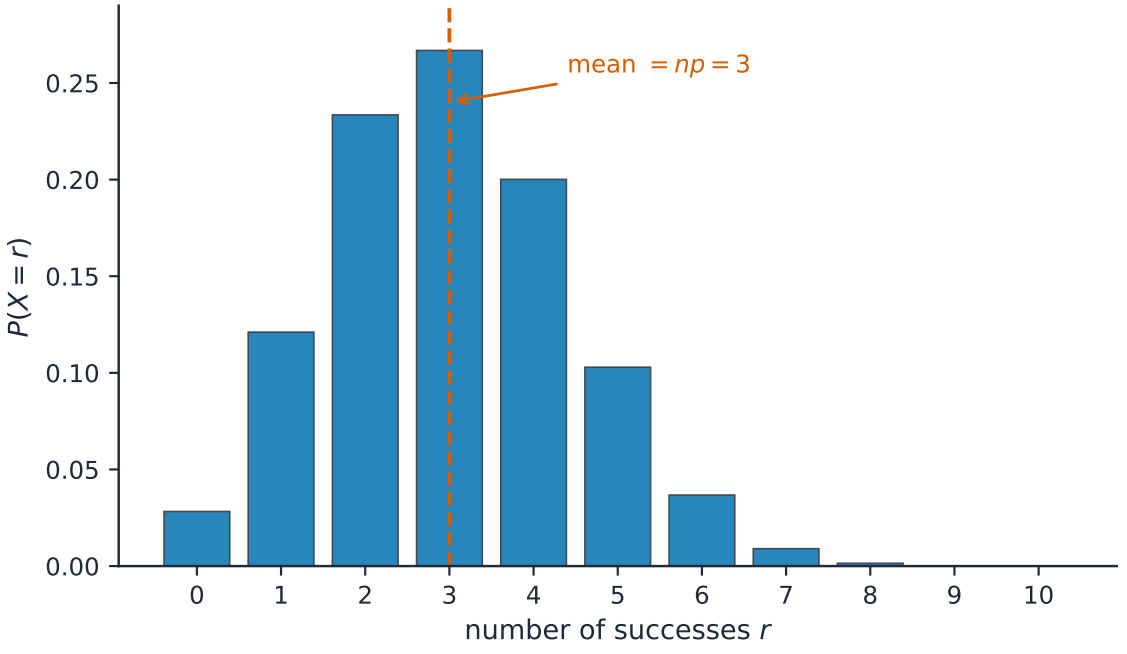
$$\sigma^2_{X \pm Y} = \sigma^2_X + \sigma^2_Y.$$

Take the square root for the standard deviation. Also, scaling: $\mu_{aX+b} = a\mu_X + b$ and $\sigma_{aX+b} = |a|\sigma_X$.

Introduction to the Binomial Distribution

A **binomial** 二项 setting (BINS): a fixed number n of **Independent** trials, each with two outcomes (success/failure) and the **same** success probability p. The random variable X = number of successes. Its probability:

$$P(X = k) = \binom{n}{k} p^k (1-p)^{n-k}.$$



The binomial distribution, with mean n times p

Parameters for a Binomial Distribution

For a binomial X with n trials and success probability p:

$$\mu_X = np, \qquad \sigma_X = \sqrt{np(1-p)}.$$

Use these for "how many successes do we expect, and how much do they vary" questions.

Worked example. A player makes 70% of free throws. In $n = 10$ shots, the probability of exactly 8 makes is

$$P(X = 8) = \binom{10}{8} (0.7)^8 (0.3)^2 = 45 \times 0.0576 \times 0.09 \approx 0.23,$$

and the expected number of makes is $\mu = np = 10(0.7) = 7$, with $\sigma = \sqrt{10(0.7)(0.3)} \approx 1.45$.

The Geometric Distribution

A **geometric** 几何 setting is the same as binomial but with **no fixed** n: you keep trying until the **first success**. The random variable $Y =$ the trial of the first success:

$$P(Y = k) = (1 - p)^{k-1} p, \qquad \mu_Y = \frac{1}{p}.$$

So the expected number of trials until the first success is $1/p$.

Exam tips

- A probability lies in $[0, 1]$; use the **complement** $(1-P)$ and add mutually exclusive events.
- For independent events multiply; for "and/or" use the general addition and conditional rules.
- **Expected value** $= \sum$(value $\times$ probability).
- Recognise **binomial** (fixed n, two outcomes, constant p) and **geometric** settings.
- Draw a tree or table for multi-stage problems and multiply along branches.

Sampling Distributions

AP Statistics

Why Two Samples Never Match: Sampling Variability

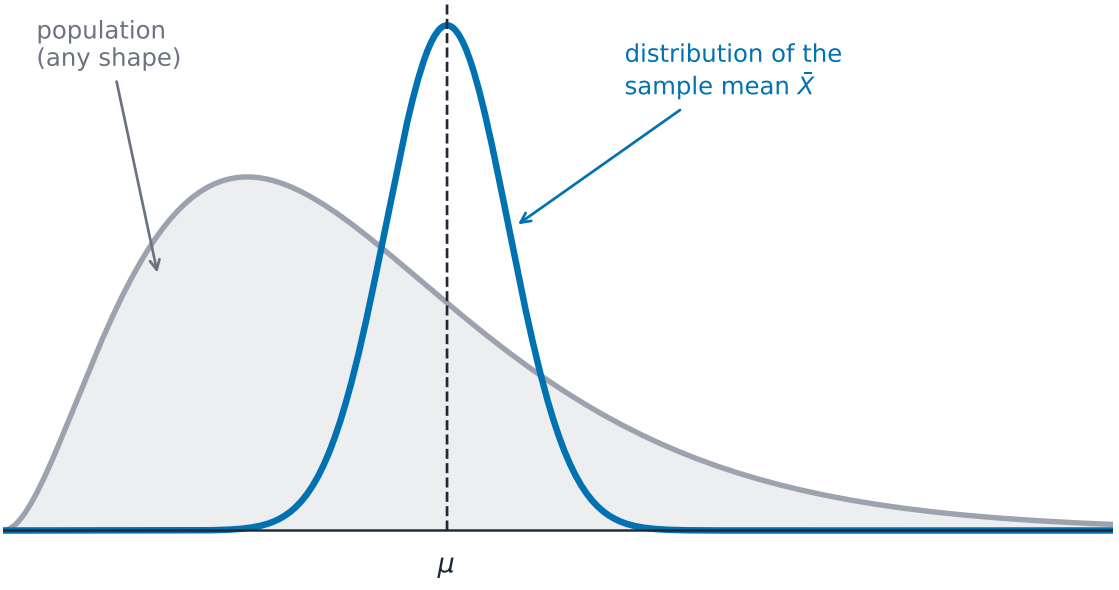
A **statistic** 统计量 (like a sample mean $\bar{x}$ or sample proportion $\hat{p}$) is computed from a sample and **varies** from sample to sample – this is **sampling variability** 抽样变异. A **parameter** 参数 (μ or p) is the fixed truth about the population. The **sampling distribution** 抽样分布 is the distribution of a statistic over all possible samples of a given size – it is the bridge from one sample to inference.

The Normal Curve as a Model for a Statistic

For large enough samples, many sampling distributions are approximately **normal**. That lets us describe a statistic by a **center** (its mean), a **spread** (its **standard error** 标准误), and a normal shape – and then compute how likely a given sample result is.

The Central Limit Theorem

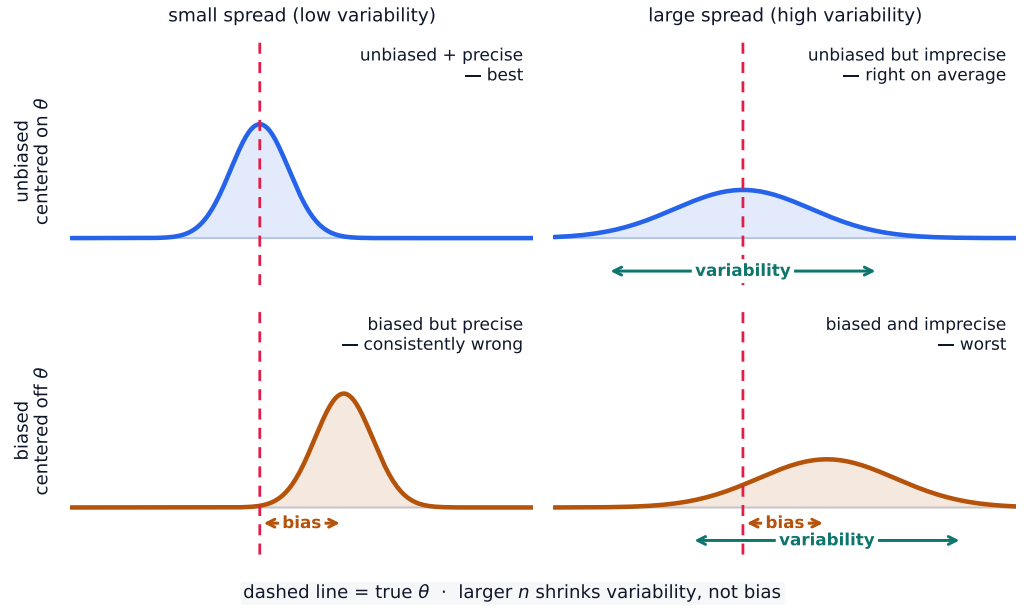
The **Central Limit Theorem** 中心极限定理 (CLT): for a **sample mean**, if the sample size n is large enough (a common rule is $n \geq 30$), the sampling distribution of $\bar{x}$ is approximately **normal**, *regardless* of the population's shape. The larger n, the more normal and the tighter the distribution.



The sample mean is nearly normal whatever the shape of the population

Good Guesses and Bad Guesses: Bias

A statistic is **unbiased** 无偏 if the mean of its sampling distribution equals the parameter – it is correct **on average**. Bias is about the center being off; **variability** is about the spread. A good estimator is both unbiased (centered right) and low-variability (precise); larger samples reduce variability but do not fix bias from bad sampling.



*Bias and variability are separate faults. Only the top-left estimator is both centered on θ and tight; the bottom-left one is precise but **consistently wrong**, which no amount of extra data will fix.*

The Sampling Distribution of a Sample Proportion

For a sample proportion $\hat{p}$ from an SRS: the mean is p (unbiased), and the standard deviation is

$$\sigma_{\hat{p}} = \sqrt{\frac{p(1-p)}{n}}.$$

This spread has two names: it is the **standard deviation** of the sampling distribution, and it is called the **standard error** once you must estimate it from the sample (replacing p by $\hat{p}$) — which is exactly what the later inference units do. It is approximately **normal** when $np \geq 10$ and $n(1-p) \geq 10$ (the **Large Counts** condition), and the 10% condition ($n \leq 0.10N$) keeps the observations near-independent.

Worked example. Suppose 40% of voters favor a measure ($p = 0.4$) and you sample

$n = 100$. The standard error is $\sigma_{\hat{p}} = \sqrt{\dfrac{0.4(0.6)}{100}} = 0.049$. The chance a sample gives $\hat{p} > 0.5$ is $z = \dfrac{0.5 - 0.4}{0.049} = 2.04$, so $P(\hat{p} > 0.5) \approx 0.02$ – a majority in the sample would be surprising.

Comparing Two Groups: Difference of Sample Proportions

For $\hat{p}_1 - \hat{p}_2$ from two independent samples: the mean is $p_1 - p_2$, and because the samples are independent the **variances add**:

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}.$$

It is approximately normal when the Large Counts condition holds in **both** samples.

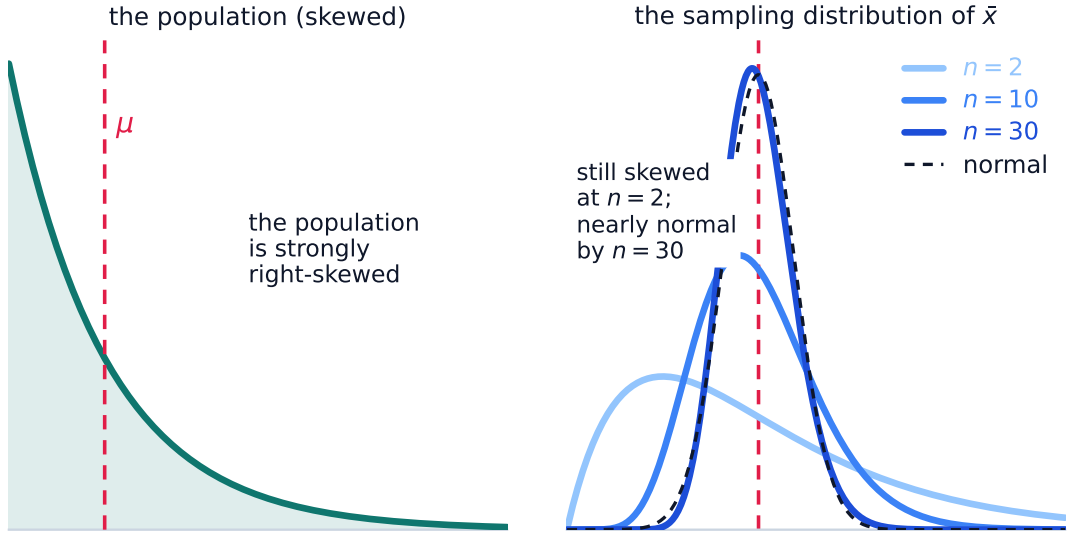
The Sampling Distribution of a Sample Mean

For a sample mean $\bar{x}$ from an SRS: the mean is μ (unbiased), and the standard deviation is

$$\sigma_{\bar{x}} = \frac{\sigma}{\sqrt{n}}.$$

Its shape is normal if the population is normal, or approximately normal for large n by the CLT. Note the spread shrinks like $\sqrt{n}$ – quadrupling the sample halves the standard error.

Worked example. A population has $\mu = 70$ and $\sigma = 12$. For samples of $n = 36$, the sampling distribution of $\bar{x}$ is centered at 70 with standard error $\dfrac{12}{\sqrt{36}} = 2$. The chance a sample mean exceeds 73 is $z = \dfrac{73 - 70}{2} = 1.5$, so $P(\bar{x} > 73) \approx 0.067$.



CLT: for large n, $\bar{x} \approx \text{Normal}(\mu,\ \sigma/\sqrt{n})$ even when the population is not normal

The population on the left is strongly skewed, yet every sampling distribution of $\bar{x}$ is centered at μ. A larger n shrinks the standard error $\sigma/\sqrt{n}$, so the curve gets taller and narrower – and it also straightens: still clearly skewed at $n = 2$, almost exactly normal (dashed) by $n = 30$.

Comparing Two Groups: Difference of Sample Means

For $\bar{x}_1 - \bar{x}_2$ from two independent samples: the mean is $\mu_1 - \mu_2$, and (independent, so variances add)

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}.$$

This is the foundation for two-sample inference in the next units.

Exam tips

- A **sampling distribution** is the distribution of a statistic over many samples, **centered on the true parameter**.
- The **Central Limit Theorem**: for a large enough sample the sample mean is approximately normal, even if the population is not.
- Larger samples give **less** variability (a smaller standard error).
- Check the conditions (random, independent/10%, large enough) before using a normal model.
- Keep straight what varies — the statistic — versus the fixed parameter.

Inference for Categorical Data: Proportions

AP Statistics

Why Be Normal?

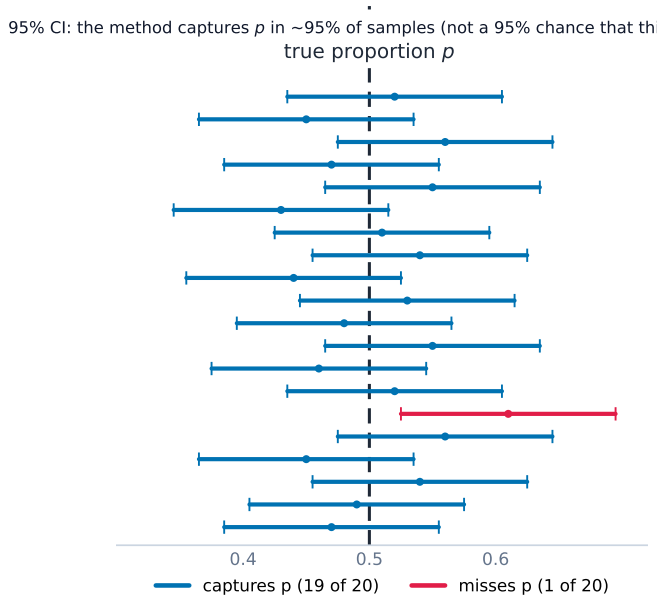
Because a sample proportion $\hat{p}$ is approximately **normally** distributed (when the conditions hold), we can measure how far a sample result is from a claimed value in standard errors, and turn that into a probability. This is what makes **inference** 推断 – drawing conclusions about a population from a sample – possible.

Confidence Interval for a Proportion

A **confidence interval** 置信区间 estimates the parameter as a range: statistic ± **margin of error** 误差幅度.

$$\hat{p} \pm z^* \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}.$$

z^* is the critical value for the **confidence level** 置信水平 (e.g. 1.96 for 95%). Conditions: **random** sample, **Large Counts** ($n\hat{p} \geq 10$ and $n(1-\hat{p}) \geq 10$), and the **10%** condition. Interpret it: "We are 95% confident the true proportion of... is between... and...". Interpret the **level**: "In 95% of samples, this method produces an interval that captures the true proportion."

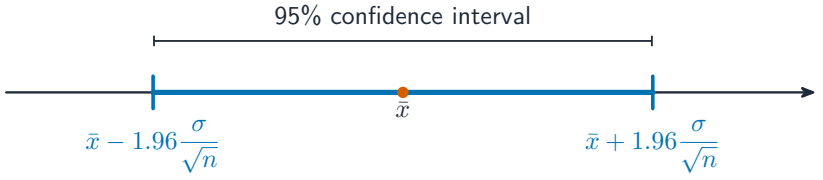


Over many samples, about 95% of 95% confidence intervals capture the true proportion

Worked example. In a random sample of 200 people, 120 support a policy, so $\hat{p} = 0.60$. A 95% interval uses $z^* = 1.96$:

$$0.60 \pm 1.96\sqrt{\frac{0.60(0.40)}{200}} = 0.60 \pm 0.068 = (0.532,\ 0.668).$$

We are 95% confident the true proportion of supporters is between 53.2% and 66.8%.



A 95% confidence interval reaches 1.96 standard errors each side of the estimate

Choosing the sample size. To keep the margin of error no larger than a target m, set $z^*\sqrt{\dfrac{\hat{p}(1-\hat{p})}{n}} \le m$ and solve for n. When you have no estimate of $\hat{p}$, use $\hat{p} = 0.5$: it makes $\hat{p}(1-\hat{p})$ as large as possible, giving the **safe (largest) required sample size**. Always round the result **up** to the next whole person.

Worked example. How many people must you survey for a 95% interval with margin of error at most 0.03? Using $\hat{p} = 0.5$ and $z^* = 1.96$:

$$n = \frac{(z^*)^2\,\hat{p}(1-\hat{p})}{m^2} = \frac{1.96^2(0.5)(0.5)}{0.03^2} = \frac{0.9604}{0.0009} \approx 1067.1,$$

so you survey 1068 people (always round **up**, since 1067 would leave the margin a shade too big).

Justifying a Claim from an Interval

To judge a claimed value: if it lies **inside** the interval, the data are consistent with it; if it lies **outside**, the data give evidence against it. Base the conclusion on whether the plausible values include the claim, in context.

Setting Up a Test for a Proportion

A **significance test** 显著性检验 weighs evidence against a claim. State a **null hypothesis** 原假设 H_0 and an **alternative hypothesis** 备择假设 H_a about the parameter p:

$$H_0 : p = p_0 \qquad H_a : p \neq p_0 \text{ (or } <, >).$$

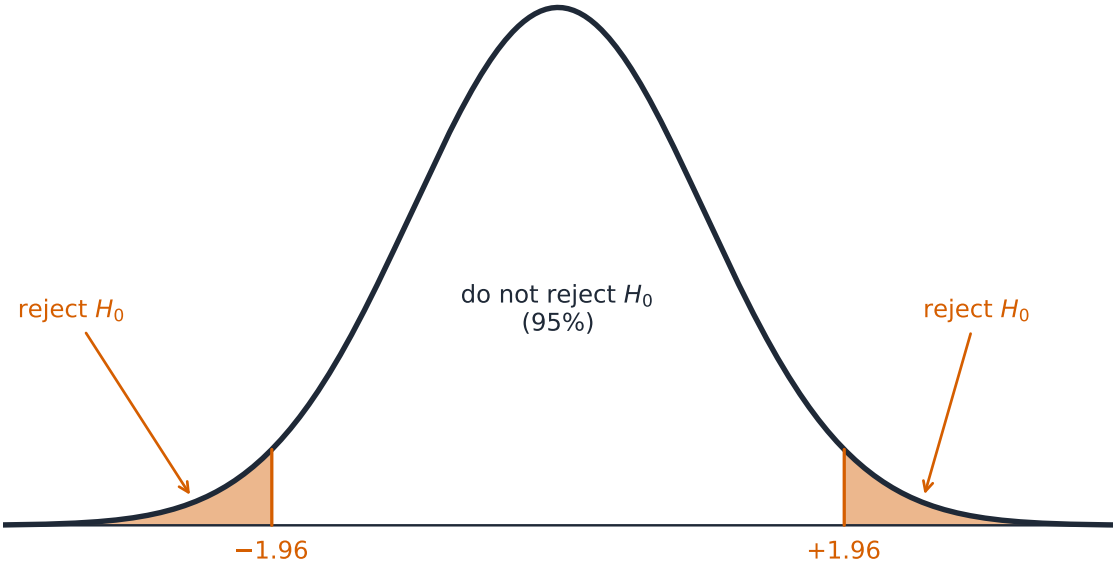
Check the same conditions (random, Large Counts using p_0, 10%). The **test statistic** 检验统计量 counts standard errors from p_0:

$$z = \frac{\hat{p} - p_0}{\sqrt{p_0(1-p_0)/n}}.$$

Worked example. A company claims 90% satisfaction ($p_0 = 0.90$); a sample of 100 finds 84 satisfied ($\hat{p} = 0.84$). Test $H_0 : p = 0.90$ vs $H_a : p \neq 0.90$ at $\alpha = 0.05$:

$$z = \frac{0.84 - 0.90}{\sqrt{0.90(0.10)/100}} = \frac{-0.06}{0.03} = -2.0,$$

giving a two-tailed p-value of about $2(0.023) = 0.046$. Since $0.046 < 0.05$, **reject** H_0 – there is evidence the true satisfaction rate differs from (is below) 90%.



A two-tailed 5% test rejects the null hypothesis in the shaded tails

Interpreting p-Values

The *p*-**value** P 值 is the probability of getting a sample result **as extreme or more extreme** than the observed one, **assuming** H_0 **is true**. A small *p*-value means the data would be surprising if H_0 held – evidence against H_0. It is *not* the probability that H_0 is true.

Concluding a Test

Compare the *p*-value to the **significance level** 显著性水平 α (often 0.05):

- $p \leq \alpha$: **reject** H_0 – there is convincing evidence for H_a.
- $p > \alpha$: **fail to reject** H_0 – not enough evidence for H_a (never "accept H_0").

Always write the conclusion **in context**, linking back to the claim.

Type I and Type II Errors

- A **Type I error** 第一类错误: rejecting a **true** H_0 (a false alarm). Its probability is α.
- A **Type II error** 第二类错误: failing to reject a **false** H_0 (a missed detection). Its probability is β.
- The **power** 检验效能 $= 1 - \beta$ is the chance of correctly detecting a real effect. Power rises with a larger sample, a larger effect, or a larger α.

Describe each error and its consequence in the problem's context.

Confidence Interval for a Difference of Proportions

To compare two proportions, estimate $p_1 - p_2$:

$$(\hat{p}_1 - \hat{p}_2) \pm z^* \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}.$$

Conditions must hold in **both** samples, and the samples must be independent.

Justifying a Claim About Two Proportions

If the interval for $p_1 - p_2$ contains 0, the data are consistent with **no difference**; if it lies entirely above or below 0, there is evidence of a difference (in that direction). State the direction and context.

Setting Up a Test for a Difference

Hypotheses compare the two proportions: $H_0 : p_1 = p_2$ versus $H_a : p_1 \neq p_2$ (or $<, >$). Because H_0 says the proportions are equal, use a **combined (pooled)** 合并 sample proportion $\hat{p}_c = \dfrac{\text{total successes}}{\text{total sample size}}$ to estimate the common p.

Carrying Out a Test for a Difference

The pooled two-proportion z statistic:

$$z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}_c(1-\hat{p}_c)\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}}.$$

Find the p-value from the normal model, compare to α, and conclude in context – the same four-step logic as the one-proportion test.

Exam tips

- State the conditions (random, 10%, large counts np, $nq \geq 10$) before any proportion inference.
- A **confidence interval** = estimate $\pm$ margin of error; "95% confident" refers to the **method's** long-run capture rate.
- For a **test**, write H_0 and H_a, compute the test statistic, find the p-value, and compare to α.
- A small p-value is evidence **against** H_0; failing to reject does **not** prove H_0.
- Larger samples shrink the margin of error; a higher confidence level widens it.

Inference for Quantitative Data: Means

AP Statistics

Should I Worry About Error?

Inference for a **mean** works like inference for a proportion, with one change: we rarely know the population standard deviation σ, so we estimate it with the sample s. That extra uncertainty means we use the ***t*-distribution** instead of the normal – a **distribution** 分布 that is bell-shaped but with heavier tails, and it depends on the **degrees of freedom** 自由度 $df = n - 1$; as n grows it approaches the normal.

Confidence Interval for a Mean

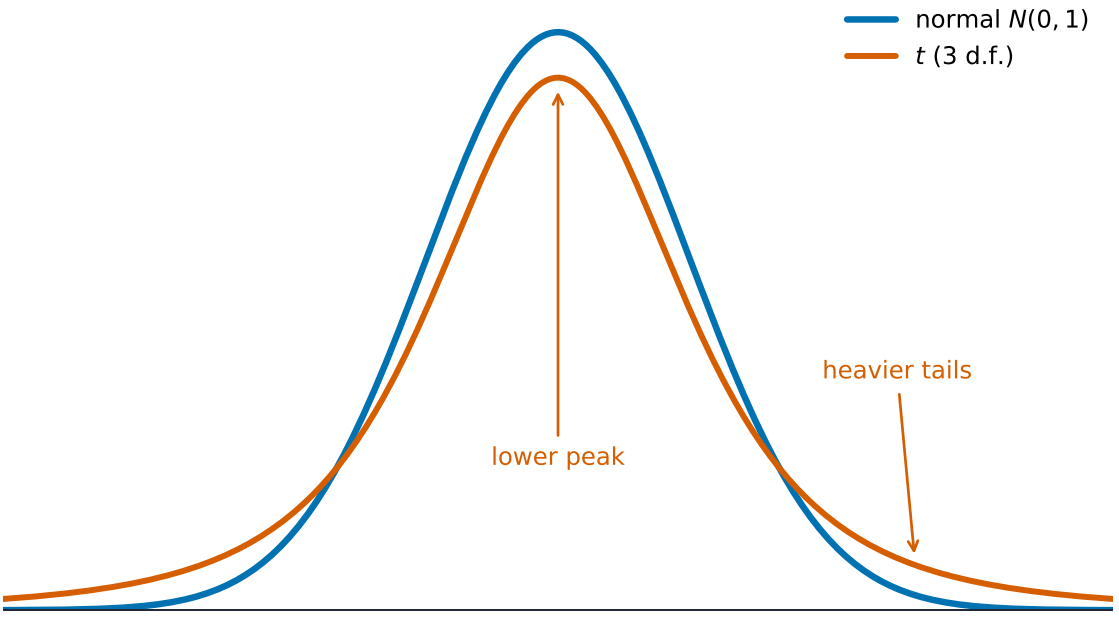
A one-sample t interval for μ:

$$\bar{x} \pm t^* \frac{s}{\sqrt{n}}.$$

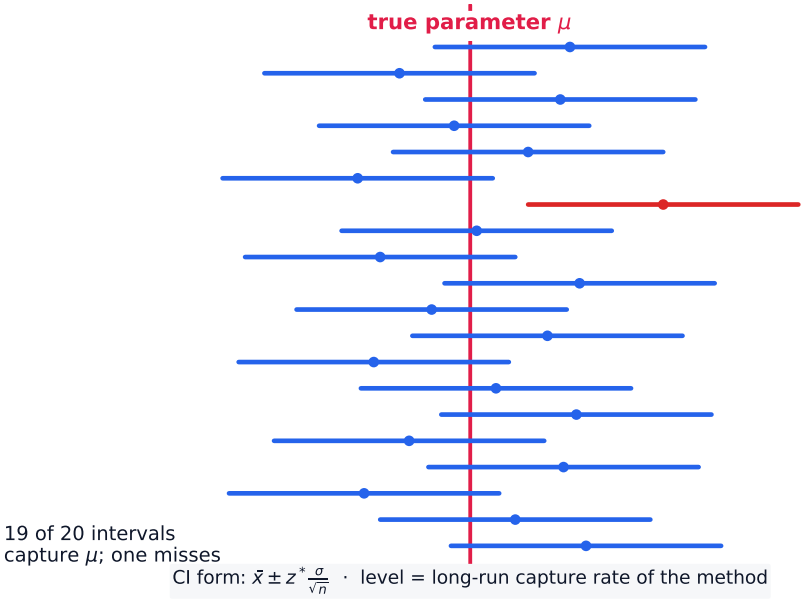
t^* is the critical value with $df = n - 1$. Conditions: **random** sample, **Normal/Large Sample** (population normal, or $n \geq 30$ by the CLT, or a roughly symmetric sample with no outliers), and the **10%** condition. Interpret the interval and the confidence level in context.

Worked example. A random sample of $n = 25$ has $\bar{x} = 50$ and $s = 8$. For a 95% interval, $df = 24$ gives $t^* = 2.064$:

$$50 \pm 2.064 \cdot \frac{8}{\sqrt{25}} = 50 \pm 2.064(1.6) = 50 \pm 3.3 = (46.7,\ 53.3).$$



The t-distribution has a lower peak and heavier tails than the normal



"95% confident" describes the method, not one interval: over many samples about 95% of the intervals contain μ and about 5% miss it.

Justifying a Claim About a Mean

As with proportions: a claimed mean **inside** the interval is plausible; **outside** the interval, the data give evidence against it. Answer in context using the plausible range.

Setting Up a Test for a Mean

State hypotheses about μ: $H_0 : \mu = \mu_0$ versus $H_a : \mu \neq \mu_0$ (or $<, >$). Check the same conditions. The one-sample t **statistic**:

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}}, \qquad df = n - 1.$$

Worked example. Test $H_0 : \mu = 45$ against $H_a : \mu \neq 45$ for the sample above ($\bar{x} = 50$, $s = 8$, $n = 25$):

$$t = \frac{50 - 45}{8/\sqrt{25}} = \frac{5}{1.6} = 3.13, \qquad df = 24.$$

This t is far out in the tail (two-tailed $p < 0.01$), so **reject** H_0 – strong evidence the mean is not 45. Notice 45 also falls outside the 95% interval $(46.7, 53.3)$, the same conclusion by two routes.

Carrying Out a Test for a Mean

Find the p-value from the t-distribution with $df = n - 1$, compare to α, and conclude in context – reject or fail to reject H_0, then state what that means for the claim. Show the test name, statistic, df, and p-value.

Confidence Interval for a Difference of Two Means

For **independent** samples, estimate $\mu_1 - \mu_2$:

$$(\bar{x}_1 - \bar{x}_2) \pm t^* \sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}.$$

Conditions must hold in both samples. (Use technology for the df; do not pool the variances on the AP exam.)

Randomisation underpins fair comparison of two groups in a mean difference test

Justifying a Claim About Two Means

If the interval for $\mu_1 - \mu_2$ contains 0, the data are consistent with equal means; if it excludes 0, there is evidence of a difference in that direction. Interpret in context.

Setting Up a Test for a Difference of Means

Hypotheses: $H_0 : \mu_1 = \mu_2$ versus $H_a : \mu_1 \neq \mu_2$ (or $<, >$). Distinguish **two independent samples** from **paired data** 配对数据 – for paired data (before/after, matched subjects), first take the **differences** and run a **one-sample** t procedure on them.

Carrying Out a Test for a Difference of Means

The two-sample t statistic:

$$t = \frac{(\bar{x}_1 - \bar{x}_2) - 0}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_2^2}{n_2}}}.$$

Get the p-value (technology for df), compare to α, conclude in context.

Selecting and Communicating a Procedure

The hardest exam skill is **choosing** the right procedure: one or two samples? proportion or mean? paired or independent? confidence interval or test? Read the question for what is being estimated or claimed, then name the procedure, check its conditions, carry it out, and communicate the conclusion clearly with numbers and context.

Exam tips

- Use **t-procedures** for means (population σ unknown) — the t-distribution has heavier tails than normal.
- Check conditions: random, independent, and roughly normal (or large n).
- Interpret an interval and a test in context, always tied to the parameter (the true mean).
- Match the right procedure: one-sample, two-sample, or paired (look for a natural pairing).
- State the degrees of freedom; for a two-sample t-test use technology's value (or, by hand, the conservative smaller $n - 1$).

Inference for Categorical Data: Chi-Square

AP Statistics

Are My Results Unexpected?

When data are **counts** spread across several categories, we test whether the observed counts differ from what a claim predicts. The tool is the **chi-square** 卡方 (χ^2) statistic, which adds up the standardized gaps between observed and expected counts:

$$\chi^2 = \sum \frac{(\text{observed} - \text{expected})^2}{\text{expected}}.$$

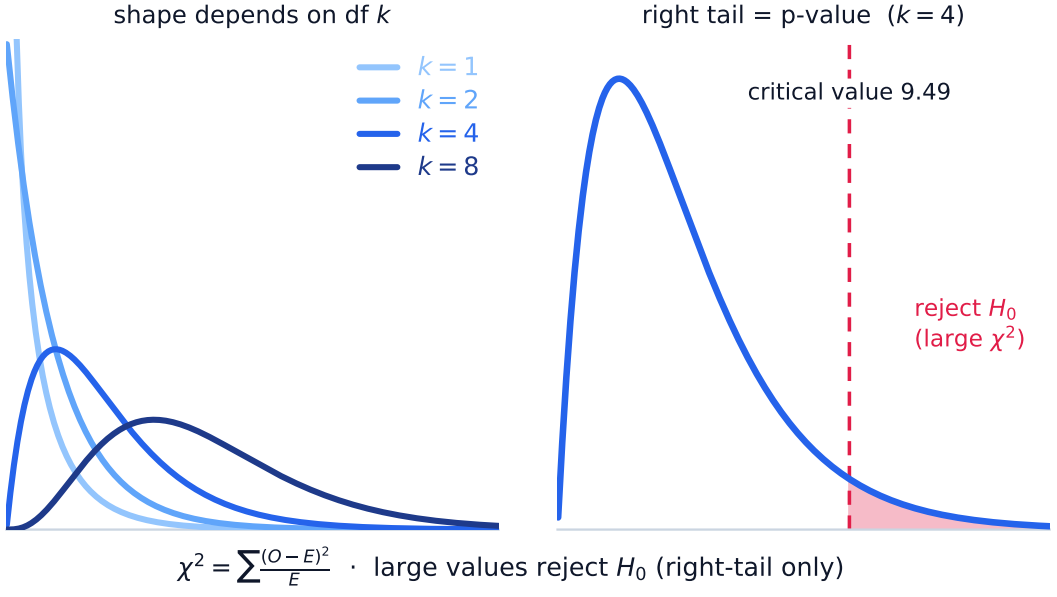
A large χ^2 means the observed counts are far from expected – evidence against the claim. The **chi-square distribution** is right-skewed and depends on its **degrees of freedom** 自由度.

Setting Up a Goodness-of-Fit Test

A **goodness-of-fit (GOF)** 拟合优度 test checks whether one categorical variable follows a **claimed distribution** (e.g. "the die is fair"). Hypotheses:

H_0 : the distribution is as claimed $\qquad$ H_a : at least one proportion differs.

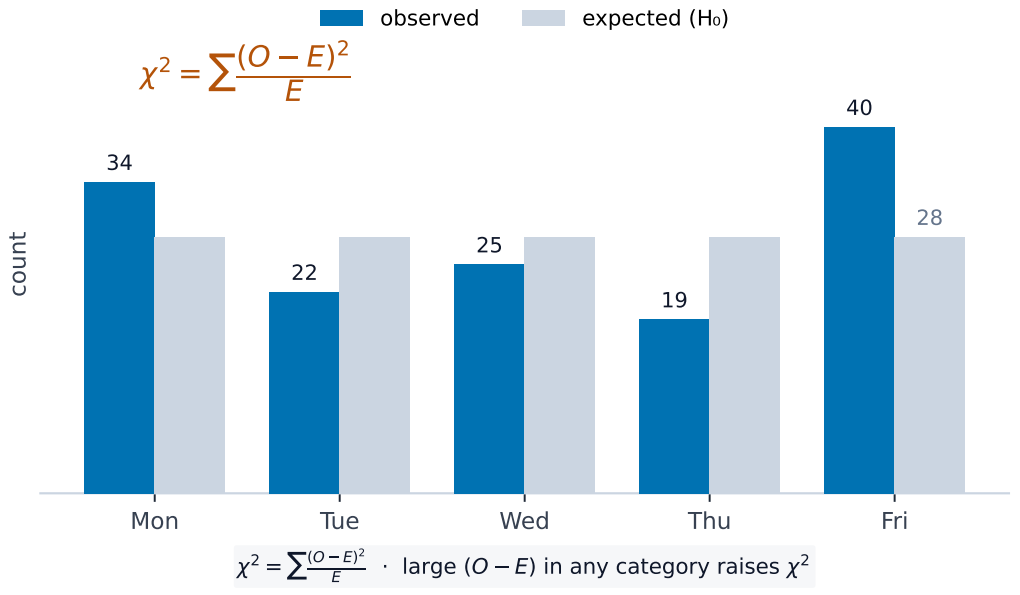
Expected count for each category $= n \times$ (claimed proportion). Conditions: random sample, all **expected counts** ≥ 5, and the 10% condition.



The chi-square distribution is right-skewed. A large statistic lands in the shaded right tail past the critical value – that is where you reject the model.

Carrying Out a Goodness-of-Fit Test

Compute $\chi^2 = \sum \dfrac{(O-E)^2}{E}$ with df = (number of categories) $-$ 1. Find the p-value from the chi-square distribution (upper tail), compare to α, and conclude in context. A large component of the sum points to the category that deviates most.



Chi-square compares observed counts with those expected under the null hypothesis

Worked example. A die rolled 60 times gives counts $8, 10, 12, 9, 11, 10$. If it is fair, each expected count is $60/6 = 10$, so

$$\chi^2 = \frac{(8-10)^2}{10} + \frac{(10-10)^2}{10} + \frac{(12-10)^2}{10} + \frac{(9-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(10-10)^2}{10} = 0.4 + 0 + 0.4 + 0.1 + 0.1 + 0 = 1.0,$$

with $df = 6 - 1 = 5$. Write out **every** category, including the two that match their expected count exactly and so add 0 – the sum runs over all six categories, and df counts categories, not just the ones that differ. This χ^2 is small (a large p-value), so we **fail to reject** H_0 – no evidence the die is unfair.

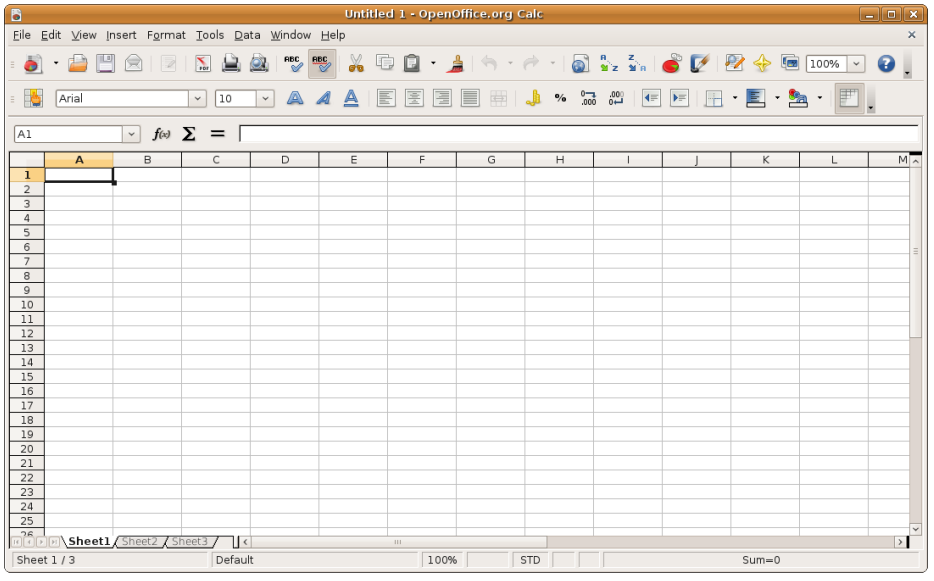
Expected Counts in Two-Way Tables

For a two-way table, the **expected count** in a cell (under "no association") is

$$E = \frac{(\text{row total}) \times (\text{column total})}{\text{grand total}}.$$

This is the count you would see if the row and column variables were unrelated.

Worked example. In a two-way table a cell's row total is 40, its column total is 50, and the grand total is 200. Its expected count is $E = \dfrac{40 \times 50}{200} = 10$. Repeating for every cell gives the expected table to compare against the observed one.



A spreadsheet organises categorical counts before a chi-square test

Homogeneity or Independence?

Two tests use the same χ^2 math but answer different questions:

- **Test for homogeneity** 同质性: are the distributions of one categorical variable the **same across several populations or groups** (separate samples/treatments)?
- **Test for independence** 独立性: are two categorical variables **associated within a single population** (one sample, two variables measured)?

The design (several samples vs one sample) decides which name and hypotheses to use.

Carrying Out a Test for Homogeneity or Independence

Compute expected counts, then $\chi^2 = \sum \dfrac{(O-E)^2}{E}$ over all cells, with

$$df = (\text{rows} - 1)(\text{columns} - 1).$$

Conditions: random data, all expected counts ≥ 5, 10% condition. Find the p-value, compare to α, and conclude in context – evidence of a difference between groups (homogeneity) or of an association (independence).

Choosing the Right Categorical Procedure

Decide by the setup: one categorical variable against a claimed distribution $\Rightarrow$ **goodness-of-fit**; one sample cross-classified by two variables $\Rightarrow$ **independence**; several samples/groups compared $\Rightarrow$ **homogeneity**. Comparing just **two** proportions can use either a two-proportion z-test or a chi-square test, but only for a **two-tailed** alternative, where they agree exactly ($\chi^2 = z^2$). A chi-square test is always two-tailed, so it cannot give a directional conclusion: if H_a is one-tailed (say $p_1 > p_2$), use the z-test.

Exam tips

- Use $\chi^2 = \sum \frac{(O-E)^2}{E}$ for **categorical** data; always divide by the **expected** count.
- Pick the right test: goodness-of-fit (one variable), independence, or homogeneity (two-way table).
- Compute expected counts as $\frac{\text{row total} \times \text{column total}}{\text{grand total}}$ and check each is ≥ 5.
- A large χ^2 (small p-value) means observed counts differ from expected by more than chance.
- State degrees of freedom correctly (categories -1, or $(r-1)(c-1)$).

Inference for Quantitative Data: Slopes

AP Statistics

Do Those Points Align?

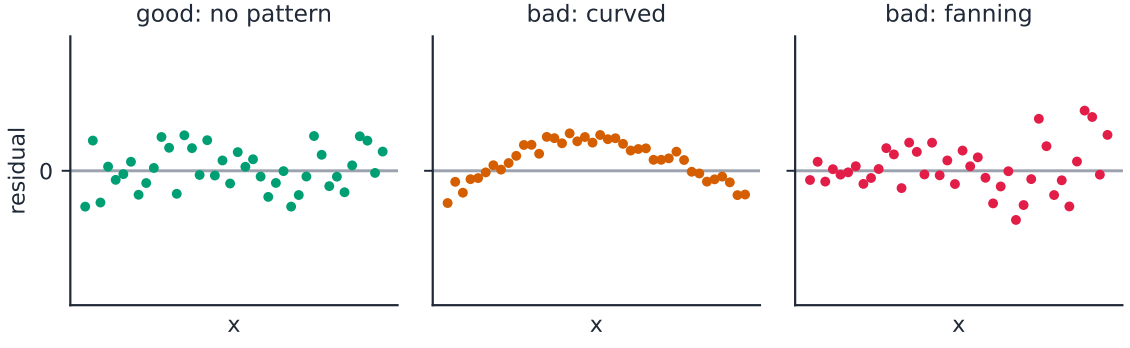
A sample **scatterplot** 散点图 gives a **sample slope** 样本斜率 b for the least-squares **regression** 回归 line – but a different sample would give a slightly different slope. So b is a statistic with **sampling variability** 抽样变异性, estimating the **true (population) slope** 总体斜率 β. This unit does **inference** 推断 for β: is there a real **linear** 线性 relationship, and how strong is it?

Confidence Interval for a Slope

A t interval for the true slope β:

$$b \pm t^* \, SE_b, \qquad df = n - 2,$$

where b is the sample slope and SE_b its **standard error** (read from computer output). Conditions (**LINER**): the true relationship is **Linear**, observations **Independent**, residuals **Normal**, and residuals have **Equal** spread (check the residual plot and a histogram of residuals), from **Random** data. Interpret the interval for β in context, with units of y per unit of x.



residual $= y - \hat{y}$ · random scatter about 0 supports a linear model (LINER)

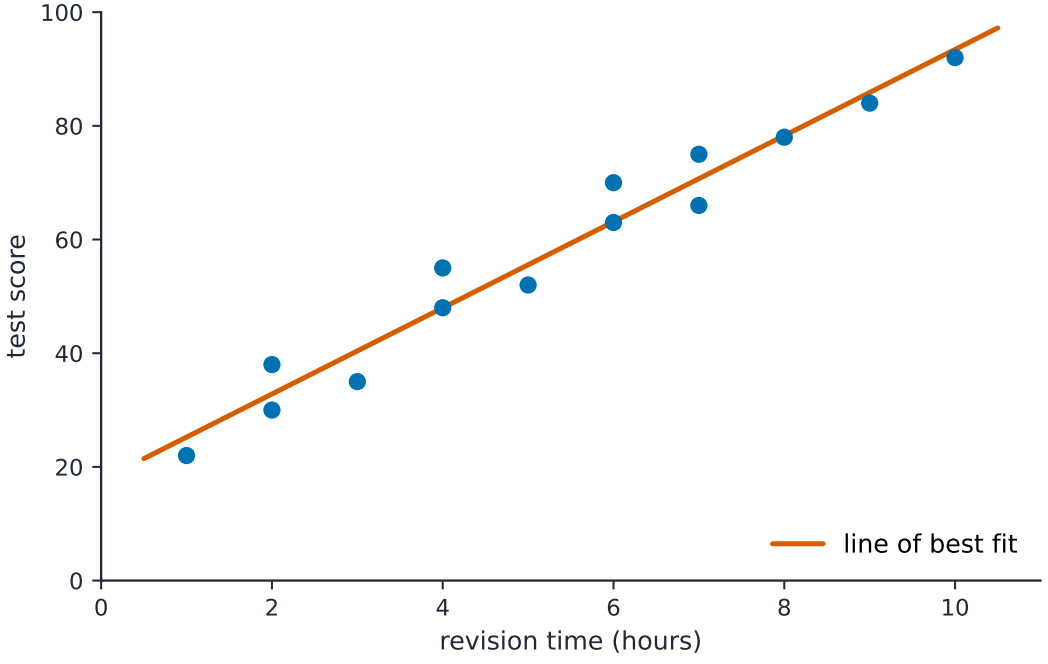
A random, patternless residual plot supports the conditions; a curve or a fan does not

The **residual plot** 残差图 is where you check Linear and Equal-spread: you want a formless cloud around zero. A **curve** means the relationship is not linear; a **fan** (spread growing with x) means the residuals do not have equal spread – both break a condition.

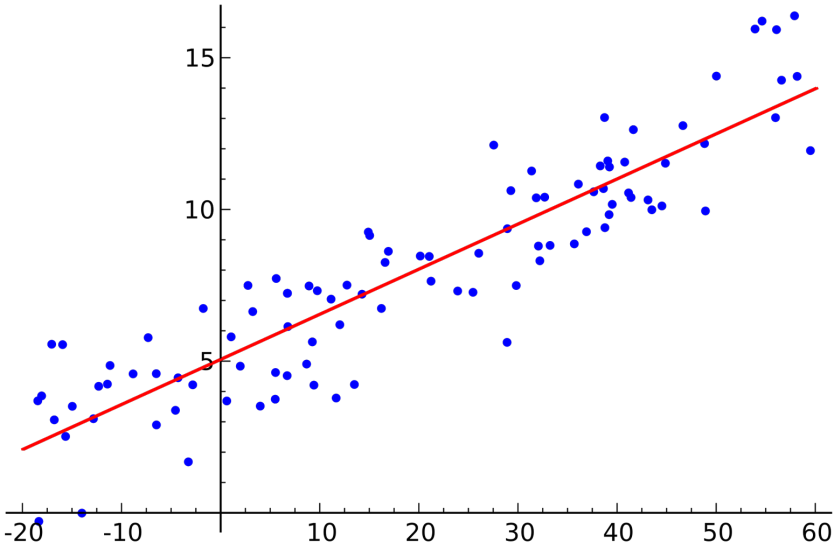
Worked example. Regression output gives slope $b = 2.5$ with $SE_b = 0.8$ from $n = 20$ points. For a 95% interval, $df = 18$ gives $t^* = 2.101$:

$$2.5 \pm 2.101(0.8) = 2.5 \pm 1.68 = (0.82,\ 4.18).$$

Because 0 is **not** in the interval, there is evidence of a positive linear relationship.



Slope inference is based on the least-squares regression line through the points



Least-squares regression: the line that minimises the sum of squared residuals

Justifying a Claim About a Slope

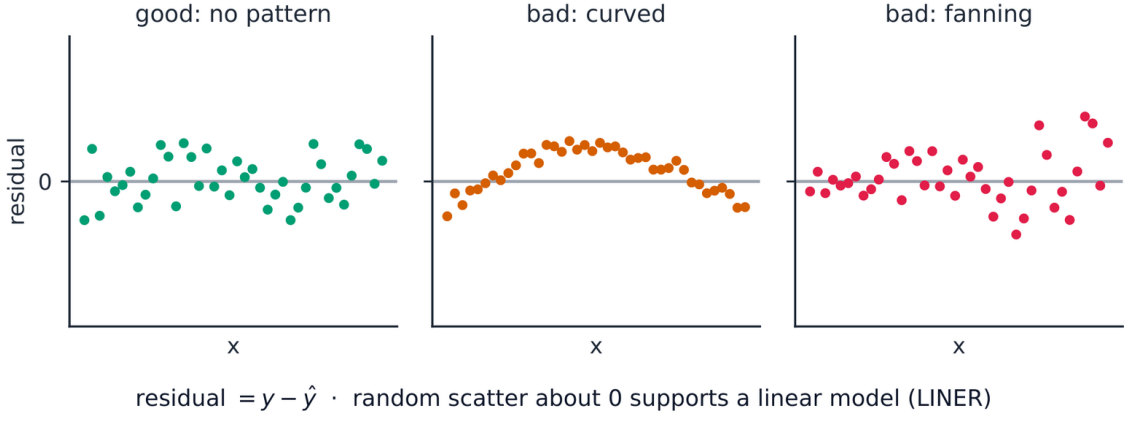
If the confidence interval for β contains 0, a slope of zero is plausible – no evidence of a linear relationship. If the interval is entirely positive or negative, there is evidence of a real (positive or negative) linear relationship. State the direction in context.

Setting Up a Test for a Slope

The usual test asks whether there is any linear relationship:

$$H_0 : \beta = 0 \quad \text{(no linear relationship)} \qquad H_a : \beta \neq 0 \text{ (or } <, >).$$

Check the LINER conditions. This is a t-test on the slope.



Check residual plots before trusting a slope CI or test

Carrying Out a Test for a Slope

The slope t statistic:

$$t = \frac{b - 0}{SE_b}, \qquad df = n - 2.$$

Both b and SE_b come straight from the regression output. Find the p-value from the t-distribution, compare to α, and conclude in context – evidence (or not) of a linear relationship between the two variables.

Watch the tails. Regression output always prints the **two-tailed** p-value (for H_a : $\beta \neq 0$). If your H_a is one-tailed, **halve** it – and first check the sample slope really points the way H_a claims; if it points the other way, the one-tailed p-value is above 0.5 and you cannot reject H_0.

Worked example. For the same output ($b = 2.5$, $SE_b = 0.8$, $n = 20$), test $H_0 : \beta = 0$:

$$t = \frac{2.5 - 0}{0.8} = 3.13, \qquad df = 18,$$

a small p-value (< 0.01), so **reject** H_0 – convincing evidence of a linear relationship. This matches the interval, which excluded 0.

Selecting the Right Procedure

Across all of inference, identify: **what** is estimated or claimed (a proportion, a mean, a difference, a distribution of counts, or a slope), **how many** samples, and **which** design (independent or paired; sample or experiment). Then name the procedure, verify its conditions, carry it out, and communicate the conclusion with the statistic, the p-value or interval, and a plain-language answer in context. This selecting-and-communicating skill is what the investigative-task question rewards most.

Exam tips

- Inference for a **slope** tests whether the true slope is 0 (no linear relationship).
- If a slope's confidence interval **includes 0**, you cannot conclude a real **linear** relationship – the variables may still be related in a curved way.
- Read the slope, standard error, t-statistic, and p-value straight from computer output – but the printed p-value is **two-tailed**, so halve it for a one-tailed H_a.
- Check the regression conditions (linearity, independence, roughly normal residuals, equal spread) via the residual plot.
- Interpret the interval and test in context, tied to the true slope.