Course Companion

AP Statistics

Every topic handout in one volume

全部专题讲义合集

exercises.lol

Contents 目录

1. **Exploring One-Variable Data** 2
2. **Exploring Two-Variable Data** 16
3. **Collecting Data** 27
4. **Probability, Random Variables, and Probability Distributions** 38
5. **Sampling Distributions** 51
6. **Inference for Categorical Data: Proportions** 61
7. **Inference for Quantitative Data: Means** 71
8. **Inference for Categorical Data: Chi-Square** 82
9. **Inference for Quantitative Data: Slopes** 91

Exploring One-Variable Data

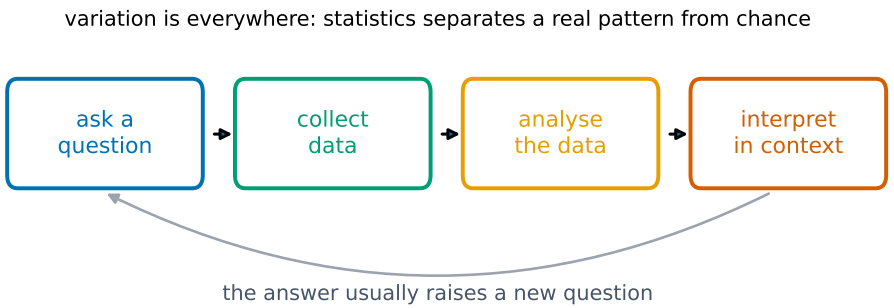
AP Statistics

Introducing Statistics: What Can We Learn from Data?

Two companies both offer you a first job at $36,000 a year. One has a higher median salary; the other has a few very high salaries and a wider spread. Which offer is better for you? A single number cannot say. To answer, you need the whole **distribution**, described with the right summaries and in context. That is what this unit teaches.

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. A number on its own means little: "42" is meaningless, but "42% of students in the sample walk to school" carries information because it has context.

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** 描述统计学 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.



Statistics answers questions by collecting, analysing and interpreting data

The Language of Variation: Variables

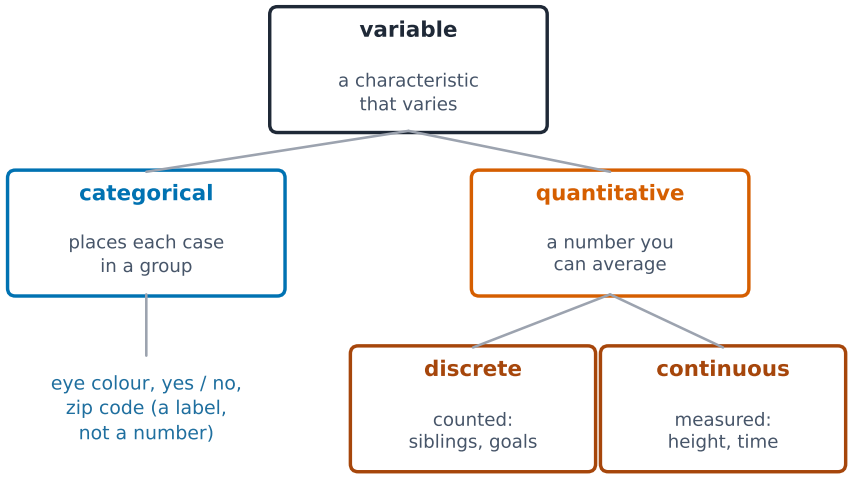
A **variable** 变量 is a characteristic that can differ between individuals. Two kinds:



Every person in a crowd differs in height, age and much else: statistics starts from that variation

- **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.



Categorical variables sort cases into groups; quantitative variables are numbers you can average

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. A proportion, a percentage and a rate all say the same thing: 30 out of 200 is 0.15, 15%, or a rate of 150 per 1000.

> **Worked example.** In a survey, 36 of 120 juniors and 45 of 200 seniors said they drive to school. More seniors drive ($45 > 36$), but that is only because more seniors were asked. Compare proportions: $36/120 = 0.30$ for juniors and $45/200 = 0.225$ for seniors. A **higher proportion** of juniors drive.

transport to school	count	relative frequency
walk	18	18/60 = 0.30
bus	27	27/60 = 0.45
car	12	12/60 = 0.20
bike	3	3/60 = 0.05
total	**60**	**1.00**

relative frequency = count ÷ total

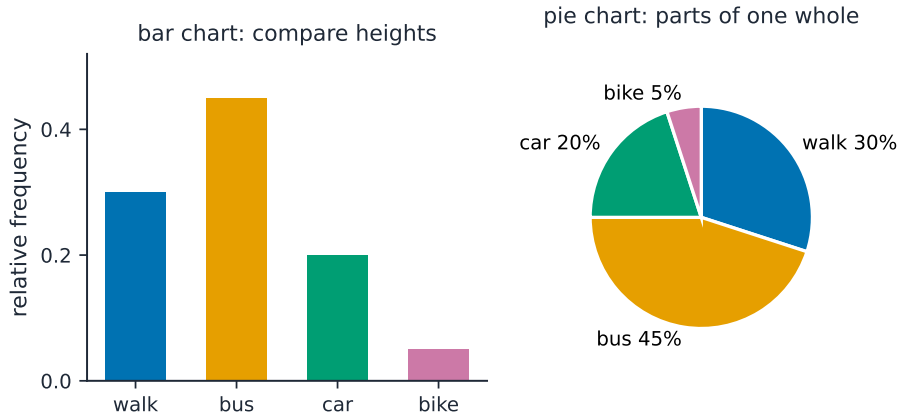
A frequency table lists counts; relative frequency is each count divided by the total

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. On a bar chart the bars do not touch, and the order of the categories carries no meaning unless the categories have a natural order.

To compare two or more groups on the same categorical variable, use relative frequencies, not counts. A **side-by-side bar graph** 并列条形图 puts each group's bars next to each other. A **segmented bar graph** 分段条形图 gives each group one bar of height 1 (100%), divided into segments for the categories. If the segments are about the same size in every bar, the distribution of the variable is similar across the groups.

exercises.lol exercises.com.cn

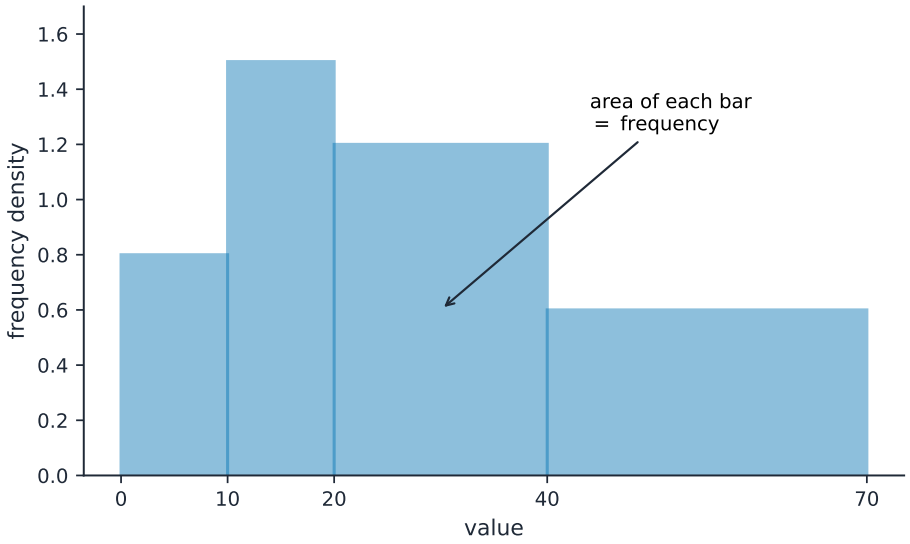


The same data as a bar chart and as a pie chart

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.

A **discrete** 离散 variable takes countable values, such as the number of siblings $(0, 1, 2, \dots)$. A **continuous** 连续 variable can take any value in an interval, such as a time or a mass: between any two values there is always another. A **dotplot** places one dot per observation above a number line, stacking equal values. A **stem-and-leaf plot** splits each value into a stem (leading digits) and a leaf (last digit), so it keeps every data value: the stem 4 with leaves 2 5 5 8 stands for $42, 45, 45, 48$. A cumulative graph plots the proportion of values at or below each value (see below).



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).

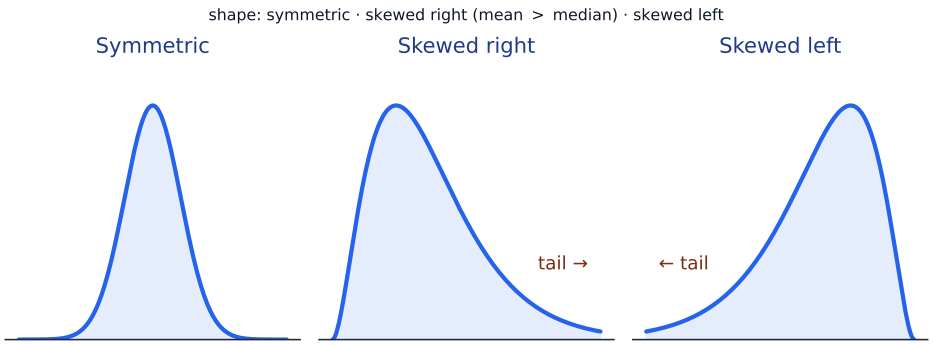
Always describe shape/center/spread **in context**, with units.

Common mistakes. Naming skew by the peak instead of the tail: a long tail to the **left** is skewed left, even though most values pile up on the right. Writing "the data are normal" from a histogram: say "roughly symmetric and unimodal", because real data are only ever approximately normal. Leaving out context: "the median is 7" earns less than "the median commute is 7 minutes".

Also look for **gaps** (stretches of the number line with no data) and clusters (groups of values separated by gaps). These often show that the data mix two kinds of individuals.

exercises.lol exercises.com.cn

Worked example (describing a distribution). A dotplot shows the tips one waiter received on one day: the minimum is \$2, the median is \$4, most tips are between \$2 and \$6, and a long right tail ends at one tip of \$18. A full answer: "The distribution of tip amounts is skewed right, with a median of about \$4. Most tips lie between \$2 and \$6 (a range of about \$16 including the high tip). The \$18 tip appears to be an outlier." It names the shape, the centre, the spread and the unusual feature, all with units. A description like this is **descriptive** only: it is about this waiter's tips on this day, not every day.



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. With an even number of values, the median is the mean of the two middle values.

What changes when one value changes? Suppose the largest tip in a set was \$8 and is changed to \$18. The **mean** goes up by $\$10/n$, and the **standard deviation** increases as this largest value moves farther from the others. The **median** and **IQR** stay unchanged only if the changed value does not enter their calculations; being above the median alone does not guarantee an unchanged upper quartile. The mean, standard deviation and range are **nonresistant**: an extreme value can move them.

exercises.lol exercises.com.cn

Changing units. Adding the same constant to every value (for example, +5 points to every test score) shifts the mean, median and quartiles by that constant, but leaves the range, IQR and standard deviation unchanged. Multiplying every value by a positive constant (for example, converting metres to centimetres, ×100) multiplies all of them: the centre and the spread.

Worked example (combining groups). At one school, 200 teachers have a mean of 9.1 years' experience, and 18 more teachers have a mean of 2.5 years. The mean for all 218 is **not** the mean of 9.1 and 2.5; weight each group by its size:

$$\bar{x} = \frac{200(9.1) + 18(2.5)}{218} = \frac{1820 + 45}{218} = 8.56 \text{ years.}$$

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.

Worked example: standard deviation. For the same data, $\bar{x} = 7$. Work in a table: each deviation from the mean, then its square.

x_i	4	6	7	8	10
$x_i - \bar{x}$	−3	−1	0	1	3
$(x_i - \bar{x})^2$	9	1	0	1	9

The squares add to 20. Divide by $n - 1 = 4$, not by n, then take the square root:

$$s_x = \sqrt{\frac{\sum(x_i - \bar{x})^2}{n-1}} = \sqrt{\frac{20}{4}} = \sqrt{5} \approx 2.24.$$

Interpretation: a typical value is about 2.24 away from the mean of 7. The deviations always add to 0 (here $-3 - 1 + 0 + 1 + 3 = 0$), which is why they are squared first.

Common mistakes. Dividing by $n = 5$ instead of $n - 1 = 4$ (that gives $\sqrt{4} = 2$, not 2.24); stopping at the variance, 5, and reporting it as the standard deviation; and describing s_x as a distance from the median, or as the largest distance from the mean, instead of a typical distance from the mean.

Graphical Representations of Summary Statistics

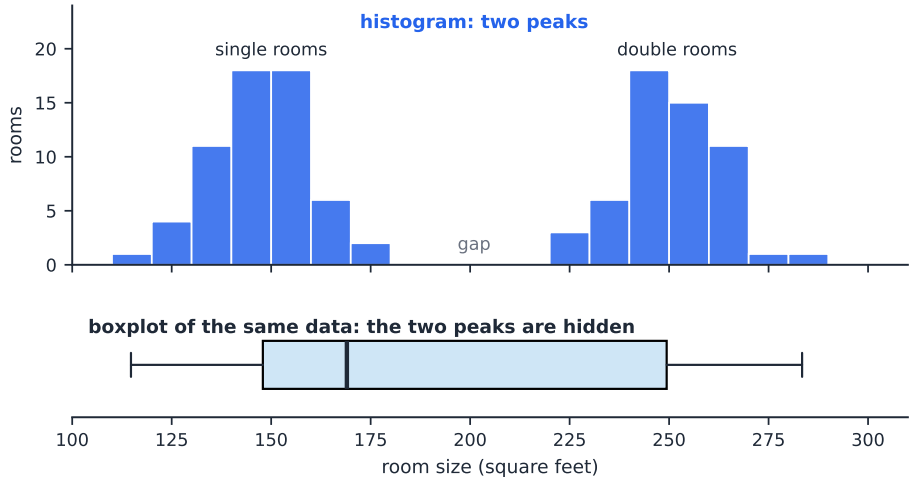
A **boxplot** 箱线图 draws the five-number summary: a box from Q_1 to Q_3 with the median inside, and whiskers to the most extreme non-outlier values. A point is an **outlier** if it lies more than $1.5 \times \text{IQR}$ beyond a quartile – a rule you may be asked to apply. Boxplots are ideal for comparing 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 second rule, also used in this course, flags a value that is **2 or more standard deviations** from the mean. The two rules can disagree. In a strongly skewed data set the standard deviation is inflated by the long tail, so the $1.5 \times$ IQR rule usually flags more points. If a question asks which rule identifies a value as an outlier, compute both fences and check.

Mean, median and shape. In a roughly symmetric distribution the mean and median are close together. In a right-skewed distribution the mean is usually to the **right** of (greater than) the median; in a left-skewed distribution, to the left.

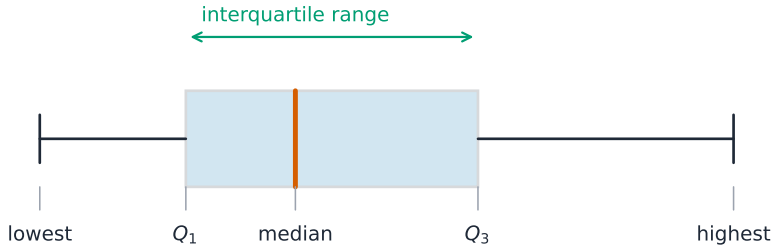
What a boxplot hides. A boxplot summarises quartiles and extremes, with outliers shown separately. It cannot show peaks, gaps or clusters, so two very different distributions can have the same boxplot.



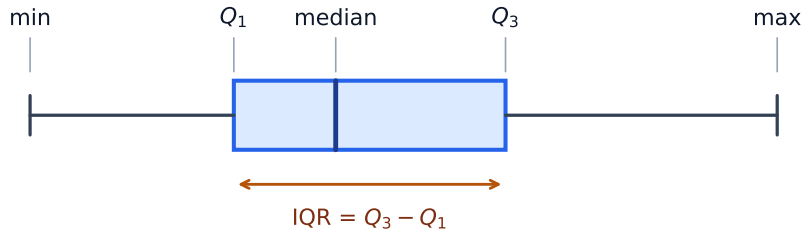
The histogram shows two clusters and a gap; the boxplot of the same data hides them

Common mistakes. Building the fences from the range instead of the IQR; adding $1.5 \times$ IQR to the median instead of to the quartiles; and drawing a whisker out to the fence (50) instead of to the largest value that is **not** an outlier.

exercises.lol exercises.com.cn



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



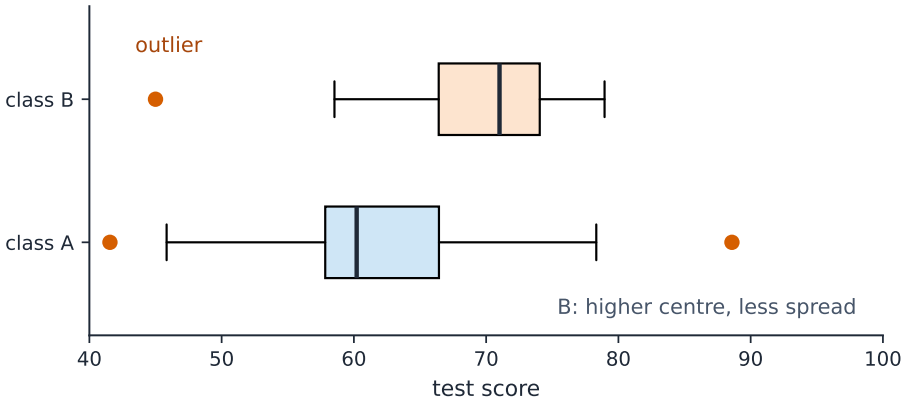
five-number summary: min, Q_1, median, Q_3, max · IQR $= Q_3 - Q_1$

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.

Worked example (comparing in context). Roughly symmetric boxplots show yearly salaries at Corporation A (median \$40,000, IQR \$12,000, no outliers) and Corporation B (median \$34,000, IQR \$20,000, two high outliers). A strong answer: "The median salary at A (\$40,000) is **higher** than at B (\$34,000). Salaries at B are **more variable**: the IQR at B is \$20,000, compared with \$12,000 at A. Both are roughly symmetric, but B has two unusually high salaries." For a job offered at \$36,000, A might appeal because more of its salaries are high (its median is above the offer). B might appeal because a few people earn far more, and \$36,000 is already above B's median.

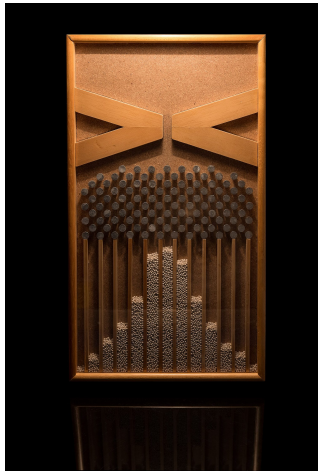


Parallel box plots compare centre, spread and outliers of two groups

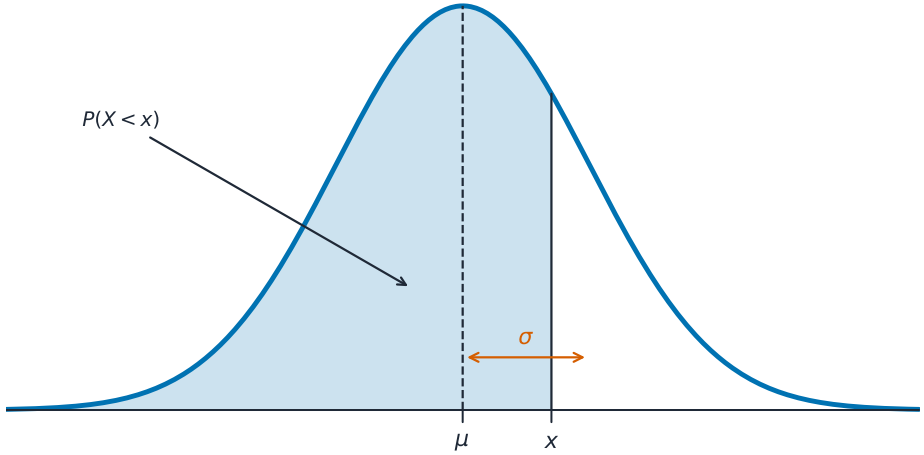
The Normal Distribution

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.

exercises.lol exercises.com.cn



A Galton board: balls bouncing left or right at random pile up into a bell-shaped, approximately Normal distribution



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.

Worked example: between two values, and backwards. With the same $\mu = 500$, $\sigma = 100$:

- Proportion between 450 and 650: $z = \dfrac{450 - 500}{100} = -0.5$ and $z = \dfrac{650 - 500}{100} = 1.5$. The area is $P(z < 1.5) - P(z < -0.5) = 0.9332 - 0.3085 = 0.6247$, so about 62.5% of scores.
- The score at the 90th percentile: the table (or invNorm) gives $z = 1.28$ for an area of 0.90 below. Solve $z = \dfrac{x - \mu}{\sigma}$ for x: $x = \mu + z\sigma = 500 + 1.28(100) \approx 628$.

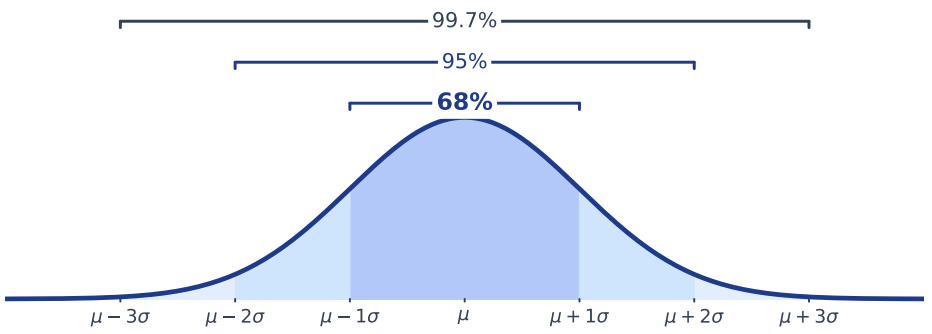
Worked example (a percentile from the model). Battery life is normal with $\mu = 30$ months and $\sigma = 6$ months. After how many months have 25% of batteries failed? That is the 25th percentile: the table gives $z = -0.674$ for an area of 0.25 below, so $x = 30 - 0.674(6) = 26.0$ months.

A z-score also compares values from different distributions. A score of 82 on a test with mean 70 and SD 8 ($z = 1.5$) is relatively better than 90 on a test with mean 80 and SD 10 ($z = 1.0$).

On the exam, show the distribution, the boundary and the area: "Normal, $\mu = 500$, $\sigma = 100$, $P(450 < X < 650) \approx 0.6247$". A bare calculator command does not earn full credit.

Common mistakes. Using the empirical rule for a z-score that is not a whole number ($z = 1.5$ needs the table or technology); giving the area **below** when the question asks for the proportion **above** (subtract from 1); and using a normal model for a distribution the question describes as strongly skewed.

exercises.lol exercises.com.cn



for approx. normal data: ~68% within 1σ, ~95% within 2σ, ~99.7% within 3σ

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

Practice

Answer each before reading the answers below.

1. For the data 3, 5, 7, 8, 9, 12, 15, find the median and the IQR, and check for outliers with the $1.5 \times IQR$ rule.
2. Household incomes in a city are strongly right-skewed. Which is larger, the mean or the median? Which better describes a typical household?
3. A test has mean 70 and standard deviation 8. Find the z-score of a student who scored 82.
4. If the scores in question 3 are approximately normal, what proportion of students scored below 58?
5. A student says: "The mean is always the best measure of center." Correct the student.
6. The 30 values in a data set have mean 20 and standard deviation 4. Every value is multiplied by 3 and then 5 is added. Find the new mean and standard deviation.
7. A class of 12 has a mean score of 78, and another class of 18 has a mean of 68. Find the mean for all 30 students.
8. A data set has $\bar{x} = 50$, $s = 5$, $Q_1 = 46$ and $Q_3 = 53$. Is a value of 61 an outlier by the $1.5 \times$ IQR rule? By the 2-standard-deviation rule?

Answers

1. Median 8. $Q_1 = 5$ and $Q_3 = 12$ (medians of the lower and upper halves), so $IQR = 7$. The fences are $5 - 10.5 = -5.5$ and $12 + 10.5 = 22.5$; no value lies outside them, so there are no outliers.
2. The mean is larger, pulled up by the few very high incomes. The median better describes a typical household, because it resists extreme values.
3. $z = \dfrac{82 - 70}{8} = 1.5$: one and a half standard deviations above the mean.
4. $z = \dfrac{58 - 70}{8} = -1.5$, so the proportion is about 0.067, or 6.7%.
5. The mean is sensitive to outliers and skew. For skewed data or data with outliers, the median is the better summary of a typical value; report the IQR with it.
6. Mean: $3(20) + 5 = 65$. Standard deviation: $3(4) = 12$; adding 5 does not change the spread.
7. $\dfrac{12(78) + 18(68)}{30} = \dfrac{936 + 1224}{30} = 72.$
8. IQR $= 7$, so the upper fence is $53 + 10.5 = 63.5$: 61 is **not** an outlier by that rule. It is $\dfrac{61 - 50}{5} = 2.2$ standard deviations above the mean, so it **is** an outlier by the 2-SD 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.
- When one value changes, say which statistics change and why: the mean and SD move; the median and IQR stay unless the middle of the data changes.
- Compare groups with **proportions**, not counts, when the groups differ in size.

exercises.lol exercises.com.cn

Exploring Two-Variable Data

AP Statistics

Are Two Variables Related?

Do students who take more semesters to graduate earn higher starting salaries? Do children choose fruit more often after a healthy-food advert? A newspaper may claim a link from a handful of people. This unit gives you the tools to check: tables and graphs to see a pattern, r to measure it, and a regression line to predict with it.

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. Also, a pattern in a sample may be **random**: with few individuals, chance alone can produce an apparent association that would vanish in more data. Units 4 to 9 show how to judge whether a pattern is larger than chance would produce.

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. A **joint relative frequency** 联合相对频率 is one cell divided by the grand total: the proportion of **all** individuals who are in that row **and** that column.

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. A **mosaic plot** 马赛克图 is a segmented bar graph whose bar **widths** are proportional to the group sizes, so it shows the marginal and the conditional distributions at once.

> **Worked example.** A school records how 200 students took a course and whether they passed.

	Passed	Failed	Total
Online	60	40	100
In person	75	25	100
Total	135	65	200

- Marginal distribution of result: $\frac{135}{200} = 0.675$ passed and $\frac{65}{200} = 0.325$ failed.
- Conditional distributions: online students passed at $\frac{60}{100} = 0.60$; in-person students at $\frac{75}{100} = 0.75$.

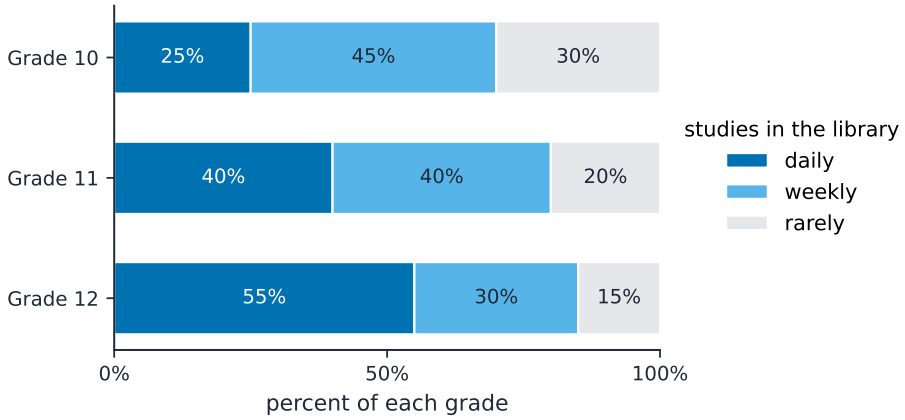
The conditional pass rates differ (0.60 vs 0.75), so the format and the result are **associated** in these data. This alone does not show that the format *causes* the difference: students chose their format.

Divide by the correct total: "the proportion of online students who passed" uses the online total (100), while "the proportion of passes that were online" uses the pass total: $\frac{60}{135} \approx 0.44$.

Three proportions from the table above, for the same cell:

- joint: $\frac{60}{200} = 0.30$ of **all** students were online and passed;
- conditional on format: $\frac{60}{100} = 0.60$ of online students passed;
- conditional on result: $\frac{60}{135} = 0.44$ of students who passed were online.

When the FRQ asks you to "describe the effect" of a factor, compare **conditional** proportions across the groups and give the numbers.



A segmented bar chart compares the conditional distribution in each group

Scatterplots for Two Quantitative Variables

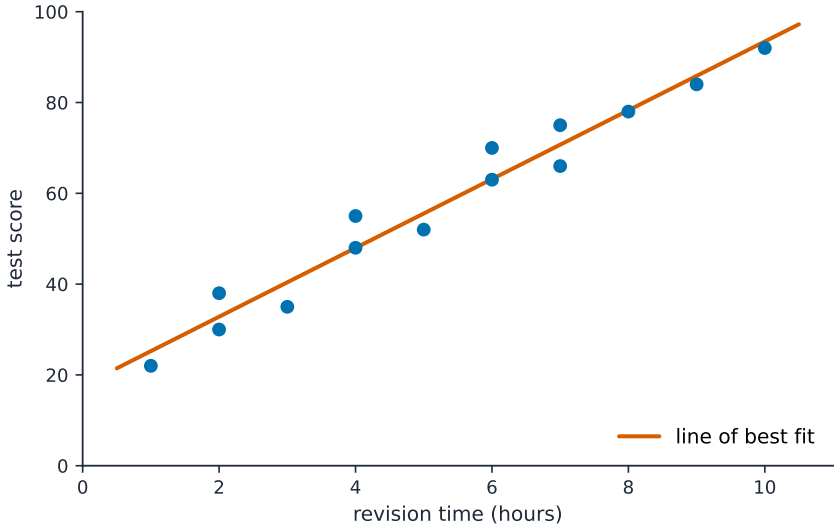
A **scatterplot** 散点图 plots each individual as a point, explanatory variable on the x-axis and response on the y-axis. Describe it with **DUFS**: **Direction** (positive/negative), **Unusual** features (outliers, clusters), **Form** (linear or curved), and **Strength** (how tightly the points follow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

exercises.lol exercises.com.cn



Old Faithful: a scatterplot of eruption length against the wait for the next one shows two clusters and a strong positive association

Worked example. A scatterplot shows height and arm span for 12 seniors. A full description: "There is a **strong, positive, linear** association between height and arm span for these seniors: taller seniors tend to have longer arm spans. There are no obvious outliers." Use "tend to": the pattern describes the group, not every pair of points.



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

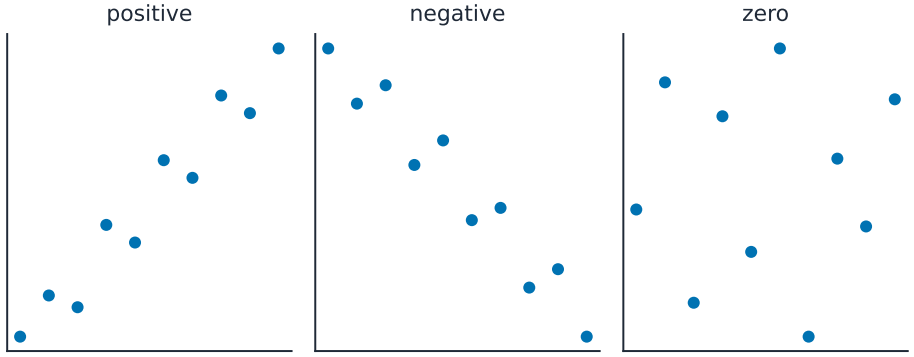
Correlation

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.

How r is built. Change each x and y to a z-score, multiply the pairs, and average (dividing by $n-1$):

$$r = \frac{1}{n-1}\sum\left(\frac{x_i-\bar{x}}{s_x}\right)\left(\frac{y_i-\bar{y}}{s_y}\right).$$

Points in the upper-right and lower-left (both z-scores of the same sign) add positive products; points in the other two corners add negative ones. Because it uses z-scores, r does not change if you change units (inches to centimetres) or add a constant to every value. $r = \pm 1$ only when all points lie exactly on a line. A value of r near ± 1 does **not** by itself show that a line is the right model: a smooth curve can also have $r = 0.95$.



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** 纵截距 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).

From summary statistics, the line comes from two formulas: the slope is $b = r\,\frac{s_y}{s_x}$, and the line always passes through $(\bar{x}, \bar{y})$, so $a = \bar{y} - b\bar{x}$.

exercises.lol exercises.com.cn

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 $\hat{y} = 20 + 3(5) = 35$. The intercept says a student who studies 0 hours is predicted to score 20. Sometimes the intercept has no sensible meaning: a car of length 0 inches, or a house of 0 square metres, does not exist.

Worked example: line from summary statistics. For another class, hours studied have $\bar{x} = 6$ and $s_x = 2$; scores have $\bar{y} = 70$ and $s_y = 10$; and $r = 0.8$.

$$b = r\,\frac{s_y}{s_x} = 0.8 \times \frac{10}{2} = 4, \qquad a = \bar{y} - b\bar{x} = 70 - 4(6) = 46.$$

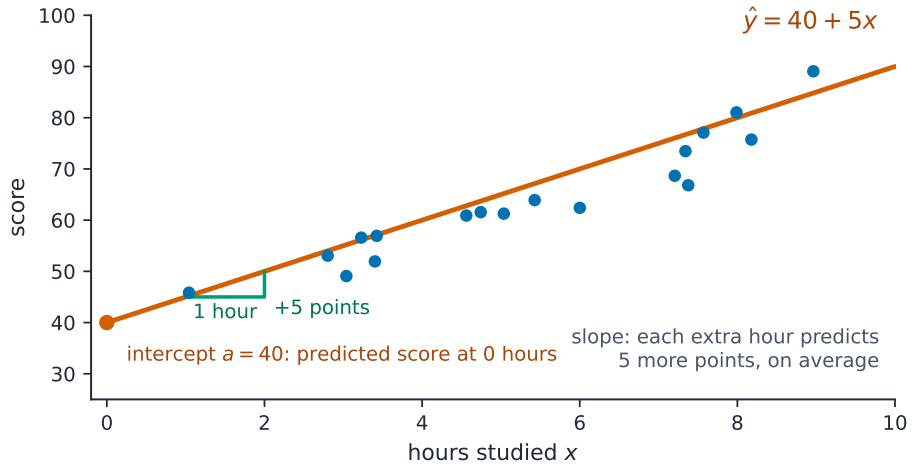
So $\hat{y} = 46 + 4x$: each extra hour is associated with a predicted 4-point increase. Also $r^2 = 0.64$: 64% of the variation in scores is explained by the linear relationship with hours studied.

Reading computer output. Exams often give regression output instead of an equation. This illustrative output is for 18 graduates.

Predictor	Coef	SE Coef	T	P
Constant	12.40	3.10	4.00	0.001
Semesters	1.85	0.42	4.40	0.000

$S = 3.61$ R-Sq $= 54.8\%$

The **Constant** row gives the intercept and the row named after the explanatory variable gives the slope: $\widehat{\text{salary}} = 12.40 + 1.85(\text{semesters})$, in thousands of dollars. Interpretation of the slope: "For each extra semester, the predicted starting salary increases by about \$1850." S is the standard deviation of the residuals: predictions are typically off by about \$3610. R-Sq is r^2. Since the slope is positive, $r = +\sqrt{0.548} \approx 0.74$. (The SE, T and P columns are for inference in unit 9.)



Reading $\hat{y} = a + bx$: the slope is the predicted change per unit of x; the intercept is the prediction at $x = 0$

Residuals

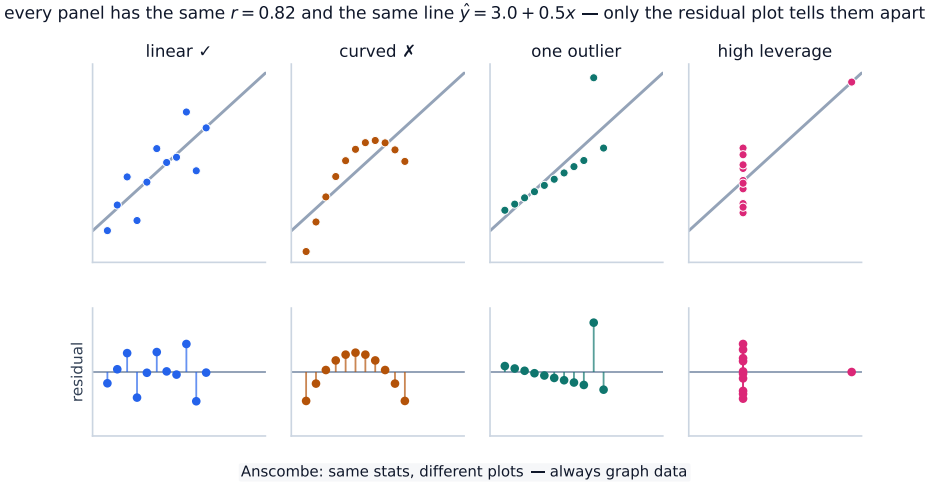
A **residual** 残差 is actual minus predicted, $y - \hat{y}$: how far a point sits above (+) or below (−) the line. A **residual plot** 残差图 graphs residuals against x. If it shows **no pattern** (random scatter), it supports a linear model; a curve suggests the wrong mean model, while a fan indicates changing residual spread.

> **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. In context: "The student scored 5 points **higher** than the line predicts for a student who studied 5 hours." A residual close to 0 means the line predicted that individual's response almost exactly.

A residual plot can put the residuals against x or against the predicted values $\hat{y}$; both are read the same way. Look for curves (wrong form) and a fan shape (the spread of the residuals grows with x, so predictions are less reliable for large x).

Common mistakes. Computing predicted minus actual ($35 - 40 = -5$): a residual is always **actual minus predicted**. Stating the slope as a certainty ("the score goes up by 3") instead of a **predicted** change, or without units and context. Interpreting the intercept when $x = 0$ lies far outside the data.

exercises.lol exercises.com.cn

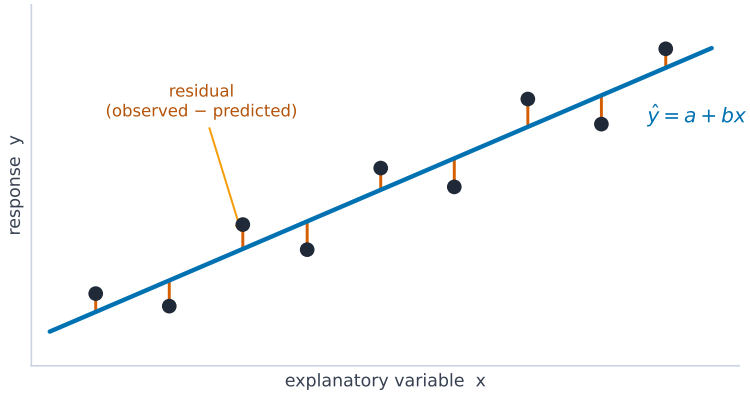


*A caution about r and the line: all four datasets have the same $r = 0.82$ and the same $\hat{y} = 3.0 + 0.5x$, yet only the first is genuinely linear. The scatterplots differ sharply; the **residual plot** below each also exposes the curve, the outlier, and the high-leverage point.*

Least-Squares Regression and Its Fit

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."



residual $= y - \hat{y}$ · least squares minimises $\sum(y-\hat{y})^2$

The least-squares line minimizes the sum of squared residuals

Common mistakes. Saying $r = 0.8$ means "80% of the variation is explained": that statement belongs to $r^2 = 0.64$. Calling $r \approx 0$ "no relationship": a strong curve can have r near 0. Saying a high r or r^2 shows a linear model is right: only the residual plot shows that.

Departures from Linearity

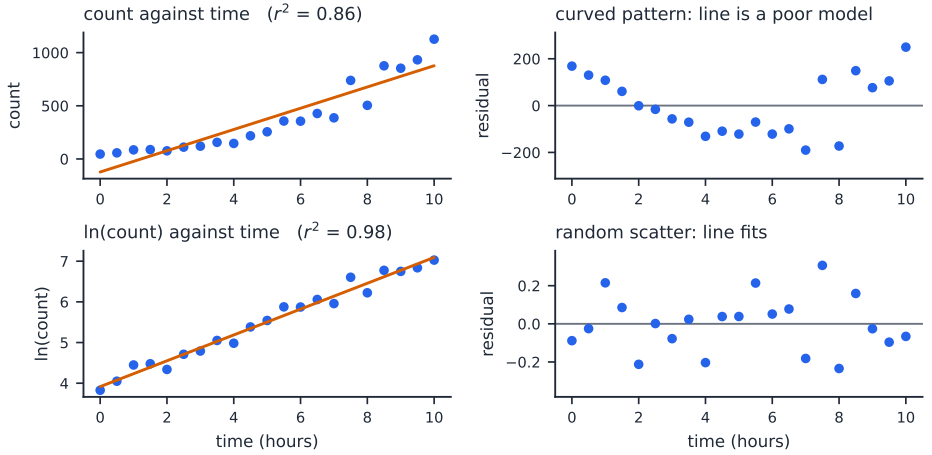
Extension: this topic was removed from the AP Statistics exam curriculum effective 2026–27. See College Board's course revisions.

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.

Worked example (an influential point). Ten countries give a weak positive scatter with $r = 0.30$. An eleventh country has an x-value far beyond the others, and it lies on the upward trend. Adding it pulls the line toward it and raises r to 0.75. This point has **high leverage** (extreme x) and is **influential** (it changes the slope and r a lot), even though its residual is small. A point with a large residual but an ordinary x-value is an **outlier**, and it usually changes the slope much less.

exercises.lol exercises.com.cn

Worked example (a transformation). A bacteria count grows quickly with time. The straight line for count against time has $r^2 = 0.86$, but its residual plot is a clear U-shape, so the linear model is wrong. Taking ln(count) gives a residual plot with no pattern and $r^2 = 0.98$. The patternless transformed residuals support the log-linear model. The two r^2 values describe different response scales, so they cannot directly compare prediction errors in counts. Predict with it, then undo the log: if $\widehat{\ln(\text{count})} = 3.9 + 0.32t$, then at $t = 5$, $\widehat{\ln(\text{count})} = 5.5$ and the predicted count is $e^{5.5} \approx 245$.



A log transformation turns a curved pattern into a straight one: compare the residual plots, not just r^2

Practice

Answer each before reading the answers below.

1. Of 120 students, 50 play a sport, and 30 of those 50 are also in a music group. What proportion of sport players are in a music group?
2. A scatterplot of hours slept against reaction time gives $r = -0.85$. Describe the relationship.
3. A regression line is $\hat{y} = 12 + 2.5x$. Predict y when $x = 8$, and find the residual for a point $(8,\ 29)$.
4. The data in question 3 have x values from 0 to 10. Why is predicting y at $x = 40$ risky?
5. A student says: "The correlation is 0.9, so x causes y." Correct the student.

6. Computer output for predicting fuel use (litres per 100 km) from car length (inches) shows Constant $= -2.50$, Length $= 0.0478$, R-Sq $= 49.0\%$. Write the equation, and find and interpret the residual for a car 175 inches long that uses 5.88 litres per 100 km.
7. A data set has $r = 0.6$. Every x-value is converted from inches to centimetres. What happens to r and to the slope?
8. In a two-way table of 400 voters, 90 are under 30 and support a proposal, and 150 are under 30 in total. Find the joint relative frequency of "under 30 and supports", and the proportion of under-30 voters who support.

Answers

1. $30/50 = 0.60$. This is a **conditional** proportion: it uses only the sport players as the whole.
2. A strong, negative, linear relationship: students who sleep more tend to have shorter reaction times.
3. $\hat{y} = 12 + 2.5(8) = 32$. Residual $= y - \hat{y} = 29 - 32 = -3$; the point lies 3 units below the line.
4. That is **extrapolation**: the pattern is known only for x from 0 to 10, and nothing guarantees it continues to 40.
5. Correlation measures the strength of a linear association, not causation. A lurking variable could drive both, or the link could run the other way. Only a randomized experiment supports a cause-and-effect claim.
6. $\hat{y} = -2.50 + 0.0478x$. At $x = 175$: $\hat{y} = -2.50 + 8.365 = 5.865$. Residual $= 5.88 - 5.865 = 0.015$: the car uses only about 0.015 litre per 100 km more than predicted, so the line predicts it almost exactly.
7. r stays 0.6 (r has no units). The slope is divided by 2.54, because each centimetre is a smaller step in x than each inch.
8. Joint: $\frac{90}{400} = 0.225$. Conditional: $\frac{90}{150} = 0.60$.

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.

exercises.lol exercises.com.cn

- Read the slope from the row named after the explanatory variable in computer output, and the intercept from **Constant**; r takes the sign of the slope.
- Choose between models by the **residual plot** first, then r^2.

Collecting Data

AP Statistics

Can We Trust the Data We Collected?

A website asks visitors "Do you like this website?" and 92% click Yes. A survey at a gym finds that most people exercise daily. Both numbers are real, and both are nearly useless, because of **who** was asked. This unit is about the decisions made before any data exist, which decide what the data can ever tell you.

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.

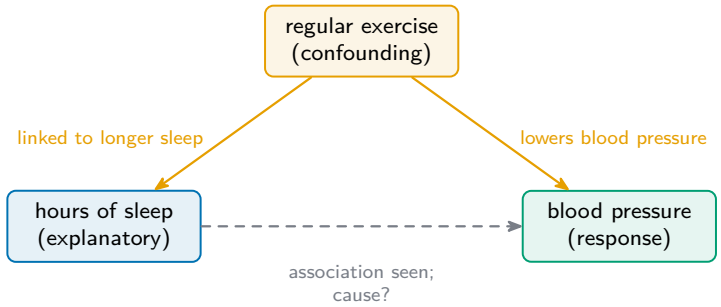
Observational Studies and Experiments

- 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.

An observational study can look back at data that already exist (**retrospective**, such as old medical records) or follow a group forward in time and collect data as things happen (**prospective**). Either way, nothing is imposed, so neither kind can show causation. Results can be generalized only to the population the sample was randomly chosen from: a random sample of one school's students says nothing about students at other schools.

> **Worked example (observational or experiment?).** An article reports that adults who sleep at least 7 hours a night have lower blood pressure. The researchers asked people how much they sleep and measured their blood pressure. The explanatory variable is hours of sleep; the response is blood pressure. It is an **observational study**: no one was told how much to sleep. Exercise status could be a **confounding variable** 混杂变量: people who exercise regularly may tend to sleep longer, **and** exercise lowers blood pressure, so the study cannot tell whether sleep or exercise causes the lower readings. A full answer always gives **both** links: to the explanatory variable and to the response.

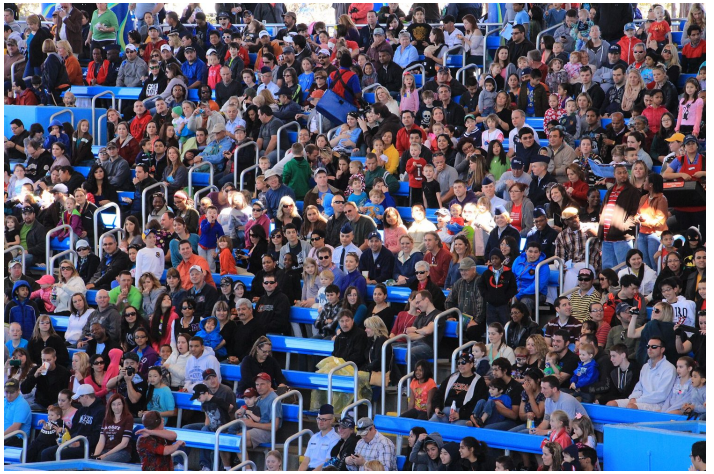
exercises.lol exercises.com.cn



A confounding variable needs both links: exercise goes with longer sleep and also lowers blood pressure, so the study cannot separate the two

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:

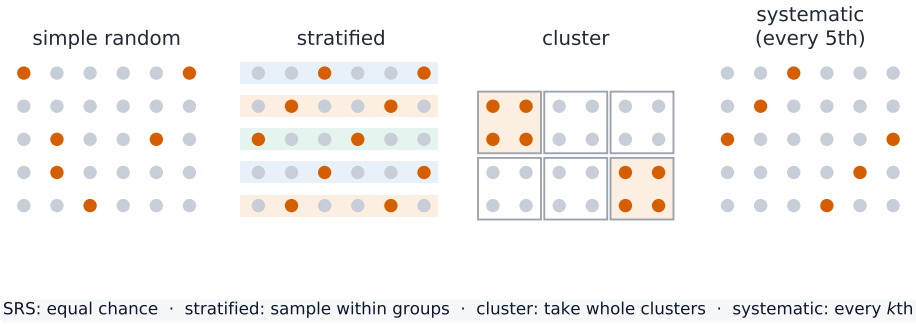


We cannot ask everyone in a population, so we study a sample, chosen at random so it represents the whole

- **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.

A **census** 普查 measures every individual in the population, which is usually too slow or expensive. Most samples are drawn **without replacement** (no one can be chosen twice); drawing **with replacement** puts each chosen individual back, so it can be picked again.

Strengths and weaknesses. A stratified sample gives more precise estimates than an SRS of the same size when the strata really differ in the response (for example, juniors and seniors answer differently), and it guarantees every stratum is represented. A cluster sample is cheaper and quicker (survey whole classrooms or whole streets), but if the clusters differ from one another, the estimate is less precise. A systematic sample is easy to carry out, but it fails if the list has a repeating pattern. If a factory's two machines alternate on a conveyor belt, "every 10th tortilla" always comes from the same machine, so the sample does not represent the day's production.



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.

Common mistakes. Mixing up **stratified** and **cluster** samples. Stratified: split into groups of similar individuals and take a random sample from **every** group. Cluster: split into groups that each look like the whole population, randomly choose **some** groups, and survey everyone in them. Also, calling a sample random when it is only convenient ("the first 50 students in the canteen").

exercises.lol exercises.com.cn



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

When Sampling Goes Wrong

Bias makes estimates systematically miss the truth:

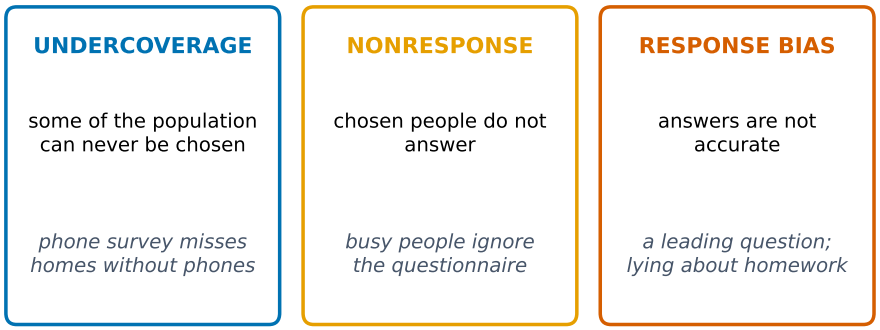
- **Undercoverage** 覆盖不足: some groups are left out of the sampling frame.
- **Nonresponse** 无回应: selected people do not answer.
- **Response bias** 回应偏差: people answer inaccurately (bad wording, sensitive topics).

Response bias has two common forms. **Question wording bias** 问题措辞偏差 comes from confusing or leading questions ("Don't you agree that the unfair new dress code should be dropped?"). **Self-reporting** is unreliable when people forget or want to look good: asked how many miles they walk, people tend to overstate it, so an activity tracker gives more accurate data than a journal.

> **Worked example (bias and its direction).** An environmental science teacher asks only the students in her own classes whether the school should install solar panels. Her students have studied renewable energy and may be more likely to say yes than the school as a whole. The sample proportion who answer yes will probably **overestimate** the true proportion for all students.

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

Common mistakes. Saying a larger sample reduces bias: it reduces **variability** only. Naming a bias without saying, in context, which way it pushes the estimate (too high or too low) and why.



Undercoverage, nonresponse and response bias

Designing an Experiment

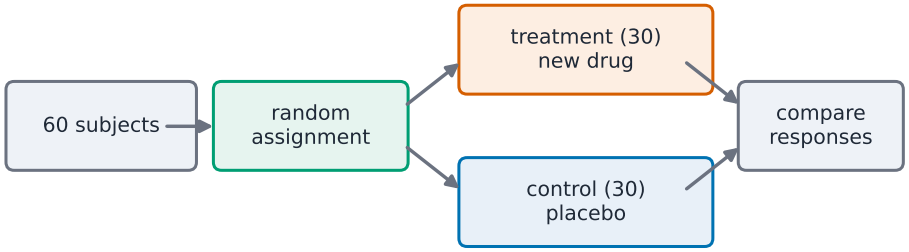
Good experiments follow three principles:



Tablets on a plate: a placebo is made to look like the real drug, so subjects cannot tell which treatment they got (blinding)

- **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.

exercises.lol exercises.com.cn



random assignment → causal comparison of responses

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

The vocabulary. The **experimental units** 实验单元 are the individuals that receive treatments (called **subjects** or **participants** when they are people). The explanatory variable in an experiment is a **factor** 因子, and its values are **levels**; each level or combination of levels is a **treatment**. The **response variable** is what is measured afterwards.

> **Worked example.** A gardener tests three fertilisers on 30 tomato plants and measures each plant's mass of fruit. Treatments: the three fertilisers. Experimental units: the 30 plants. Response: mass of fruit. With two factors, such as fertiliser (3 levels) and watering (2 levels), there are $3 \times 2 = 6$ treatments.

Is there a control group? A **control group** gets no active treatment, or a **placebo** that looks like the real treatment, so that the treatment effect can be separated from everything else. The **placebo effect** is a real response to an inactive treatment, caused by expecting to improve. An experiment comparing two new drugs has **no** control group, but it can still compare the two drugs with each other.

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, reducing differences in expectations between groups and biased assessment; placebo responses can still occur in both groups. **Blocking** 区组 groups similar subjects and randomizes within each block to reduce variability.

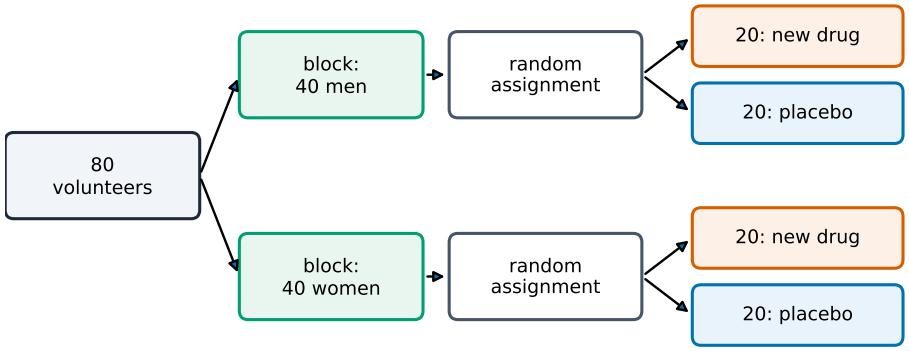
Worked example: describing random assignment. Forty volunteers will test two study apps, A and B, 20 per app. A full-credit description names the method precisely enough that someone else could carry it out:

> *Number the volunteers 1 to 40. Use a random number generator to choose integers from 1 to 40, ignoring repeats, until 20 different numbers are chosen. Those 20 volunteers use app A; the remaining 20 use app B.*

"Randomly put them into two groups" earns nothing: it does not say **how**. Drawing names from a hat also works, if you say the slips are the same size and well mixed.

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.



A randomized block design: group similar subjects, then randomize within each block

Worked example: blocking. The study-app volunteers are 24 Grade 10 and 16 Grade 11 students, and grade level surely affects test scores. A **randomized block design** 随机区组设计 separates the two grades first. Within each grade, randomly assign half to app A and half to app B: 12 and 12 in Grade 10, 8 and 8 in Grade 11. Then compare the apps within each grade. Grade differences can no longer hide, or exaggerate, the effect of the app.

Worked example (matched pairs with twins). A dermatologist tests two acne creams on 10 pairs of twins. Each twin's acne is similar to the other twin's, but severity differs a lot between pairs. In a **matched-pairs design** 配对设计, for each pair, flip a coin: heads, the first-listed twin gets cream A and the other gets cream B; tails, the reverse. Then compare the two twins in each pair. Because twins start alike, a difference within a pair is mainly due to the cream, not to how severe the acne was to begin with. A single person can also be their own pair: each subject uses both treatments, in a random order.

What an Experiment Lets You Conclude

Two questions decide the scope of a conclusion:

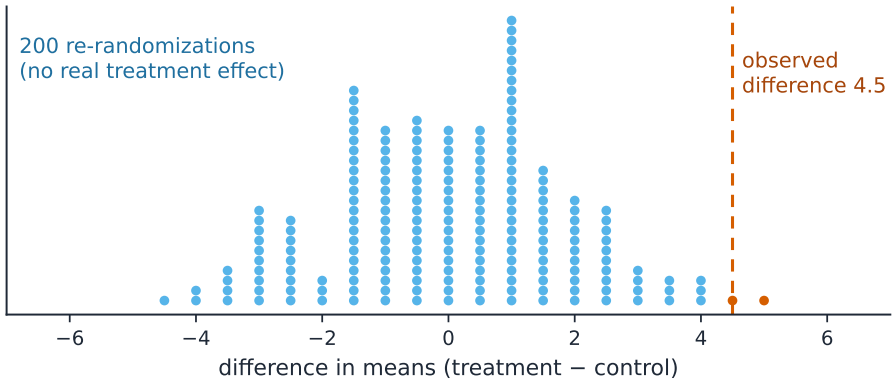
- **Random assignment** used? Then a significant difference can be attributed to the treatment (**causation**) – for these subjects.
- **Random sampling** from a population? Then results **generalize** to that population.

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

Statistically significant 统计显著 means the difference between the treatment groups is too large to be explained easily by the chance variation of random assignment alone. In a well-designed experiment, a statistically significant difference is evidence that the treatments **caused** the difference.

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.

exercises.lol exercises.com.cn



chance rarely gives a difference this large, so the result is statistically significant

If random assignment alone rarely produces a difference this large, the result is statistically significant

The scope of a conclusion

	Random assignment	**No random assignment**
Random sampling	cause and effect, for the whole population	association only, for the whole population
No random sampling	cause and effect, for these subjects only	association only, for these subjects only

Read the design once for each question: was the **sample** chosen at random (generalize?), and were the **treatments** assigned at random (causation?).

Common mistakes. Claiming cause and effect from an observational study; generalizing an experiment on volunteers to a whole population; and confusing random **sampling** (who is in the study) with random **assignment** (who gets which treatment).

Practice

Answer each before reading the answers below.

1. A teacher surveys the first 30 students to arrive at school. Name the sampling method and one likely bias.
2. A school has 900 students in 30 homerooms. The principal randomly selects 3 homerooms and surveys every student in them. Name the design.
3. Describe how to randomly assign 30 tomato plants to two fertilisers, 15 each.

exercises.lol exercises.com.cn

4. A study finds that people who drink more coffee live longer, from medical records. Can it conclude that coffee causes longer life? Explain.
5. A student says: “A sample of 10,000 people from a voluntary online poll is better than a random sample of 1,000.” Correct the student.
6. A study records whether 500 adults walk daily and their cholesterol level over five years. Is it an experiment? What could be a confounding variable, and why?
7. A researcher wants to estimate the mean height of students in a school with very different grade levels (Grades 7 to 12). Suggest a sampling method and explain why it is better than an SRS of the same size.

Answers

1. A convenience sample. Early arrivers may differ systematically, for example by living closer or by being more organised, so the results may be biased.
2. A cluster sample: the homerooms are clusters, and every member of each chosen cluster is surveyed.
3. Number the plants 1 to 30. Use a random number generator to choose 15 different numbers from 1 to 30; those plants get fertiliser A and the other 15 get fertiliser B.
4. No. This is an observational study. Coffee drinkers may differ in other ways, such as income or health, which could explain the longer lives. Only random assignment of coffee could support causation.
5. A voluntary response sample is biased: people with strong opinions are more likely to respond. A larger biased sample is still biased; it only makes the wrong answer look more precise. The random sample of 1,000 is better.
6. No, it is an observational study: nobody was assigned to walk. Diet could confound: adults who walk daily may also eat more healthily, and diet affects cholesterol, so the effect of walking cannot be separated from the effect of diet.
7. A stratified random sample by grade. Heights differ a lot between grades, so randomly sampling within each grade and weighting grade means by their population shares gives a more precise estimate than an SRS, which might by chance contain too many students from one grade.

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.
- To explain confounding, link the variable to **both** the explanatory variable and the response, in context.
- Describe randomization so precisely that someone else could carry it out: label, random method, what to do with repeats, who gets which treatment.

exercises.lol exercises.com.cn

Probability, Random Variables, and Probability Distributions

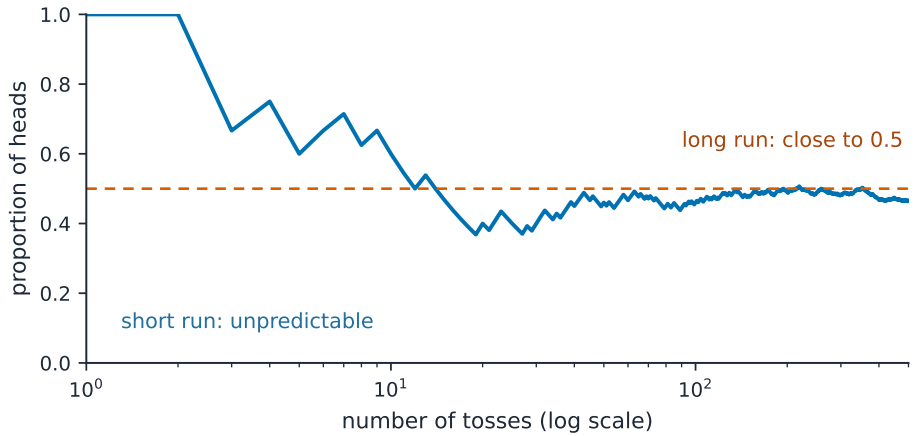
AP Statistics

Random and Non-Random Patterns

A manager says three people were chosen at random for a training trip, and all three are women, from a group of 6 men and 3 women. Bad luck, or unfair? Probability turns that hunch into a number: here about 1 chance in 84. This unit builds the rules that let you calculate chances like this and judge whether a result is surprising.

Something is **random** 随机 if individual **outcomes** 结果 are uncertain but a regular pattern emerges over **many repetitions** 重复. Short-run results look erratic; long-run **relative frequencies** 相对频率 settle down. This long-run stability is what makes probability useful. A streak or a pattern in a short run of data does not by itself show that the process is not random: random processes produce streaks quite often.

The **law of large numbers** 大数定律: as the number of trials grows, the relative frequency of an event gets closer and closer to its true probability. It says nothing about the short run: after five heads in a row, a fair coin is still 50% likely to land heads next.

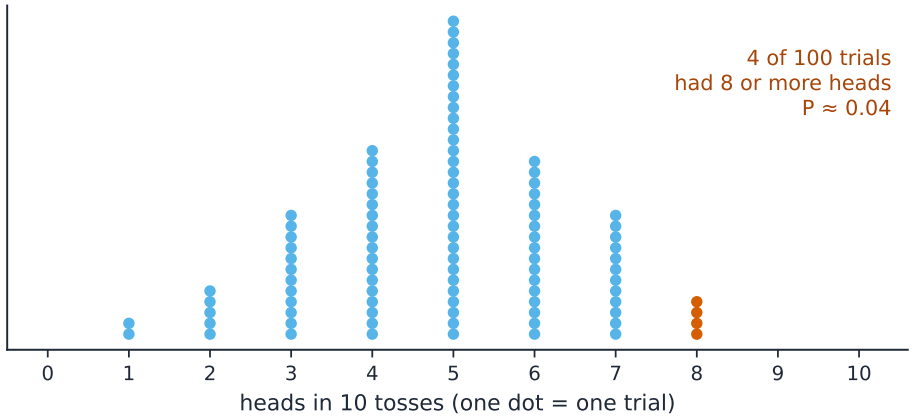


Random in the short run, predictable in the long run: the proportion of heads settles near 0.5

Estimating Probabilities Using Simulation

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

> **Worked example (a complete simulation).** Estimate the probability that choosing 3 people at random from 6 men and 3 women gives 3 women. Label the people 1 to 9, with 1–3 the women. Use a random number generator to pick integers from 1 to 9, **ignoring repeats**, until 3 different people are chosen; record whether all three are women. Repeat this many times (say 1000). The proportion of trials with three women estimates the probability. A simulation description must say what the numbers stand for, how repeats are handled, what is recorded, and that the process is repeated many times.



Repeat a chance process many times and count how often the event happens

Introduction to Probability

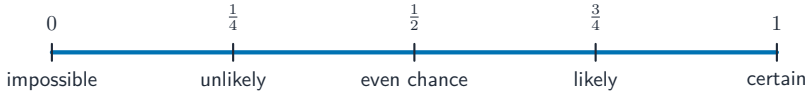
The **probability** 概率 of an **event** 事件 is a number from 0 to 1 giving its long-run relative frequency. The **sample space** 样本空间 is the set of all outcomes. For an event A, the **complement** 补 rule: $P(A^c) = 1 - P(A)$. Probabilities of all outcomes sum to 1. When all outcomes are equally likely,

$$P(A) = \frac{\text{number of outcomes in } A}{\text{total number of outcomes}}.$$

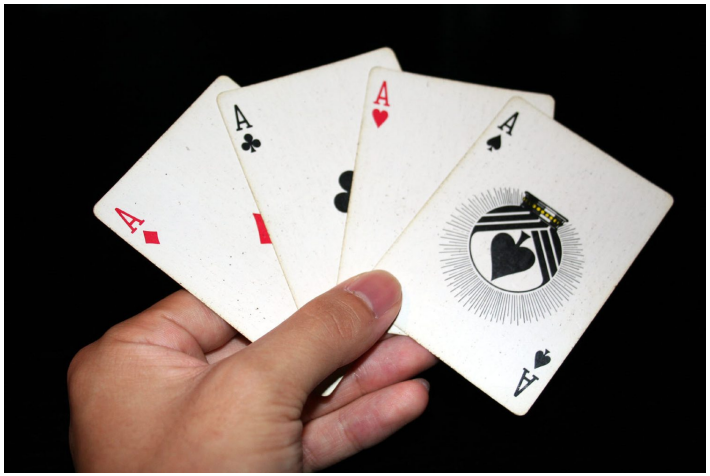
A probability is a **long-run** relative frequency. "$P(\text{rain}) = 0.3$" means that over many days like this one, it rains on about 30% of them.

exercises.lol exercises.com.cn

> **Worked example (is the claim believable?).** For the manager's choice above, there are $\binom{9}{3} = 84$ equally likely groups of 3, and only one of them is all three women: $P = \frac{1}{84} = 0.012$. An outcome this unlikely under random selection gives **reason to doubt** that the choice was random. It does not prove unfairness: unlikely things do happen, about 1 time in 84.



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** 不相交) if they cannot both happen. Then the **addition rule** 加法法则 simplifies:

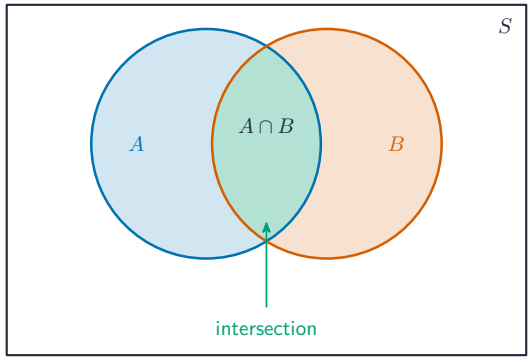
$$P(A \text{ or } B) = P(A) + P(B) \quad \text{(if mutually exclusive)}.$$

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

Worked example. Of 200 students at a school, 120 play a sport, 80 play an instrument and 50 do both. A student is chosen at random. Find $P(\text{sport or instrument})$, and say whether the events are mutually exclusive.

$$P(S \text{ or } I) = \frac{120}{200} + \frac{80}{200} - \frac{50}{200} = \frac{150}{200} = 0.75.$$

The events are **not** mutually exclusive: $P(S \text{ and } I) = \frac{50}{200} = 0.25 \neq 0$. Adding $0.60 + 0.40 = 1.00$ without the overlap would count the 50 students who do both twice.



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

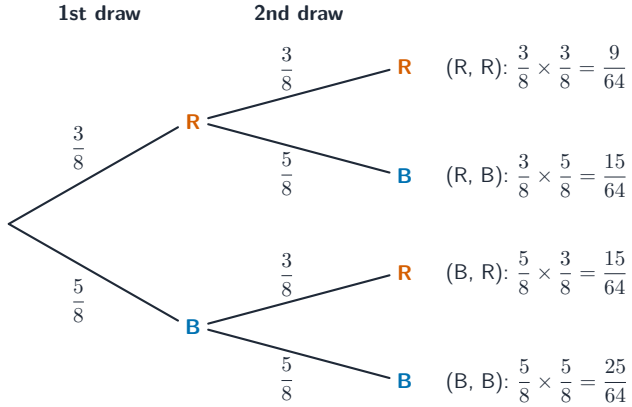
Conditional Probability

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 defined when $P(B) > 0$: the chance of A once you know B happened. **Two-way tables** 双向表 make these easy: restrict to the row/column for B, then find A's share.

exercises.lol exercises.com.cn



On a tree diagram, multiply the probabilities along the branches

Worked example: tree diagram. 2% of people have a disease. A test is positive for 95% of people who have it, and (a **false positive** 假阳性) for 10% of people who do not. A person tests positive. What is the probability that they have the disease?

- Branch "disease, positive": $0.02 \times 0.95 = 0.019$.
- Branch "no disease, positive": $0.98 \times 0.10 = 0.098$.
- So $P(\text{positive}) = 0.019 + 0.098 = 0.117$.

$$P(\text{disease} \mid \text{positive}) = \frac{P(\text{disease and positive})}{P(\text{positive})} = \frac{0.019}{0.117} \approx 0.162.$$

Only about 16% of positive results come from people with the disease: so few people have it that the false positives outnumber the true ones. Do not confuse $P(\text{positive} \mid \text{disease}) = 0.95$ with $P(\text{disease} \mid \text{positive})$.

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** 乘法法则 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** 不独立 (if one happens, the other cannot).

"At least one": use the complement. A mall has three ATMs, each working at opening time with probability 0.9, independently. The only way to have **none** working

is all three failing:

$$P(\text{at least one works}) = 1 - P(\text{none}) = 1 - (0.1)^3 = 0.999.$$

Given that at least one works, the probability all three work is

$$P(\text{all} \mid \text{at least one}) = \frac{P(\text{all and at least one})}{P(\text{at least one})} = \frac{0.9^3}{0.999} = \frac{0.729}{0.999} = 0.730.$$

The "and" in the numerator is just $P(\text{all})$, because "all three work" is already inside "at least one works".

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

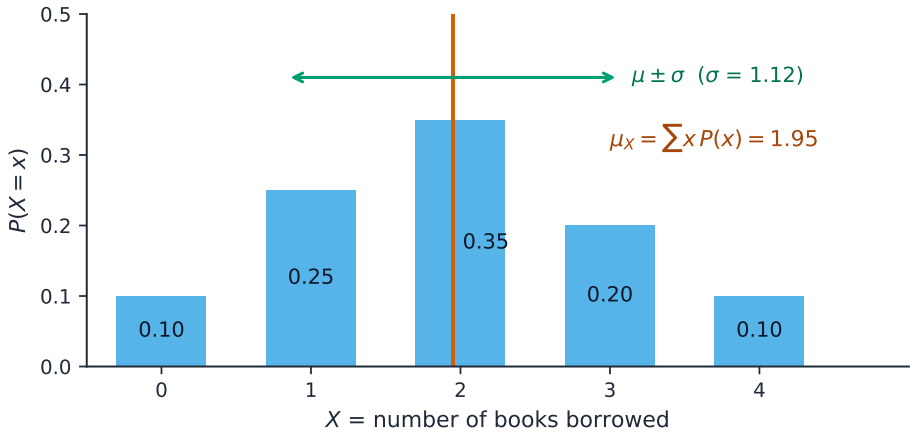
A sample space diagram lists every equally likely outcome

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** 离散 (a table of values) or **continuous** 连续 (an area-under-a-curve model like the normal).

A **cumulative** probability distribution gives $P(X \le x)$ for each value, adding the probabilities up from the smallest value. For a finite discrete table, its last entry is 1. Two random variables are **independent** if knowing the value of one does not change the probability distribution of the other.

exercises.lol exercises.com.cn



A discrete probability distribution with its mean and standard deviation

Mean and Standard Deviation of Random Variables

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

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

The **standard deviation** 标准差 $\sigma_X = \sqrt{\sum (x_i - \mu_X)^2 \, P(x_i)}$ 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.

> **Worked example (the ATMs again).** The number of working ATMs, X, has $E(X) = 3(0.9) = 2.7$. Interpretation in context: "Over many days, the mean number of ATMs working when the mall opens is about 2.7." If we are told that at least one ATM works, the case $X = 0$ is removed, so the expected value **increases** slightly: the remaining values are all 1 or more.

Combining Random Variables

Extension: combining random variables was removed from the AP Statistics exam curriculum effective 2026–27. See College Board's revisions.

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

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

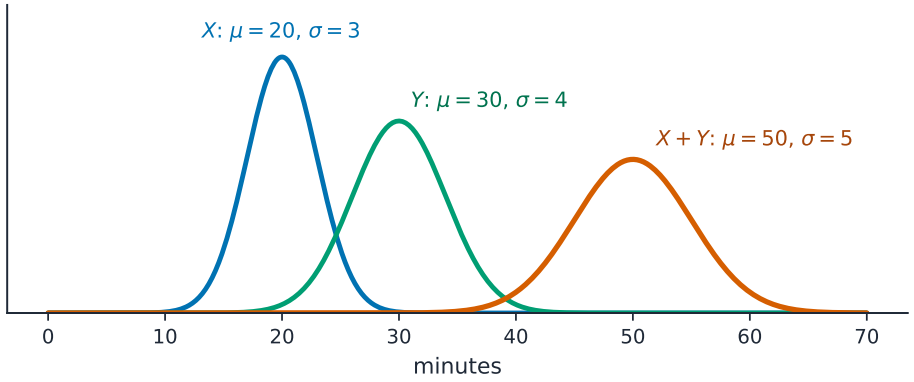
Take the square root for the standard deviation. Also, scaling: $\mu_{aX+b} = a\mu_X + b$ and $\sigma_{aX+b} = |a|\sigma_X$. A linear change $Y = a + bX$ does not change the **shape** of the distribution (for $b > 0$): a right-skewed X gives a right-skewed Y.

> **Worked example.** Task A takes X minutes with $\mu_X = 20$, $\sigma_X = 3$; task B takes Y minutes with $\mu_Y = 15$, $\sigma_Y = 4$; the times are independent.

- Total time: $\mu_{X+Y} = 20 + 15 = 35$ minutes, and $\sigma_{X+Y} = \sqrt{3^2 + 4^2} = \sqrt{25} = 5$ minutes.
- Difference: $\mu_{X-Y} = 20 - 15 = 5$ minutes, and $\sigma_{X-Y} = \sqrt{3^2 + 4^2} = 5$ minutes as well.

Two common errors: adding the standard deviations ($3 + 4 = 7$), and subtracting them for a difference ($4 - 3 = 1$). Add the **variances**, then take the square root.

A third error: treating $3X$ like $X_1 + X_2 + X_3$. Tripling **one** value triples the spread, $\sigma_{3X} = 3\sigma_X$; adding **three independent** values gives $\sigma = \sqrt{3}\,\sigma_X$, because the variances add.



$\mu_{X\pm Y} = \mu_X \pm \mu_Y \quad \sigma^2_{X\pm Y} = \sigma^2_X + \sigma^2_Y$, so $\sigma_{X+Y} = \sqrt{3^2 + 4^2} = 5$

For independent random variables means add and variances add

exercises.lol exercises.com.cn

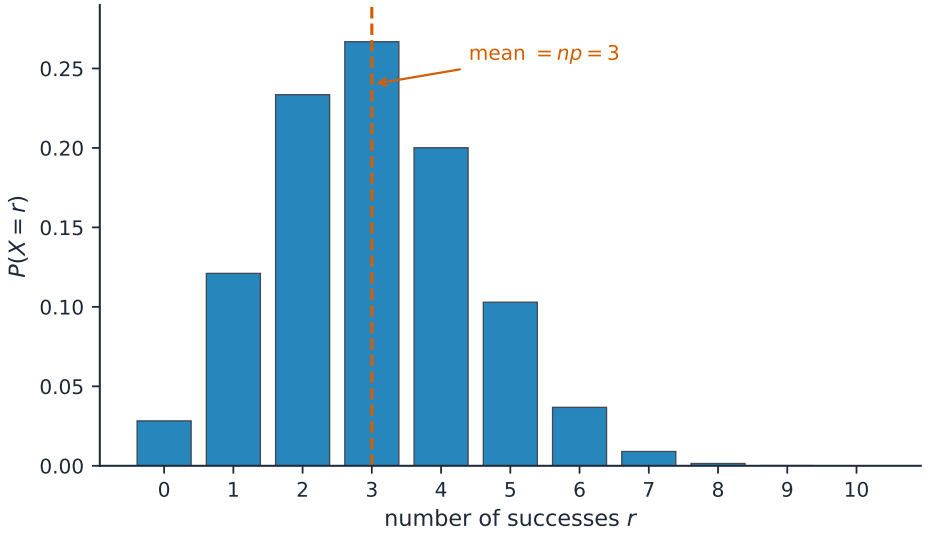
Introduction to the Binomial Distribution



Free throws: a fixed number of independent shots, each a make or a miss with the same probability, is a binomial setting

A **binomial** 二项 setting (BINS): a fixed number n of **Independent** trials, each with two outcomes (**success** 成功/**failure** 失败) 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. Interpret them in context: "If we test many batches of n, the mean number of successes per batch is about np." A binomial distribution can also be built from the formula, or estimated by simulation.

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 $\sigma = \sqrt{10(0.7)(0.3)} \approx 1.45$.

Worked example: "at least". For the same player ($n = 10$, $p = 0.7$), the probability of **at least** 8 makes adds the top three values:

$$P(X \geq 8) = P(8) + P(9) + P(10) = 0.2335 + 0.1211 + 0.0282 = 0.3828.$$

exercises.lol exercises.com.cn

By the complement, the probability of **at most** 7 makes is $P(X \leq 7) = 1 - 0.3828 = 0.6172$.

Common mistakes. Writing $P(X \geq 8) = 1 - P(X \leq 8)$: the complement of "at least 8" is "at most 7", so $P(X \geq 8) = 1 - P(X \leq 7)$. Using a binomial model when n is not fixed, or when the trials are not independent (drawing without replacement from a small group).

The Geometric Distribution

Extension: geometric distributions were removed from the AP Statistics exam curriculum effective 2026–27; the binomial distribution remains assessed.

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:

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

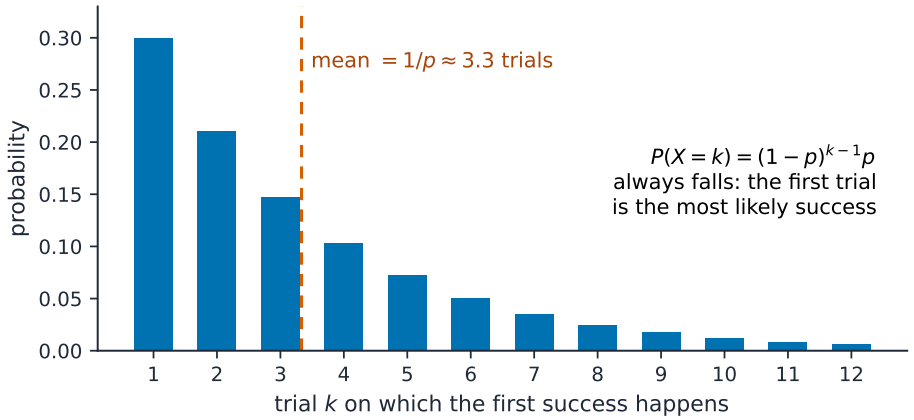
So the expected number of trials until the first success is $1/p$.

Worked example. A player makes 30% of shots, independently. The probability that the first make is on the **third** shot needs two misses, then a make:

$$P(Y = 3) = (0.7)^2(0.3) = 0.147.$$

The probability of a make within the first three shots uses the complement (three misses in a row): $P(Y \leq 3) = 1 - 0.7^3 = 0.657$. On average the first make comes on shot $\mu_Y = \frac{1}{0.3} \approx 3.33$. The standard deviation is $\sigma_Y = \dfrac{\sqrt{1-p}}{p} = \dfrac{\sqrt{0.7}}{0.3} = 2.79$ shots.

Worked example (rocket igniters). A company claims a failure rate of 15%. The probability that the first 30 igniters tested all work is $0.85^{30} = 0.0076$. Getting 30 successes in a row would be very unusual if the failure rate really were 15%, so it is reasonable to believe the failure rate is **lower**. Given those 30 successes, the chance that the first failure is on test 31 or 32 is $0.15 + 0.85(0.15) = 0.278$: past successes do not change the next trials, because they are independent.



The geometric distribution: trials until the first success

Practice

Answer each before reading the answers below.

1. $P(A) = 0.5$, $P(B) = 0.4$ and $P(A \text{ and } B) = 0.2$. Find $P(A \text{ or } B)$.
2. Are A and B in question 1 independent? Show why.
3. A basketball player makes 30% of three-point shots, independently. In 10 shots, find the probability of exactly 2 makes.
4. A game costs \$1 to play. You win \$5 (and keep your \$1) with probability 0.2; otherwise you lose the \$1. Find the expected gain per game.
5. A student says: "Mutually exclusive events are independent, because one has nothing to do with the other." Correct the student.
6. Each of 4 independent smoke alarms works with probability 0.95. Find the probability that at least one works.
7. In a large school, 40% of students walk to school. If 8 students are chosen at random, find the mean and standard deviation of the number who walk, and interpret the mean.
8. Machine A makes 60% of a factory's bolts and 2% of its bolts are faulty; machine B makes the rest, and 5% of B's bolts are faulty. A bolt is faulty. Find the probability it came from machine A.

Answers

1. $P(A \text{ or } B) = 0.5 + 0.4 - 0.2 = 0.7$.
2. Yes: $P(A)P(B) = 0.5 \times 0.4 = 0.2 = P(A \text{ and } B)$.
3. Binomial with $n = 10$, $p = 0.3$: $\binom{10}{2}(0.3)^2(0.7)^8 = 45(0.09)(0.0576) = 0.233$.

4. $E = 0.2(5) + 0.8(-1) = 0.20$: on average the player gains \$0.20 per game.
5. If A and B are mutually exclusive and both have positive probability, then knowing A happened tells you B did **not**. So $P(B \mid A) = 0 \neq P(B)$: they are dependent.
6. $1 - 0.05^4 = 1 - 0.00000625 = 0.99999$.
7. $\mu = 8(0.4) = 3.2$, $\sigma = \sqrt{8(0.4)(0.6)} = 1.39$. In many random samples of 8 students, the mean number who walk would be about 3.2.
8. $P(\text{faulty}) = 0.6(0.02) + 0.4(0.05) = 0.012 + 0.020 = 0.032$, so $P(A \mid \text{faulty}) = \dfrac{0.012}{0.032} = 0.375$.

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** $= \sum(\text{value} \times \text{probability})$.
- Recognise **binomial** (fixed n, independent trials, two outcomes, constant p); geometric settings are extension content for exams from 2027.
- Draw a tree or table for multi-stage problems and multiply along branches.
- For "at least one", subtract "none" from 1.
- When a result has a small probability under a claim, say it gives **reason to doubt** the claim, and quote the probability.
- Interpret an expected value as a **long-run mean** over many repetitions, in context.

exercises.lol exercises.com.cn

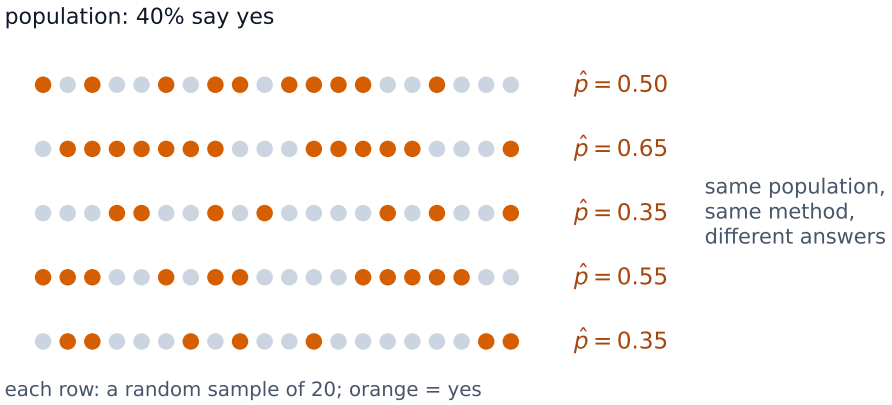
Sampling Distributions

AP Statistics

Why Two Samples Never Match: Sampling Variability

You and a friend each survey 50 random students about the canteen. You find 62% like it; your friend finds 54%. Neither of you made a mistake: random samples simply differ. How much they differ is predictable, and that prediction is the key to every confidence interval and test in the rest of the course.

A **statistic** 统计量 (like a **sample mean** 样本均值 $\bar{x}$ or **sample proportion** 样本比例 $\hat{p}$) is computed from a sample and **varies** from sample to sample – this is **sampling variability** 抽样变异性. A **parameter** 参数 (μ or p) is the fixed truth about the **population** 总体. 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** 推断. Differences between samples can be random (sampling variability) or not (for example, one sample was chosen badly). The sampling distribution tells you how large the random differences usually are.



Five random samples from one population give five different sample proportions

The Normal Curve as a Model for a Statistic

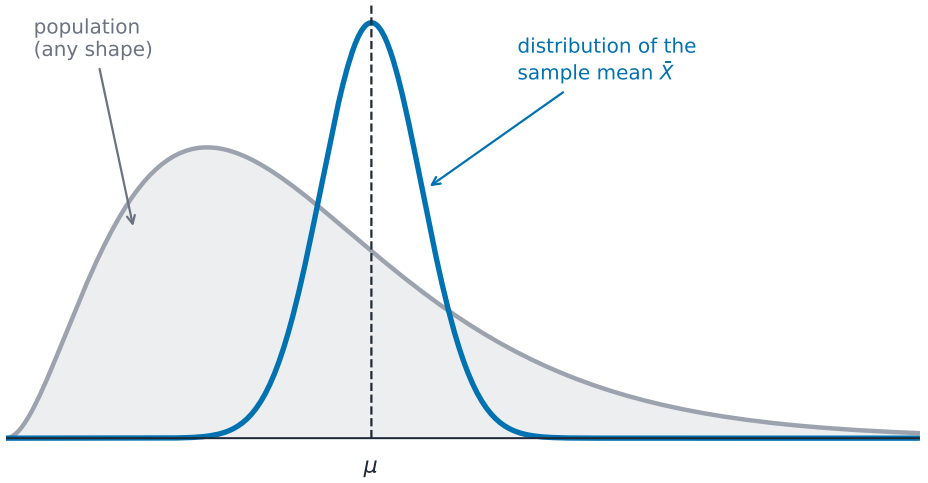
For large enough samples, many sampling distributions are approximately **normal** 正态的. That lets us describe a statistic by a **center** 中心 (its mean), a **spread** 离散程度 (its **standard error** 标准误), and a normal shape – and then compute how likely a given sample result is.

A **continuous random variable** 连续随机变量 can take any value in an interval. Its probabilities are **areas** under a density curve, so the probability of any single exact value is 0, and $P(X < a) = P(X \le a)$. To find an area, write the inequality for the interval, change the boundaries to z-scores, and use a table or technology. Normal models also work well for real distributions that are roughly symmetric and mound-shaped.

The Central Limit Theorem

The **Central Limit Theorem** 中心极限定理 (CLT): for **independent** 独立 observations from the same population with finite, positive **variance** 方差, the sampling distribution of a **sample mean** becomes approximately **normal** as n increases. A strongly **skewed** 偏斜 population may need a much larger sample than 30; there is no universal sample-size cutoff. The standard error decreases as n increases.

A sampling distribution can also be **simulated**: take many random samples from a population (or a model of it) and compute the statistic for each one. In a randomized experiment, a **randomization distribution** 随机化分布 is built the same way, by re-randomizing the units to groups many times as if the treatment had no effect, and recording the difference each time.



With finite positive population variance, the sample mean becomes approximately normal as the sample size grows

exercises.lol exercises.com.cn

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** 偏差 is about the center being off; **variability** 变异性 is about the spread. A good **estimator** 估计量 is both unbiased (centered right) and low-variability (**precise** 精确的); larger samples reduce variability but do not fix bias from bad sampling. Any sample statistic used to estimate a parameter is a **point estimate** 点估计; its sampling distribution shows how far from the parameter it tends to land.

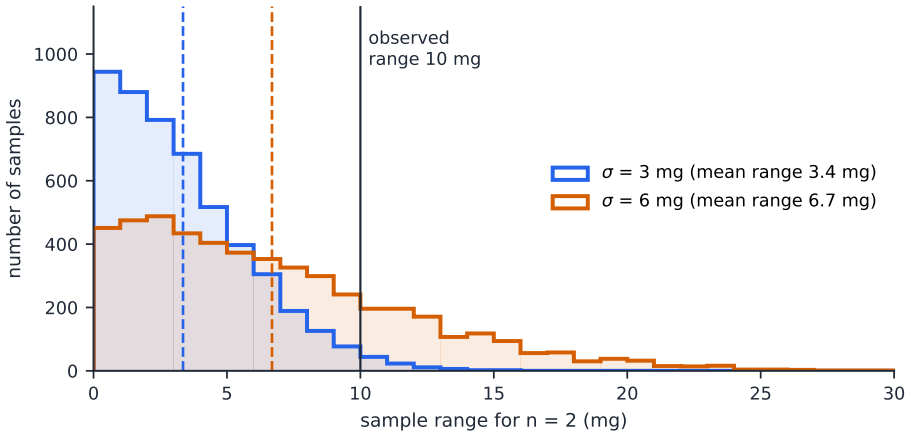


Arrows in a target: bias is how far the centre of the group is from the bullseye, variability is how spread out the group is

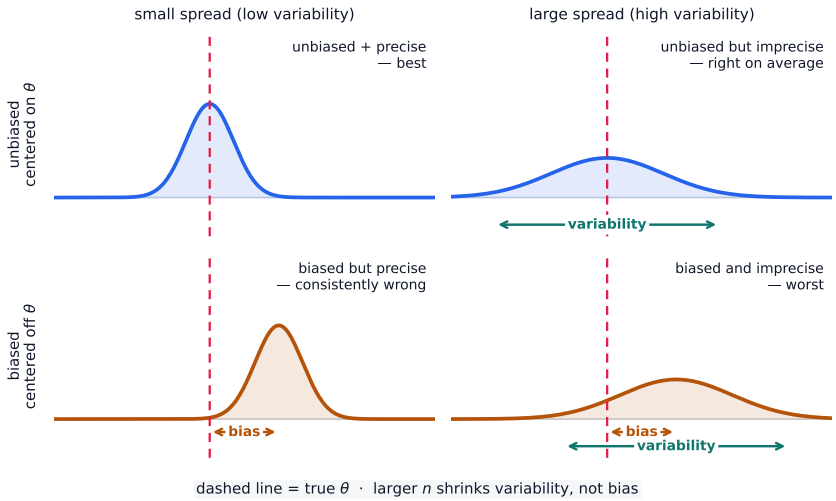
Other statistics. Every statistic has a sampling distribution, not only $\bar{x}$ and $\hat{p}$. The sample **median** estimates the population median, not generally the population mean; its finite-sample distribution need not be centred exactly on either parameter. The sample **range** of small samples is biased low (a sample rarely contains the most extreme values), and its distribution is skewed right.

> **Worked example (the sample range).** A machine puts gold on necklaces; the amount is normal with $\mu = 300$ mg. A statistician measures 2 necklaces a day and records the **range**. The figure shows 5000 simulated ranges for $\sigma = 3$ mg and for $\sigma = 6$ mg. As σ grows, the distribution of the range shifts to the **right** and spreads out. A range of at least 10 mg happens in under 2% of samples when $\sigma = 3$, but in about a quarter of them when $\sigma = 6$: so an observed range of 10 mg is evidence that σ is larger than 3 mg.

exercises.lol exercises.com.cn



A larger population standard deviation moves the sampling distribution of the range to the right and spreads it out



*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** 简单随机样本: the mean is p (unbiased), and the **standard deviation** 标准差 is

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

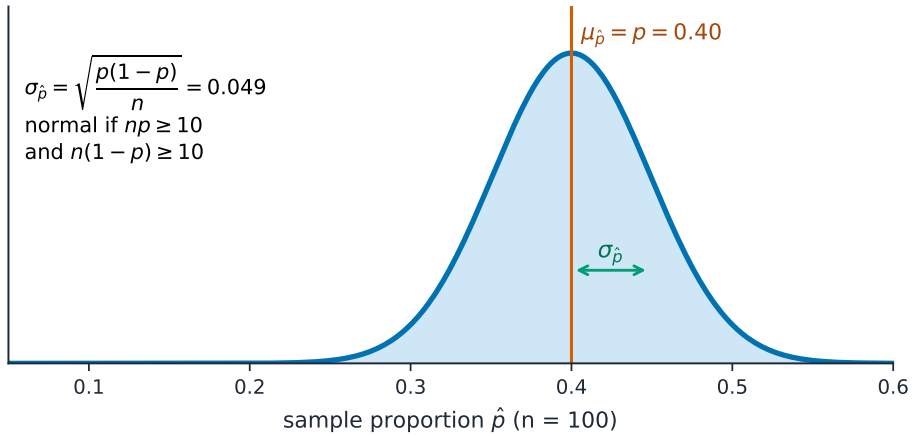
This spread is the **standard error**: the standard deviation of the statistic's sampling distribution. When p is unknown, replace it by $\hat{p}$ to estimate that standard error, as in the later inference units. It is approximately **normal** when $np \geq 10$ and $n(1-p) \geq 10$ (the **Large Counts condition** 大计数条件), and the 10% condition ($n \leq 0.10N$) keeps the observations near-independent. Strictly, sampling **without** replacement makes the standard deviation a little **smaller** than the formula gives; when n is under 10% of the population, the difference is too small to matter.

> **Worked example (a point estimate).** In a random sample of 300 students, 87 say yes. The point estimate of the population proportion is $\hat{p} = \frac{87}{300} = 0.29$. Interpret the standard deviation in context: if p were 0.29, sample proportions from samples of 300 would typically differ from 0.29 by about $\sqrt{0.29(0.71)/300} = 0.026$.

> **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.

Common mistakes. Checking "$n \geq 30$" for a proportion: that rule belongs to means; for $\hat{p}$ check $np \geq 10$ and $n(1-p) \geq 10$ (here 40 and 60). Treating the 10% condition as a normality check: it keeps the observations close to independent, so the formula for $\sigma_{\hat{p}}$ is valid.

exercises.lol exercises.com.cn



The sampling distribution of $\hat{p}$ is centred at p with standard deviation $\sqrt{p(1-p)/n}$

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** 方差相加:

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

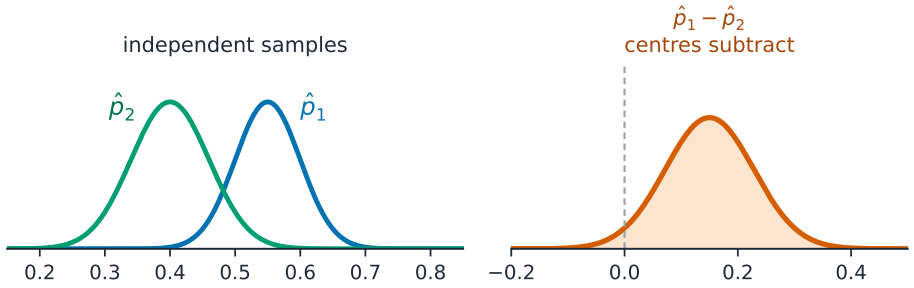
It is approximately normal when the Large Counts condition holds in **both** samples.

> **Worked example.** In town 1, 60% of adults cycle to work; in town 2, 50%. Independent random samples of 100 adults are taken from each town. How likely is it that the town 1 sample shows a **lower** proportion than town 2?

The mean of $\hat{p}_1 - \hat{p}_2$ is $0.6 - 0.5 = 0.10$, and

$$\sigma_{\hat{p}_1 - \hat{p}_2} = \sqrt{\frac{0.6(0.4)}{100} + \frac{0.5(0.5)}{100}} = \sqrt{0.0049} = 0.07.$$

Large Counts: 60, 40, 50 and 50 are all at least 10, so use a normal model. $z = \dfrac{0 - 0.10}{0.07} \approx -1.43$, so $P(\hat{p}_1 - \hat{p}_2 < 0) \approx 0.077$. About 8% of such pairs of samples would put town 1 behind, even though town 1 really is ahead.



$$\sigma = \sqrt{\frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}}$$

The distribution of $\hat{p}_1 - \hat{p}_2$: centres subtract and variances add

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. The spread of the **data** in one sample does not shrink with n: a sample of 200 tortillas has about the same spread of diameters ($\approx \sigma$) as a sample of 20. What shrinks is the spread of the **sample means** from one day to the next.

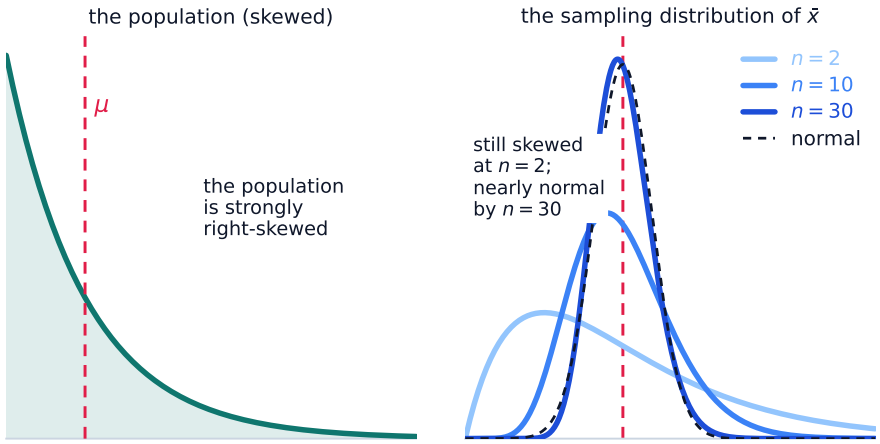
Worked example (three days instead of one). A school's daily absences are normal with $\mu = 120$ and $\sigma = 11$. It loses funding if absences on a randomly chosen day exceed 140: $z = \frac{140-120}{11} = 1.82$, so $P \approx 0.034$. If the state instead uses the **mean** of 3 randomly chosen days (assume independent), $\sigma_{\bar{x}} = \frac{11}{\sqrt{3}} = 6.35$ and $z = \frac{140-120}{6.35} = 3.15$, so $P \approx 0.0008$. Averaging several days makes an extreme result much less likely.

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 $P(\bar{x} > 73) \approx 0.067$.

Common mistakes. Using $\sigma = 12$ instead of $\sigma_{\bar{x}} = 2$ for a question about a **mean** of 36 values. For **one** individual, $z = (73 - 70)/12 = 0.25$ and $P(X > 73) \approx 0.40$: a single value above 73 is common, a sample mean above 73 is not. Saying the CLT

exercises.lol exercises.com.cn

makes the **population** (or the sample) normal: it is the sampling distribution of $\bar{x}$ that becomes approximately normal.



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

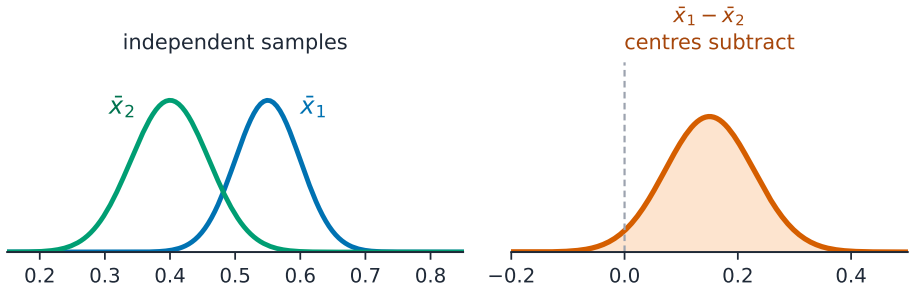
The population on the left is strongly skewed, yet every sampling distribution of $\bar{x}$ is centered at μ. A larger n shrinks the standard error $\sigma/\sqrt{n}$, so the curve gets taller and narrower – and it also straightens: still clearly skewed at $n = 2$, almost exactly normal (dashed) by $n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

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.



$$\sigma = \sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}$$

The distribution of $\bar{x}_1 - \bar{x}_2$: centres subtract and variances add

Worked example. Machine 1 fills bottles with $\mu_1 = 500$ mL and $\sigma_1 = 4$ mL; machine 2 with $\mu_2 = 498$ mL and $\sigma_2 = 3$ mL. Both fill amounts are normal. Independent random samples of 16 bottles from machine 1 and 9 from machine 2 are taken. The mean of $\bar{x}_1 - \bar{x}_2$ is $500 - 498 = 2$ mL, and

$$\sigma_{\bar{x}_1 - \bar{x}_2} = \sqrt{\frac{4^2}{16} + \frac{3^2}{9}} = \sqrt{1 + 1} \approx 1.41 \text{ mL}.$$

Both populations are normal, so the difference is normal even for these small samples. The machine 2 sample has the larger mean when $\bar{x}_1 - \bar{x}_2 < 0$:

$$z = \frac{0 - 2}{1.41} \approx -1.41, \qquad P(\bar{x}_1 - \bar{x}_2 < 0) \approx 0.079.$$

The variances add even though the means subtract: $1+1$ under the square root, never $1 - 1$.

Practice

Answer each before reading the answers below.

1. A population proportion is $p = 0.30$ and samples of $n = 50$ are taken. Give the mean and standard deviation of $\hat{p}$, and check the Large Counts condition.
2. A normal population has $\mu = 100$ and $\sigma = 15$. For independent random samples of $n = 25$, find $P(\bar{x} > 106)$.
3. By what factor must the sample size grow to halve the standard deviation of $\bar{x}$?
4. A population of incomes is strongly right-skewed. Is $\bar{x}$ approximately normal for samples of $n = 5$? For $n = 50$?

exercises.lol exercises.com.cn

5. A student says: "A bigger sample removes bias." Correct the student.
6. Heights of a plant species are normal with $\mu = 40$ cm and $\sigma = 6$ cm. Find the probability that one plant is taller than 43 cm, and that the mean of 9 plants is more than 43 cm.
7. Explain why the sample maximum is a biased estimator of the population maximum.

Answers

1. Mean 0.30; standard deviation $\sqrt{0.30(0.70)/50} = 0.065$. Large Counts: $np = 15$ and $n(1-p) = 35$, both at least 10.
2. $\sigma_{\bar{x}} = 15/\sqrt{25} = 3$, so $z = \dfrac{106 - 100}{3} = 2.0$ and $P(\bar{x} > 106) \approx 0.023$.
3. Four times as large, because $\sigma_{\bar{x}} = \sigma/\sqrt{n}$.
4. Not for $n = 5$: with a skewed population, a small sample's mean is still skewed. For $n = 50$, probably yes: $n \geq 30$ is the usual AP guideline, so by the Central Limit Theorem $\bar{x}$ is approximately normal. An extremely skewed population (a few huge incomes) could still need a larger sample.
5. A larger sample reduces **variability** (the spread of the sampling distribution), not **bias** (where its center sits). A biased method such as a **voluntary-response** 自愿回应 poll stays biased, however large.
6. One plant: $z = \frac{43-40}{6} = 0.5$, $P \approx 0.309$. Mean of 9: $\sigma_{\bar{x}} = \frac{6}{3} = 2$, $z = 1.5$, $P \approx 0.067$.
7. A sample can never contain a value larger than the population maximum, and it usually misses the largest values, so the sample maximum is usually **below** the true maximum. Its sampling distribution is centred below the parameter, so it is biased (low).

Exam tips

- A **sampling distribution** describes a statistic over repeated samples. It is centered on the true parameter when the statistic is **unbiased**.
- 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.
- Distinguish the spread of **individual values** (σ) from the spread of **sample means** ($\sigma/\sqrt{n}$); read which one the question asks about.
- To judge a sample result, place it in the sampling distribution: a value far in the tail is unlikely if the claimed parameter is true.

Inference for Categorical Data: Proportions

AP Statistics

Why Be Normal?

A poll says 52% of voters back a candidate, "with a margin of error of 3 points". Is the candidate really ahead? A company claims 90% of customers are satisfied, and your sample finds 84%. Is the claim false, or was your sample unlucky? This unit gives the two tools that answer such questions: confidence intervals and significance tests.

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

Confidence Interval for a Proportion

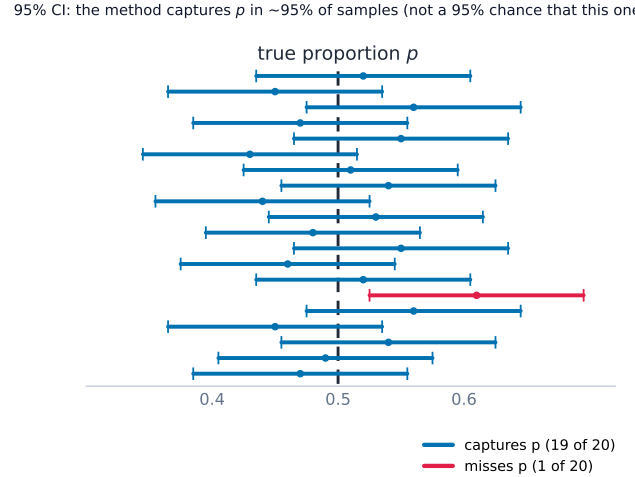
A **confidence interval** 置信区间 estimates the **parameter** 参数 as a range: statistic ± **margin of error** 误差幅度.

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

z^* is the **critical value** 临界值 for the **confidence level** 置信水平 (e.g. 1.96 for 95%). Conditions: **random sample** 随机样本, **Large Counts condition** 大计数条件 ($n\hat{p} \geq 10$ 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."

Why the conditions matter. The interval assumes the sampling distribution of $\hat{p}$ is approximately normal, which is what makes z^* the right multiplier. With fewer than 10 successes or failures, that distribution is skewed and the interval does not capture p as often as claimed. The **random** condition lets you generalize to the population; the **10%** condition keeps the observations close to independent, so the standard error formula is correct.

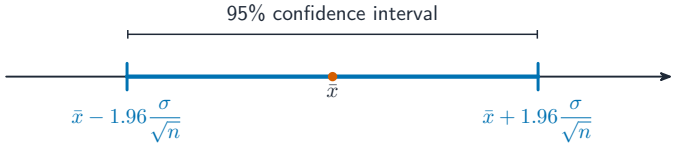
exercises.lol exercises.com.cn



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%.

Common mistakes. Saying "there is a 95% chance that p is between 0.532 and 0.668": p is fixed, and this interval either contains it or does not; the 95% describes the **method**. Describing individuals or the sample instead of the population proportion. Skipping the Large Counts check: here $n\hat{p} = 120$ and $n(1-\hat{p}) = 80$, both at least 10.



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

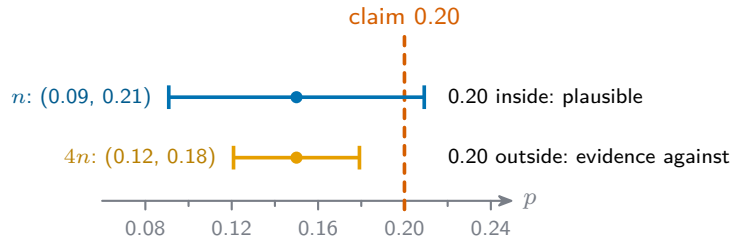
Width. The width of the interval is twice the margin of error. With everything else fixed, a **larger sample** gives a narrower interval (four times the sample halves the width), and a **higher confidence level** gives a wider one (99% uses $z^* = 2.576$).

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** 可信 values include the claim, **in context** 结合情境.

> **Worked example (a discount program).** A store's program should give a discount on 20% of bills. A random sample gives the 95% interval 0.15 ± 0.06, that is $(0.09,\ 0.21)$.

- Is there convincing evidence the program is **not** working as intended? No: 0.20 is inside the interval, so 20% is a plausible value.
- Is there convincing evidence that it **does** give the discount with probability 0.20? Also no: many other values, such as 0.10, are plausible too. An interval can never prove a single value.
- A second sample **four times as large** also gives $\hat{p} = 0.15$. Its margin of error is half as big, 0.03, so the interval is $(0.12,\ 0.18)$. Now 0.20 is outside: there is convincing evidence that the program gives fewer discounts than intended.



Inside the interval: plausible. Outside: evidence against. Four times the sample halves the margin of error and moves 0.20 outside

exercises.lol exercises.com.cn

> **Worked example (an interval in other units).** A 95% interval for the proportion of water-cup customers who take soft drinks is $(0.24,\ 0.38)$. Each such customer costs \$0.25, and 3000 customers ask for water cups in June. Multiply both ends: $3000 \times 0.25 \times (0.24,\ 0.38) = (\$180,\ \$285)$. We are 95% confident the expected June cost, assuming the same customer proportion, is between \$180 and \$285. This estimates the mean cost, not a prediction interval for the realised cost.

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:

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

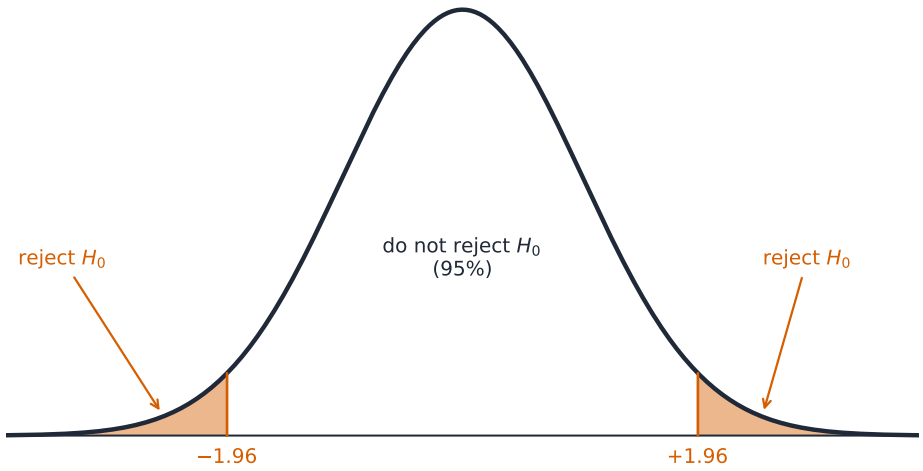
Check the same conditions (random, Large Counts using p_0, 10%). The **test statistic** 检验统计量 counts standard errors from p_0:

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

> **Worked example.** A company claims 90% satisfaction; a random sample of 100 of its customers finds 84 satisfied ($\hat{p} = 0.84$). Is there convincing evidence, at $\alpha = 0.05$, that the true rate is not 90%? A full free-response answer has four parts, and each one earns credit.

- **State.** $H_0 : p = 0.90$ and $H_a : p \neq 0.90$, where p is the true proportion of **all** the company's customers who are satisfied. Define p in words: hypotheses with an undefined symbol lose credit.
- **Plan.** A one-sample z-test for a proportion. Random: the customers were a random sample. 10%: 100 is less than 10% of all customers. Large Counts, **using** p_0: $np_0 = 100(0.90) = 90 \geq 10$ and $n(1-p_0) = 100(0.10) = 10 \geq 10$.
- **Do.**
  $$z = \frac{0.84 - 0.90}{\sqrt{0.90(0.10)/100}} = \frac{-0.06}{0.03} = -2.0,$$
  giving a **two-tailed** 双尾 p-value of about $2(0.023) = 0.046$.
- **Conclude.** Since $0.046 < 0.05$, we **reject** 拒绝 H_0. There is convincing evidence that the true proportion of satisfied customers is not 90% (the sample suggests it is lower).

The conclusion must compare the p-value with α **and** be stated about the parameter, in context.



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** 观测到的 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.

A full interpretation has three parts: the assumption, the probability, and the observed result, in context. For the satisfaction test: "**Assuming** the true satisfaction rate is 90%, there is a 0.046 probability of getting a sample proportion at least as far from 0.90 as 0.84 (in either direction) by chance in a random sample of 100 customers." The distribution of the test statistic when H_0 is true is the **null distribution**; it may come from the normal model or from a simulation.

Common mistakes. Writing hypotheses about the statistic ($H_0 : \hat{p} = 0.90$): hypotheses are always about the **parameter** p. Using $\hat{p}$ in the test's standard error: the test assumes H_0, so it uses p_0 (the interval uses $\hat{p}$). Reading $p = 0.046$ as "a 4.6% chance 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):

- $p \le \alpha$: **reject** H_0 – there is **convincing evidence** 有力证据 for H_a.

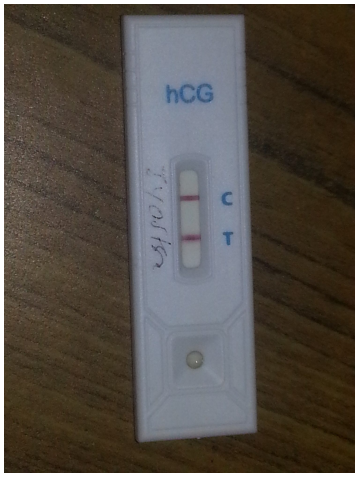
exercises.lol exercises.com.cn

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

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

Common mistakes. Writing "accept H_0" or "this proves $p = 0.90$" after a large p-value: failing to reject only means the evidence is not strong enough. Concluding without the comparison of the p-value with α, or without context.

Type I and Type II Errors



A diagnostic test can err both ways: a false positive is like a Type I error, a false negative like a Type II error

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

Describe each error and its **consequence** 后果 in the problem's context. Which error is worse depends on the situation, and that decides α. If a Type I error is very costly (approving a drug that does not work), choose a **small** α, such as 0.01. If a Type II error is worse (missing a dangerous fault), a larger α such as 0.10 raises the power.

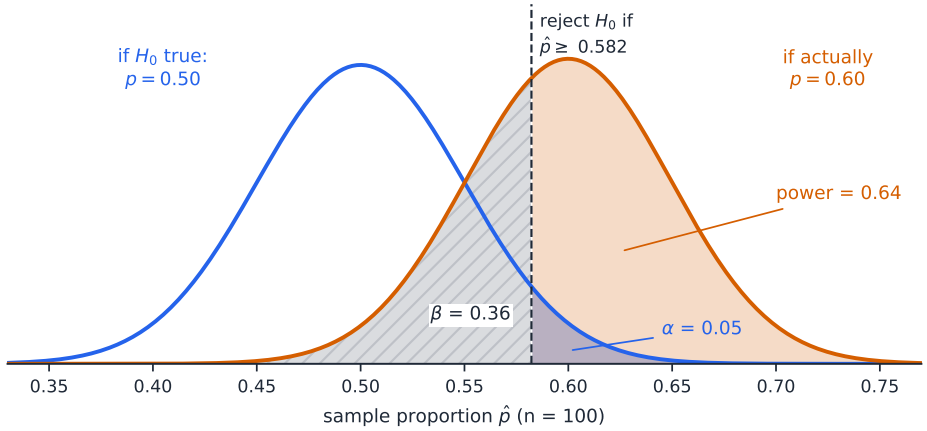
> **Worked example.** In the satisfaction test above ($H_0 : p = 0.90$):

- **Type I error:** concluding that the satisfaction rate is not 90% when it really is 90%. Consequence: the company is wrongly accused of overstating its rate.

- **Type II error:** failing to find evidence against 90% when the true rate is different (say 85%). Consequence: customers keep trusting a rate that is too high.

A good answer names the error **and** says what happens because of it, in the words of the problem.

Power in a picture. Test $H_0 : p = 0.50$ against $H_a : p > 0.50$ with $n = 100$ and $\alpha = 0.05$. Using the normal approximation, the rejection cut-off is approximately $\hat{p} = 0.582$. If the true proportion is really 0.60, the chance of rejecting is the area of the orange curve beyond 0.582: the power is 0.64, so $\beta = 0.36$.



The cut-off is set by the null distribution; the power is the area of the true distribution beyond it

With a **larger sample**, both curves get narrower, so they overlap less: the power **increases** and β falls. Power also increases with a larger α (the cut-off moves left) and with a true value farther from p_0.

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.

> **Worked example.** For the two teaching methods below ($\hat{p}_1 = \frac{45}{75} = 0.60$, $\hat{p}_2 = \frac{32}{80} = 0.40$), a 95% interval for $p_1 - p_2$ is

$$(0.60 - 0.40) \pm 1.96\sqrt{\frac{0.60(0.40)}{75} + \frac{0.40(0.60)}{80}} = 0.20 \pm 1.96(0.0787) = 0.20 \pm 0.154,$$

exercises.lol exercises.com.cn

which is (0.046, 0.354). We are 95% confident that method 1's true pass rate is between 4.6 and 35.4 **percentage points** 百分点 higher than method 2's. The interval uses the **separate** $\hat{p}$ values; the pooled $\hat{p}_c$ belongs only to the test.

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.

> **Worked example: a two-proportion test.** Two teaching methods are compared with random samples of students. Method 1: 45 of 75 pass. Method 2: 32 of 80 pass. Is there convincing evidence, at $\alpha = 0.05$, of a difference in pass rates?

- **State.** $H_0 : p_1 = p_2$ and $H_a : p_1 \neq p_2$, where p_1, p_2 are the true pass rates under each method.
- **Plan.** Two-sample z-test for a difference of proportions. Independent random samples; each is less than 10% of its population; using $\hat{p}_c = 77/155$, the pooled expected success/failure counts are about 37.3, 37.7, 39.7, 40.3, all at least 10.
- **Do.** $\hat{p}_1 = 0.60$, $\hat{p}_2 = 0.40$, pooled $\hat{p}_c = \dfrac{77}{155} = 0.497$.

$$z = \frac{0.60 - 0.40}{\sqrt{0.497(0.503)\left(\frac{1}{75} + \frac{1}{80}\right)}} = \frac{0.20}{0.0804} = 2.49, \qquad p \approx 0.013.$$

- **Conclude.** Because $0.013 < 0.05$, reject H_0. There is convincing evidence that the true pass rates under the two methods differ; method 1's is higher.

Practice

Answer each before reading the answers below.

1. In a random sample of 160 students, 64 walk to school. Construct a 95% confidence interval for the true proportion who walk.
2. How large a sample gives a 95% margin of error of at most 0.04, with no prior estimate of p?
3. A coin is tossed 400 times and lands heads 220 times. Test $H_0 : p = 0.5$ against $H_a : p \neq 0.5$.
4. Which is the correct meaning of "95% confident" for the interval in question 1? (a) 95% of students walk between 32% and 48% of the time. (b) The method used captures the true proportion in about 95% of samples.
5. A student says: "My interval $(0.532,\ 0.668)$ has a 95% chance of containing p." Correct the student.
6. A 95% interval for the proportion of a school's students who read daily is $(0.31,\ 0.43)$. The principal claims that 40% read daily. Does the interval give convincing evidence against the claim? Does it prove the claim?
7. In a test of $H_0 : p = 0.30$, describe a Type II error in context if p is the proportion of a company's employees who work from home.
8. How would increasing the sample size from 100 to 400 change the power of the test in the power figure? Explain.

Answers

1. $\hat{p} = 0.40$, $SE = \sqrt{0.40(0.60)/160} = 0.0387$, so $0.40 \pm 1.96(0.0387) = 0.40 \pm 0.076$, giving $(0.324,\ 0.476)$.
2. $n = \dfrac{1.96^2(0.5)(0.5)}{0.04^2} = 600.25$, so survey **601** people.
3. $\hat{p} = 0.55$, $z = \dfrac{0.55 - 0.50}{\sqrt{0.25/400}} = \dfrac{0.05}{0.025} = 2.0$, two-sided $p \approx 0.046$. At $\alpha = 0.05$, reject H_0: convincing evidence that the coin is not fair.
4. (b). The confidence level describes the method's long-run success rate.
5. Once computed, the interval either contains p or it does not. The 95% describes the method: about 95% of intervals made this way capture p. Say "we are 95% confident", not "a 95% probability".
6. No: 0.40 is inside the interval, so it is plausible. But the interval does not prove $p = 0.40$, because values such as 0.35 are plausible too.
7. A Type II error would be failing to find convincing evidence that the proportion who work from home differs from 0.30 when it really does differ.
8. The power would increase. With $n = 400$ the standard deviation of $\hat{p}$ halves, so the null and true sampling distributions overlap less, and a true value of 0.60 is much more likely to give a $\hat{p}$ beyond the cut-off.

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.
- A claimed value inside the interval is **plausible**, not proven; a value outside it is evidence against the claim.
- Interpret a p-value with all three parts: "assuming H_0 is true", the probability, and "a result at least as extreme as the one observed", in context.

exercises.lol exercises.com.cn

Inference for Quantitative Data: Means

AP Statistics

Should I Worry About Error?

Inference for a **mean** 均值 works like inference for a proportion, with one change: we rarely know the **population standard deviation** 总体标准差 σ, so we estimate it with the sample s. That extra uncertainty means we use the ***t*-distribution** t 分布 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.

A packing machine is supposed to fill bags of chips with 28.0 g. A quality checker weighs 20 bags and finds a mean of 27.4 g. Is the machine underfilling, or is 0.6 g just the luck of which bags were picked? This unit turns that question into a number, and makes you say what could go wrong with your answer.

Confidence Interval for a Mean



A blood-pressure reading of 120/74: one measurement varies, so a study estimates the mean with a confidence interval

A one-sample t **confidence interval** 置信区间 for μ:

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

t^* is the **critical value** 临界值 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** 异常值), and the **10%** condition. Interpret the interval and the **confidence level** 置信水平 in context.

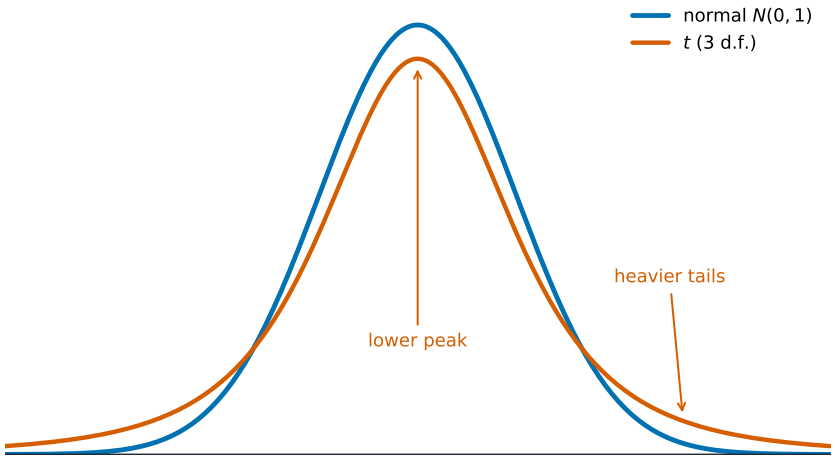
> **Worked example.** Assume the population is approximately normal. A random sample of $n = 25$ has $\bar{x} = 50$ and $s = 8$. For a 95% interval, $df = 24$ gives $t^* = 2.064$:
>
> $$50 \pm 2.064 \cdot \frac{8}{\sqrt{25}} = 50 \pm 2.064(1.6) = 50 \pm 3.3 = (46.7,\ 53.3).$$

Interpretation: "We are 95% confident that the interval from 46.7 to 53.3 captures the true mean of the population the sample was taken from." Name the variable and the units in a real problem.

About *t*. The t-distributions have heavier tails than the normal because s varies from sample to sample. As the degrees of freedom increase, the tails get thinner and t^* shrinks toward z^*: for 95%, $t^* = 2.776$ with $df = 4$, 2.064 with $df = 24$ and 1.984 with $df = 100$. Holding s fixed, the width is proportional to $t^*/\sqrt{n}$: four times the sample size roughly halves it, with an additional change because t^* depends on df.

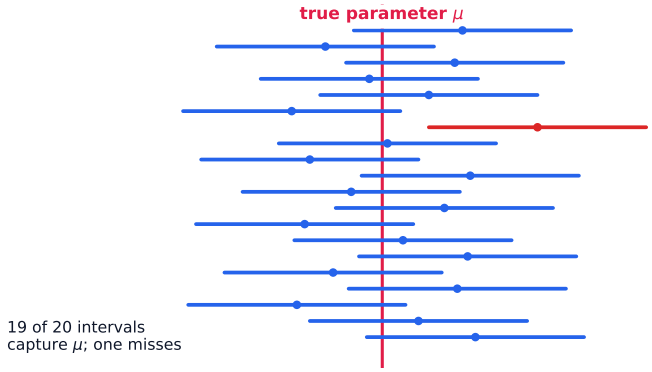
The Normal condition, three ways. Any one of these is enough: (1) the population is stated to be approximately normal; (2) $n \geq 30$, the usual AP large-sample guideline (extreme skew or outliers may require more data); (3) for a small sample, a dotplot or boxplot of the data shows no strong skew and no outliers. For two samples, check each sample separately.

Common mistakes. Using $z^* = 1.96$ when σ is unknown, or the wrong degrees of freedom (n instead of $n - 1$). Writing "$n \geq 30$" for the Normal condition when $n = 25$: with a small sample, say what a graph of the sample shows (no strong skew, no outliers).



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

exercises.lol exercises.com.cn



CI form: $\bar{x} \pm z^* \frac{\sigma}{\sqrt{n}}$ · level = long-run capture rate of the method

"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** 可信; **outside** the interval, the data give evidence against it. Answer in context using the plausible range.

A two-sided test at level α and a $(1-\alpha)$ confidence interval agree. If the 95% interval for the mean number of bedrooms in new houses is $(3.0,\ 3.4)$, then 2.9 is outside it, and a two-sided test of $H_0 : \mu = 2.9$ at $\alpha = 0.05$ would reject H_0. The interval also shows how big the difference might be, which a test alone does not.

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 统计量:

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

exercises.lol exercises.com.cn

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** 拒绝 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.

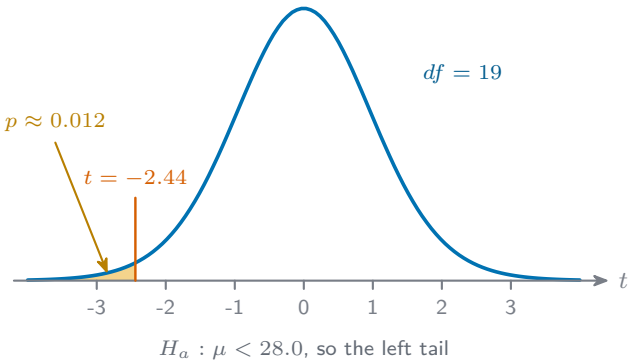
Carrying Out a Test for a Mean

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.

Worked example: a complete four-step test. The chip bags: a random sample of $n = 20$ bags has $\bar{x} = 27.4$ g and $s = 1.1$ g. At $\alpha = 0.05$, is there convincing evidence that the machine underfills?

- **State.** $H_0 : \mu = 28.0$ and $H_a : \mu < 28.0$, where μ is the true mean fill of all bags from this machine.
- **Plan.** One-sample t-test for a mean. Random: the bags were randomly selected. 10%: 20 bags is less than 10% of all bags filled. Normal: $n = 20 < 30$, so check the sample's dotplot: it shows no strong **skew** 偏态 and no outliers.
- **Do.** $t = \frac{27.4 - 28.0}{1.1/\sqrt{20}} = -2.44$, $df = 19$, $p \approx 0.012$.
- **Conclude.** Because $p \approx 0.012 < \alpha = 0.05$, reject H_0. There is convincing evidence that the true mean fill of bags from this machine is less than 28.0 g.

Every step names the **parameter** 参数 in context. "Reject H_0" alone, with no sentence about bags of chips, does not earn the conclusion point.



The alternative is mu < 28.0, so the p-value is the left-tail area beyond t = -2.44: about 0.012

What could go wrong. A **Type I error** 第一类错误 here means concluding the machine underfills when its true mean really is 28.0 g: the factory stops a good machine. A **Type II error** 第二类错误 means failing to detect underfilling that is really happening: customers keep getting short bags. The probability of a Type I error is α. **Power** 检验效能, the chance of correctly detecting a real difference, rises with a larger sample, a larger α, or a larger true difference.

> **Worked example: paired data.** Six students take a test before and after a revision course. Their score changes (after minus before) are 4, 2, 6, −1, 3, 4. These are paired: the same students twice. So analyse the six **differences** as one sample: $\bar{d} = 3.0$, $s_d = 2.37$, so
>
> $$t = \frac{3.0 - 0}{2.37/\sqrt{6}} = 3.11, \qquad df = 5, \qquad p \approx 0.013 \text{ (one-sided)}.$$
>
> Provided the population score differences are approximately normal, at $\alpha = 0.05$ there is convincing evidence that the true mean score change is positive. With only six students, check the differences for strong skew and outliers. This before-and-after comparison alone does not establish that the course caused the improvement.

Define the parameter and the order. For paired data, the parameter is μ_d, the **mean difference** in the population of pairs, and you must say which way you subtract: "μ_d = the true mean of (after − before) scores for all students who take the course". The sign of $\bar{d}$ only makes sense with that order.

exercises.lol exercises.com.cn

> **Worked example (car prices).** A researcher picks car models at random and, for each model, records the price paid by one randomly chosen man and one randomly chosen woman. Each man's price is paired with a woman's price for the **same model**, and models differ a lot in price. So analyse the differences (woman − man) with a paired t-test of $H_0 : \mu_d = 0$ against $H_a : \mu_d > 0$. A two-sample test would let the big differences between models hide the small difference between men and women.

Common mistakes. Running a two-sample test on paired data: the two lists are not independent, because each value in one list belongs with one value in the other (the same student). The reverse error is pairing two independent groups just because they have the same size.

Confidence Interval for a Difference of Two Means

For **independent samples** 独立样本, 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** 合并 the **variances** 方差 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.

> **Worked example: two independent samples.** Thirty plants get a new fertiliser and 30 different plants get the old one, assigned at random. New: $\bar{x}_1 = 24.1$ cm, $s_1 = 3.2$. Old: $\bar{x}_2 = 22.0$ cm, $s_2 = 3.6$.
>
> $$t = \frac{24.1 - 22.0}{\sqrt{\frac{3.2^2}{30} + \frac{3.6^2}{30}}} = \frac{2.1}{0.879} = 2.39, \qquad df \approx 57, \qquad p \approx 0.020 \text{ (two-sided)}.$$
>
> At $\alpha = 0.05$ there is convincing evidence of a difference in mean height. Because the fertiliser was **randomly assigned** 随机分配, the experiment supports a **cause-and-effect** 因果关系 conclusion: the new fertiliser causes taller plants, on average, for plants like these.

The matching 95% interval uses the same standard error, with $t^* \approx 2.002$ for $df \approx 57$:

$$(24.1 - 22.0) \pm 2.002(0.879) = 2.1 \pm 1.76 = (0.34,\ 3.86) \text{ cm}.$$

It excludes 0, which agrees with the test: the new fertiliser's mean height is plausibly between 0.34 and 3.86 cm greater.

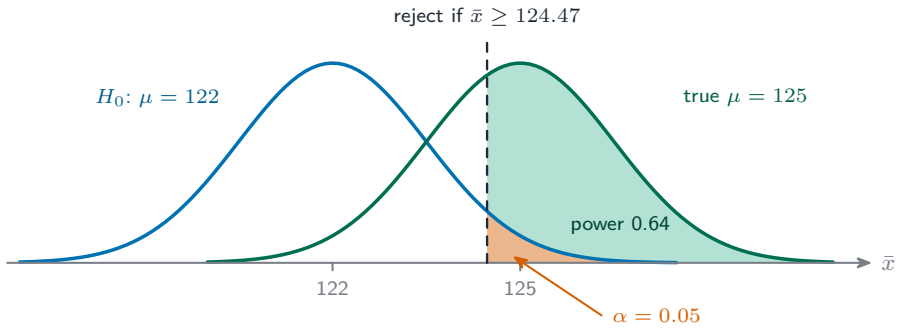
Degrees of freedom for two samples. Technology gives a df from a long formula (often not a whole number). By hand, a safe choice is the smaller of $n_1 - 1$ and $n_2 - 1$; it gives a slightly wider interval or larger p-value.

Two-sample or paired? Use two independent samples when the two groups contain **different** individuals with no natural link: for example, 20 patients randomly assigned to dark chocolate and 20 **other** patients to white chocolate. Use paired data only when each value in one group belongs with one value in the other: the same person twice, twins, or matched units.

exercises.lol exercises.com.cn

Statistical versus practical significance. A tiny p-value says the difference is unlikely to be chance; it does not say the difference is large enough to matter. With 5000 students per group, a new method might raise the mean test score by 0.4 points out of 100 with $p = 0.001$: statistically significant, but too small to be worth changing the course. Report the size of the effect (a confidence interval) as well.

> **Worked example (power for a mean).** A company tests $H_0 : \mu = 122$ against $H_a : \mu > 122$ for employees' mean blood pressure, with $\sigma = 15$, $n = 100$ and $\alpha = 0.05$. The test rejects H_0 when $\bar{x} \geq 122 + 1.645\left(\frac{15}{\sqrt{100}}\right) = 124.47$. If the true mean is 125, then
>
> $$P(\bar{x} \geq 124.47) = P\left(z \geq \frac{124.47 - 125}{1.5}\right) = P(z \geq -0.35) = 0.64.$$
>
> This probability of correctly rejecting the false H_0 is the **power**. With a larger sample, $\sigma_{\bar{x}}$ shrinks and the power increases.



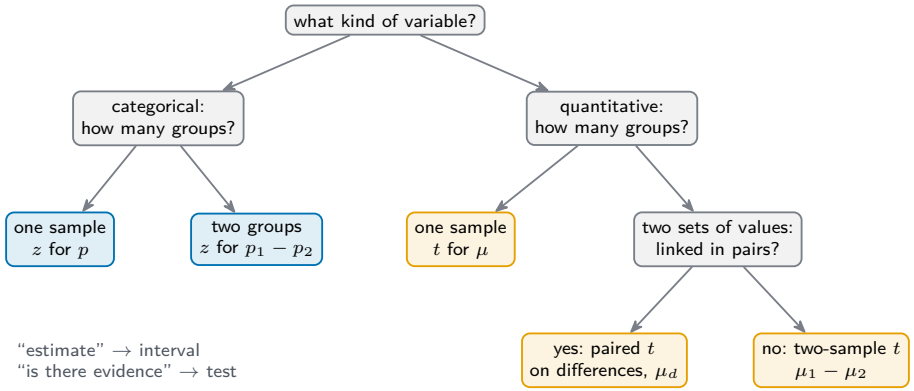
Power is the area of the true-mean curve beyond the rejection cut-off: 0.64 when the true mean is 125

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.

Data	Parameter	Procedure
one categorical variable, one sample	p	one-sample z for a proportion
one categorical variable, two groups	$p_1 - p_2$	two-sample z for a difference of proportions
one quantitative variable, one sample	μ	one-sample t for a mean
paired measurements	μ_d	paired t (one-sample t on the differences)
one quantitative variable, two independent groups	$\mu_1 - \mu_2$	two-sample t for a difference of means

"Estimate" or "how large" asks for an **interval**; "is there evidence" asks for a **test**. Always define the parameter in words, in context, before you start.



The table as a decision tree: type of variable, number of groups, then paired or independent

The investigative task (Question 6 through 2026). Starting with the 2027 exam, there are four free-response questions; the old Question 6 format no longer applies. See College Board's current exam format. The older investigative task took you a step past the course. It gives you something new, such as a new statistic with its formula, a table for judging its size, or the results of a simulation, and asks you to use it with ideas you already know. You are not expected to have seen it before. Read the definition carefully, show your working, and explain each answer in context.

exercises.lol exercises.com.cn

> **Worked example (an original investigative task).** A café tests whether a new grinder makes espresso shot times more **consistent** 一致的. With the old grinder, 12 randomly chosen shots had mean 27.0 s and standard deviation 3.0 s; with the new grinder, 12 shots had mean 26.4 s and standard deviation 1.8 s.

(a) *Why would a two-sample t-test not answer the café's question?* A t-test compares **means** (the centre); the café wants to know about **spread**. Two grinders could give the same mean shot time with very different variability.

(b) *The manager defines the statistic $R = \frac{s_{old}}{s_{new}}$. Calculate R and say what a large value would suggest.* $R = \frac{3.0}{1.8} = 1.67$. A value well above 1 suggests the new grinder's shot times vary less than the old grinder's.

(c) *A simulation assumed the two grinders are equally consistent and produced 200 values of R. Nine of them were 1.67 or greater. Is there convincing evidence, at $\alpha = 0.05$, that the new grinder gives more consistent shot times?* The estimated p-value is $\frac{9}{200} = 0.045$. Because $0.045 < 0.05$, there is convincing evidence that the true standard deviation of shot times is smaller with the new grinder than with the old one.

(d) *Suppose the new grinder's standard deviation had been 2.5 s instead. How would R and the evidence change?* $R = \frac{3.0}{2.5} = 1.2$, which is smaller. Many more simulated values would be 1.2 or greater, so the p-value would be larger and the evidence weaker.

Part (d) illustrates the older investigative-task reasoning: say which way the statistic moves, explain why from its formula, then say what that does to the conclusion.

Practice

Answer each before reading the answers below.

1. Assume the population is approximately normal. A random sample of 16 values has $\bar{x} = 12.5$ and $s = 2.0$. Construct a 95% confidence interval for μ ($t^* = 2.131$ for $df = 15$).
2. Interpret the interval from question 1 in context, for a sample of 16 phone batteries measured in hours of use.
3. Name the procedure: (a) 20 people's reaction times, with and without caffeine; (b) heights of a random sample of 50 men and a separate random sample of 50 women; (c) the proportion of voters who support a plan.
4. In the chip-bag test, describe a Type II error and its consequence.
5. A student says: "The p-value of 0.012 means there is a 1.2% chance that H_0 is true." Correct the student.
6. Twenty runners each run 5 km once in old shoes and once in new shoes, in a random order. Name the procedure, define the parameter, and say why it is appropriate.

7. A 95% interval for a mean is $(12.1,\ 14.3)$. What would a two-sided test of $H_0: \mu = 14$ at $\alpha = 0.05$ conclude?

Answers

1. The **standard error** 标准误 is $SE = 2.0/\sqrt{16} = 0.50$, so $12.5 \pm 2.131(0.50) = 12.5 \pm 1.07$, giving $(11.4,\ 13.6)$.
2. We are 95% confident that the interval from 11.4 to 13.6 hours captures the true mean battery life of all phones of this type.
3. (a) Paired t on the differences; (b) two-sample t for $\mu_1 - \mu_2$; (c) one-proportion z procedure.
4. Failing to find convincing evidence of underfilling when the machine really does underfill. The fault goes unfixed and customers keep receiving underweight bags.
5. The p-value assumes H_0 is true. It is the probability of getting a sample mean at least as low as 27.4 g **if** the true mean were 28.0 g. It says how surprising the data would be under H_0, not how likely H_0 is.
6. A paired t-test (or interval) on the differences. Parameter: μ_d, the true mean of (old − new) times for all runners like these. Each runner gives two times, so the times are paired; using differences removes the large variation between runners.
7. 14 is inside the interval, so the test would **fail to reject** H_0: 14 is a plausible value for μ.

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 $n - 1$).
- For paired data, define μ_d with the order of subtraction, then run a one-sample t on the differences.
- Separate **statistical** significance (small p-value) from **practical** importance (a difference big enough to matter).
- In older investigative-task practice, use the new statistic exactly as it is defined, show the calculation, and when a value changes, say which way the statistic moves and why.

exercises.lol exercises.com.cn

Inference for Categorical Data: Chi-Square

AP Statistics

Are My Results Unexpected?

Do Grade 10 and Grade 11 students study in different places? Ask 200 of them and the counts will never split perfectly evenly, even if grade makes no difference at all. The question this unit answers is how uneven the counts must be before chance stops being a good explanation.

When data are **counts** 频数 spread across several **categories** 类别, we test whether the **observed counts** 观测频数 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** 卡方分布 is **right-skewed** 右偏的 and depends on its **degrees of freedom** 自由度.

Setting Up a Goodness-of-Fit Test

Extension: chi-square goodness-of-fit was removed from the AP Statistics exam curriculum effective 2026–27. Homogeneity and independence remain assessed. See College Board's revisions.

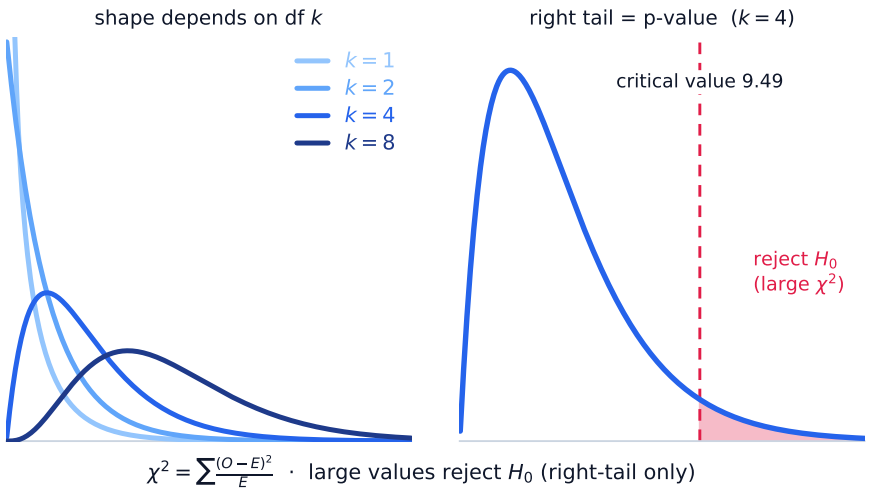


Mixed jelly beans: do the colours appear in the proportions the maker claims? A goodness-of-fit test compares observed and expected counts

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

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

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

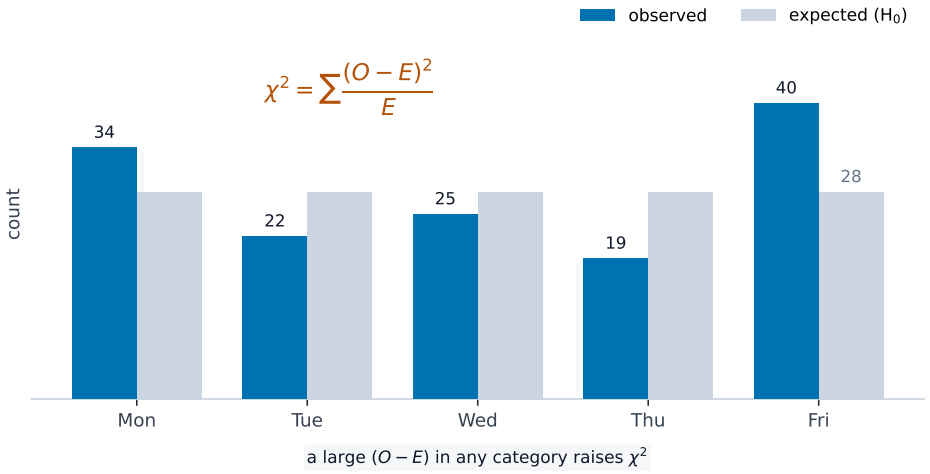


The chi-square distribution is right-skewed. A large statistic lands in the shaded right tail past the critical value – that is where you reject the model.

Carrying Out a Goodness-of-Fit Test

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

exercises.lol exercises.com.cn



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** 不拒绝 H_0 – no evidence the die is unfair.

Worked example: unequal claimed proportions. A sweet maker claims its bags are 30% red, 20% green and 50% yellow. A random sample of 200 sweets has 70 red, 30 green and 100 yellow.

Colour	Red	Green	Yellow
Observed	70	30	100
Expected	$200(0.30) = 60$	$200(0.20) = 40$	$200(0.50) = 100$

$$\chi^2 = \frac{(70-60)^2}{60} + \frac{(30-40)^2}{40} + \frac{(100-100)^2}{100} = 1.67 + 2.50 + 0 = 4.17, \qquad df = 3 - 1 = 2.$$

The p-value is about $0.12 > 0.05$, so fail to reject H_0: there is not convincing evidence that the colour distribution differs from the maker's claim. The expected counts come from the **claimed** percentages, not from dividing 200 equally by 3.

Interpreting the *p*-value. "Assuming the colours really are 30%, 20% and 50%, there is about a 0.12 probability of getting a χ^2 statistic of 4.17 or larger in a random

sample of 200 sweets." The null distribution can be the χ^2 model or a simulation of many samples from the claimed distribution; either gives the right-tail area beyond the observed statistic.

Expected Counts in Two-Way Tables

For a **two-way table** 双向表, 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 (no **association** 关联).

> **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.

f_x `=$D2*B$4/$D$4`

	A	B	C	D
1		**likes PE**	**dislikes PE**	**total**
2	**boys**	36	14	50
3	**girls**	24	26	50
4	**total**	60	40	100

expected counts

	likes	**dislikes**
boys	30	20
girls	30	20

expected = row total × column total ÷ grand total = 50 × 60 ÷ 100 = 30

type the formula once, then fill it across: the spreadsheet does every cell

A spreadsheet organises categorical counts before a chi-square test

exercises.lol exercises.com.cn

	likes PE	**dislikes PE**	**total**
boys	36 (30)	14 (20)	50
girls	24 (30)	26 (20)	50
total	60	40	100

$$\text{expected (boys, likes)} = \frac{50 \times 60}{100} = 30$$

observed (expected)

Expected count = row total × column total ÷ grand total

Homogeneity or Independence?

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

- **Test for homogeneity** 同质性检验: are the distributions of one categorical variable the **same across several populations or groups** (separate samples/treatments)?
- **Test for independence** 独立性检验: 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.

Writing the hypotheses. For homogeneity, the null says the proportions in each category are the **same** in every group. For three cities and one yes/no question: $H_0 : p_A = p_B = p_C$, where each p is the proportion of all teens in that city who drank a soft drink in the past week; H_a: at least two of these proportions differ. For independence: H_0: there is no association between the two variables in the population; H_a: there is an association.

Experiments. A randomized experiment with a categorical response is analysed with the same chi-square test for homogeneity: the treatment groups play the part of the populations. If children are randomly assigned to watch one of three ads and then choose a snack, a small p-value is evidence that the type of ad **causes** a difference in snack choice, for children similar to those in the study.

HOMOGENEITY	INDEPENDENCE
separate random samples from several populations (or groups); one categorical variable	one random sample from one population; two categorical variables recorded
H_0: the distribution is the same in every population	H_0: the two variables are independent (no association)

Homogeneity compares populations on one variable; independence asks about two variables in one sample

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).

Worked example: a test for independence. A random sample of 200 students at one school is classified by grade and by where they usually study.

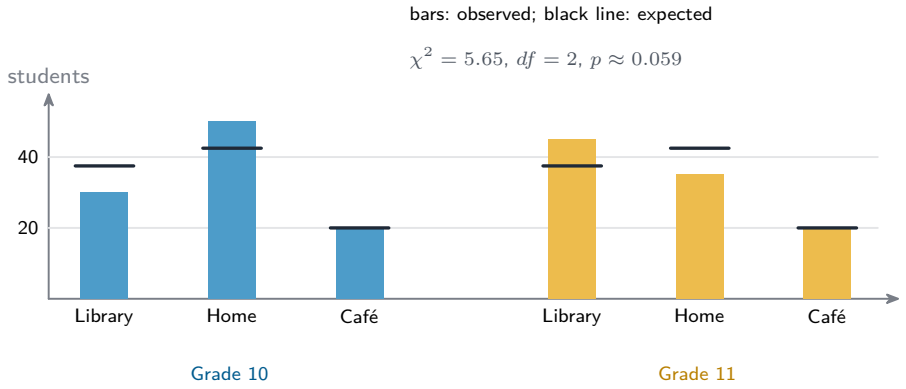
	Library	Home	Café	Total
Grade 10	30	50	20	100
Grade 11	45	35	20	100
Total	75	85	40	200

- **State.** H_0: grade and study place are independent for students at this school. H_a: they are associated.
- **Plan.** Chi-square test for independence: **one** sample, **two** variables. Random sample; 200 is less than 10% of the students if the school has more than 2000; expected counts below are all at least 5.
- **Do.** Expected counts: $\dfrac{100 \times 75}{200} = 37.5$, then 42.5 and 20 in each row. The six components are 1.50, 1.32, 0, 1.50, 1.32, 0, so $\chi^2 = 5.65$ with $df = (2-1)(3-1) = 2$, giving $p \approx 0.059$.

exercises.lol exercises.com.cn

- **Conclude.** Because $p \approx 0.059 > \alpha = 0.05$, fail to reject H_0. There is **not** convincing evidence of an association between grade and study place for students at this school.

The library cells contribute most: Grade 11 students chose the library more often than expected. That is worth noting as a pattern, but with this p-value it is not convincing evidence.



Observed against expected: the library and home counts miss by 7.5 in each grade, giving chi-square 5.65, not enough at the 0.05 level

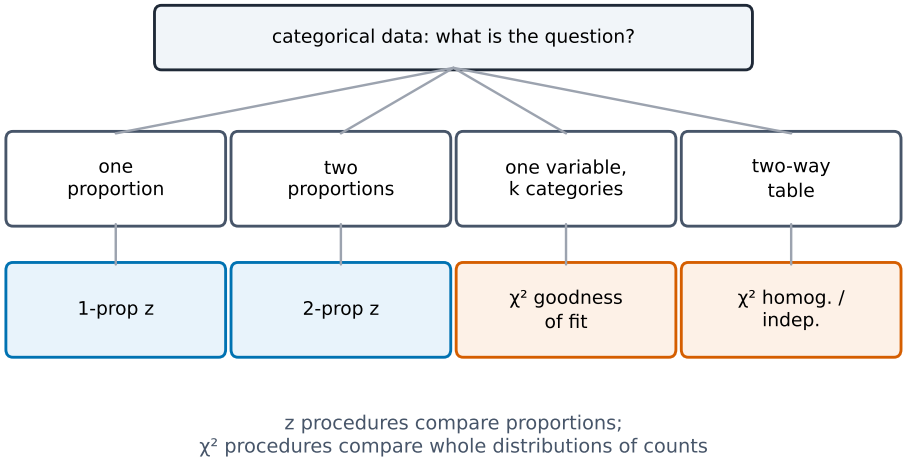
Common mistakes. Checking the condition with **observed** counts: it is the **expected** counts that must all be at least 5. Running chi-square on percentages or means instead of **counts**. Using $df = n - 1 = 199$ instead of $(r-1)(c-1) = 2$. Writing the conclusion about the sample ("these students are associated") instead of the population the sample represents.

Is the test appropriate? Before any chi-square test, compute the smallest expected count. If a table of athletes has a cell with expected count $\frac{12 \times 18}{80} = 2.7$, the large-counts condition fails, so the χ^2 p-value would not be reliable. Collect more data, or combine categories in a sensible way (for example, merge two small age groups) until every expected count is at least 5. Writing H_0 as "grade and study place are associated": the null hypothesis always says no association. Calling a one-sample, two-variable test "homogeneity".

Choosing the Right Categorical Procedure

Decide by the setup: one categorical variable against a claimed distribution ⇒ **goodness-of-fit**; one sample cross-classified by two variables ⇒ **independence**; several samples/groups compared ⇒ **homogeneity**. Comparing just **two** proportions can use either a two-proportion z-test or a chi-square test, but only for a **two-tailed** 双尾 alternative, where they agree exactly ($\chi^2 = z^2$). The chi-square statistic is

compared with the **upper tail** of its reference distribution. For two proportions it tests a **non-directional** alternative, so if H_a is **one-sided** 单侧 (say $p_1 > p_2$), use the z-test.



Choosing a categorical inference procedure

Practice

Answer each before reading the answers below.

1. In a two-way table, a cell's row total is 60, its column total is 50 and the grand total is 200. Find its expected count.
2. A two-way table has 3 rows and 4 columns. Find the degrees of freedom for a chi-square test.
3. A die is rolled 60 times, giving counts 6, 10, 8, 12, 14, 10 for faces 1 to 6. Find χ^2 for a goodness-of-fit test of a fair die. Its p-value is about 0.55. Conclude.
4. Name the test: (a) one random sample of 500 adults, classified by phone brand and age group; (b) separate random samples from three cities, comparing favourite sport; (c) the colours in a bag of sweets, compared with the maker's claimed percentages.
5. A student says: "In the worked example $p = 0.059$, so grade and study place are independent." Correct the student.
6. A randomized experiment assigns 150 patients to three treatments and records whether each improves. Name the test and write its hypotheses.
7. A two-way table has row totals 30 and 70, column totals 15, 40 and 45, and grand total 100. Is a chi-square test appropriate? Check the smallest expected count.

exercises.lol exercises.com.cn

Answers

1. $\dfrac{60 \times 50}{200} = 15$.
2. $(3-1)(4-1) = 6$.
3. Each expected count is 10. $\chi^2 = \dfrac{16+0+4+4+16+0}{10} = 4.0$, $df = 5$. With $p \approx 0.55$, fail to reject H_0: no convincing evidence that the die is unfair.
4. (a) Independence; (b) homogeneity; (c) goodness-of-fit.
5. Failing to reject H_0 means the data are not convincing evidence of an association. It does not prove independence: a real but small association could still exist. Write "not convincing evidence of an association", never "they are independent".
6. A chi-square test for homogeneity. H_0: the proportion of patients who improve is the same for all three treatments; H_a: at least two of the proportions differ. With random assignment, a significant result supports a cause-and-effect conclusion.
7. Smallest expected count $= \frac{30\times15}{100} = 4.5 < 5$, so the condition fails and the chi-square test is not appropriate as it stands.

Exam tips

- Use $\chi^2 = \sum \frac{(O-E)^2}{E}$ for **categorical** data; always divide by the **expected** count.
- For exams from 2027, distinguish independence and homogeneity (two-way tables); goodness-of-fit is extension content.
- 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 $(r-1)(c-1)$).
- Before testing, find the **smallest expected count**; if it is below 5, the chi-square test is not appropriate.
- Write the conclusion about the **population** (or the treatments), never just about the sample.

exercises.lol exercises.com.cn

Inference for Quantitative Data: Slopes

AP Statistics

Do Those Points Align?

Extension: inference for regression slopes was removed from the AP Statistics exam curriculum effective 2026–27. This handout retains the earlier unit numbering. See College Board's revisions.

Students who revise longer tend to score higher, on average, in one class's data. But a class is one sample. Would another class show the same upward slope, or could the trend be an accident of who happened to be in this one? Inference for a slope answers that, and it starts from the computer output used in regression analysis.

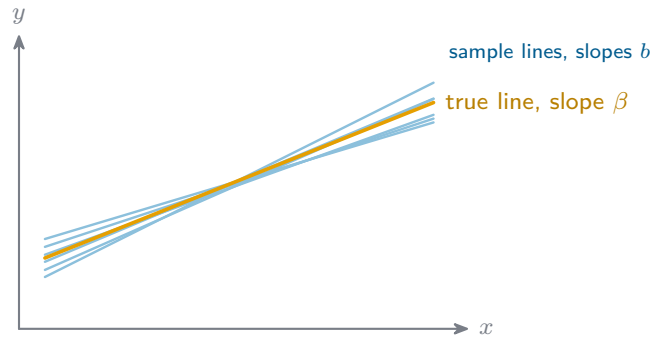
A sample **scatterplot** 散点图 gives a **sample slope** 样本斜率 b for the **least-squares regression line** 最小二乘回归线 – but a different sample would give a slightly different slope. So b is a statistic with **sampling variability** 抽样变异性, estimating the **true (population) slope** 总体斜率 β. This unit does **inference** 推断 for β: is there a real **linear** 线性 relationship, and how strong is it?

The model is $\mu_y = \alpha + \beta x$: for each x, the responses y vary around a true line, normally and with the same standard deviation σ. Points scatter around the line partly at random. The question is whether the trend in the sample is larger than that random scatter could produce.

The sampling distribution of b. Over many random samples, b is centred on β (it is unbiased). Its standard deviation is estimated by

$$SE_b = \frac{s}{s_x\sqrt{n-1}},$$

where s is the standard deviation of the residuals and s_x the standard deviation of the x-values. The slope is estimated more precisely when the points sit close to the line (small s), when the x-values are spread out (large s_x), and when n is large.



schematic: each random sample gives a slightly different b, centred on β

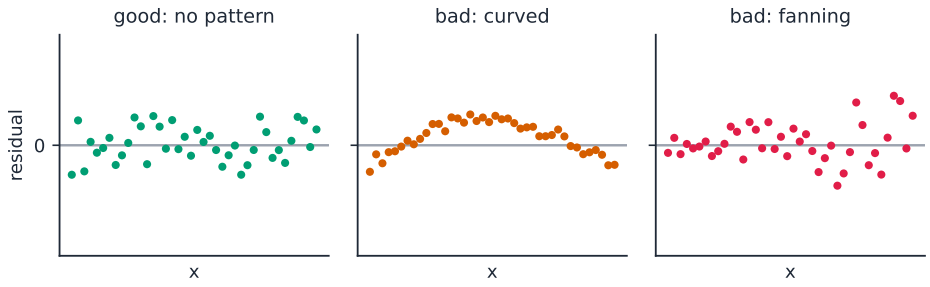
Schematic: the sample slope b varies from sample to sample around the true slope beta, which is why b needs a standard error

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** 标准误 (read from computer output). **Conditions** 条件 (**LINER**): the true relationship is **Linear**, observations **Independent**, **residuals** 残差 **Normal**, and residuals have **Equal** spread (check the residual plot and a **histogram** 直方图 of residuals), from **Random** data. Interpret the interval for β **in context** 结合情境, with units of y per unit of x.



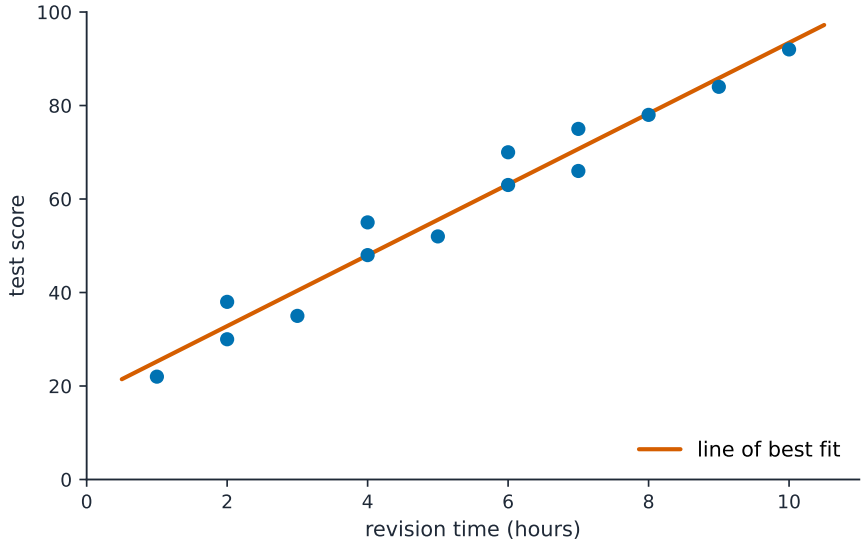
residual $= y - \hat{y}$ · random scatter about 0 supports a linear model (LINER)

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

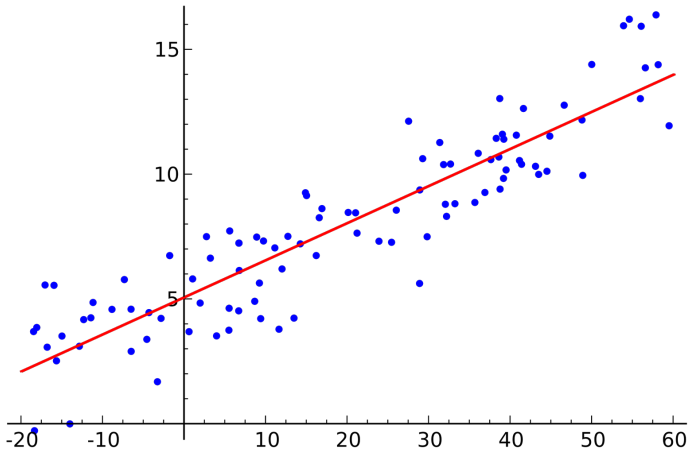
The **residual plot** 残差图 is where you check Linear and Equal-spread: you want a formless cloud around zero. A **curve** 曲线 means the relationship is not linear; a **fan** 扇形 (spread growing with x) means the residuals do not have equal spread – both break a condition.

exercises.lol exercises.com.cn

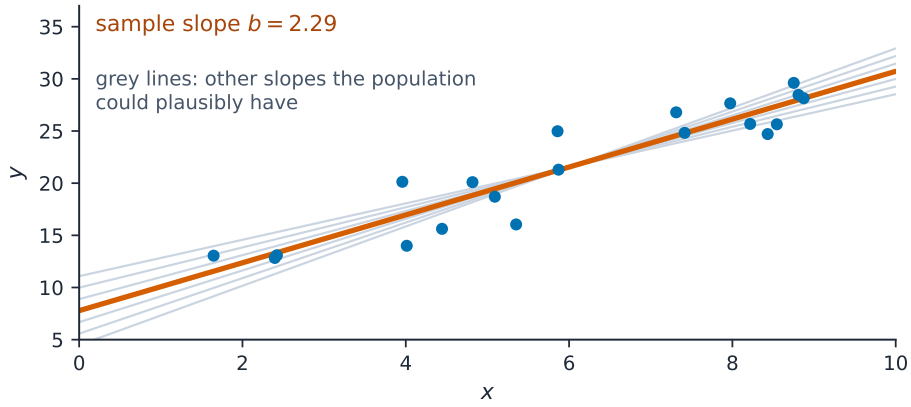
> **Worked example.** Regression output gives slope $b = 2.5$ with $SE_b = 0.8$ from $n = 20$ points. For a 95% interval, $df = 18$ gives $t^* = 2.101$:
>
> $$2.5 \pm 2.101(0.8) = 2.5 \pm 1.68 = (0.82,\ 4.18).$$
>
> Because 0 is **not** in the interval, there is evidence of a positive linear relationship.



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



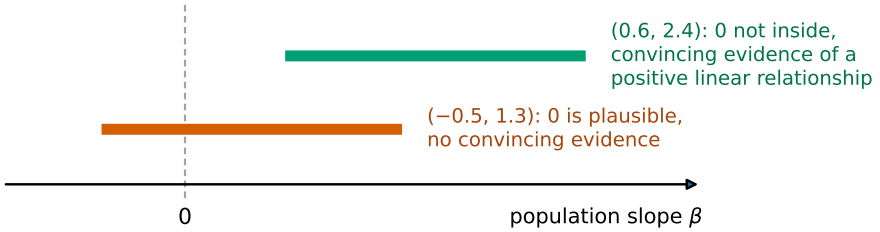
$b \pm t^* SE_b$ with $df = n - 2$: an interval for the population slope β

A confidence interval gives the plausible values of the population slope

Justifying a Claim About a Slope

If the **confidence interval** 置信区间 for β contains 0, a slope of zero is **plausible** 可信 – 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.

exercises.lol exercises.com.cn



If 0 is outside the interval for β there is convincing evidence of a linear relationship

Interpreting the interval. "We are 95% confident that the interval from 0.82 to 4.18 captures the slope of the population regression line: for each one-unit increase in x, the mean of y increases by between 0.82 and 4.18 units." Larger samples give narrower intervals, other things equal.

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.

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** 回归 output. Find the p-**value** P 值 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. Standard regression output usually 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.

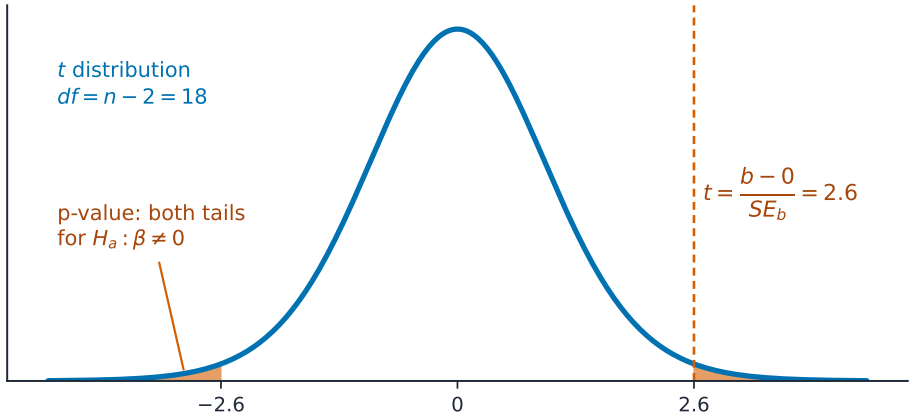
> **Worked example.** For the same output ($b = 2.5$, $SE_b = 0.8$, $n = 20$), test $H_0 : \beta = 0$:
>
> $$t = \frac{2.5 - 0}{0.8} = 3.13, \qquad df = 18,$$
>
> a small p-value (< 0.01), so **reject** H_0 – **convincing evidence** 有力证据 of a linear relationship. This matches the interval, which excluded 0.

Testing a slope other than 0. For all male sambar deer, the population slope of weight against chest circumference is 4.5 kg/cm. A biologist measures 20 male tule elk and finds $b = 5.4$ kg/cm with $SE_b = 0.35$. Is the elk slope different from 4.5? Test $H_0 : \beta = 4.5$ against $H_a : \beta \neq 4.5$:

$$t = \frac{5.4 - 4.5}{0.35} = 2.57, \qquad df = 18, \qquad p \approx 0.019.$$

Since $0.019 < 0.05$, reject H_0: there is convincing evidence that the slope for male tule elk differs from 4.5 kg/cm. Note that the printed p-value in computer output is for $\beta = 0$ only; for any other value, compute t yourself.

Interpreting the p-value. "Assuming the true slope is 4.5, there is a 0.019 probability of getting a sample slope at least as far from 4.5 as 5.4 in a random sample of 20 elk."



The t statistic for the slope and its two-sided p-value

Reading regression output

Regression software can print output like this, from a random sample of 24 students, with x = hours of revision and y = test score:

Predictor	Coef	SE Coef	T	P
Constant	48.2	4.10	11.76	0.000
Hours	3.05	0.92	3.32	0.003

$S = 6.84$ and $R\text{-}Sq = 33.3\%$.

- **The line:** predicted score $= 48.2 + 3.05$ (hours). The slope row, labelled with the x variable, is the one that matters for inference.
- **The interval:** $df = 24 - 2 = 22$ and $t^* = 2.074$, so $3.05 \pm 2.074(0.92) = 3.05 \pm 1.91$, giving (1.14, 4.96). *We are 95% confident that, for each extra hour of revision, the predicted score rises by between 1.1 and 5.0 points, on average, for students like these.*
- **The test:** $t = 3.05/0.92 = 3.32$ and the printed $p = 0.003$ (**two-sided** 双侧). That is convincing evidence of a linear relationship between revision hours and score.
- S is the typical size of a residual: predictions miss by about 6.8 points. (From $SE_b = \frac{s}{s_x\sqrt{n-1}}$, the revision hours had $s_x = \frac{6.84}{0.92\sqrt{23}} \approx 1.55$ hours.) R^2: 33.3% of the variation in scores is explained by the linear relationship with hours. The rest comes from everything else.

> **Worked example: checking the conditions in context.** Suppose the residual plot for the revision data is a patternless band of roughly constant width, and a histogram of the residuals is roughly symmetric with no outliers. A full-credit check names each condition and the evidence for it:

- **Linear:** the residual plot shows no curved pattern, so a straight line is a suitable model.
- **Independent:** the 24 students were a random sample, and 24 is less than 10% of all students at the school, so the observations are independent.
- **Normal:** the histogram of residuals is roughly symmetric with no outliers, so Normal residuals are plausible.
- **Equal spread:** the residuals do not fan out as hours increase, so the spread about the line is roughly constant.
- **Random:** the data come from a random sample.

"Conditions are met" or a list of the letters LINER earns nothing on its own: each condition needs its evidence from the plot or the design.

This was an **observational** 观察性的 sample, not an **experiment** 实验. The evidence shows an **association** 关联, not that revising causes higher scores: more motivated students might both revise longer and score higher.

Common mistakes. Reading the Constant row instead of the slope row; using $df = n - 1$ instead of $n - 2$; taking $S = 6.84$ as the standard error of the slope (that is $SE_b = 0.92$, in the SE Coef column); writing hypotheses about b (the sample slope) instead of β; and halving the printed p-value when H_a is two-sided.

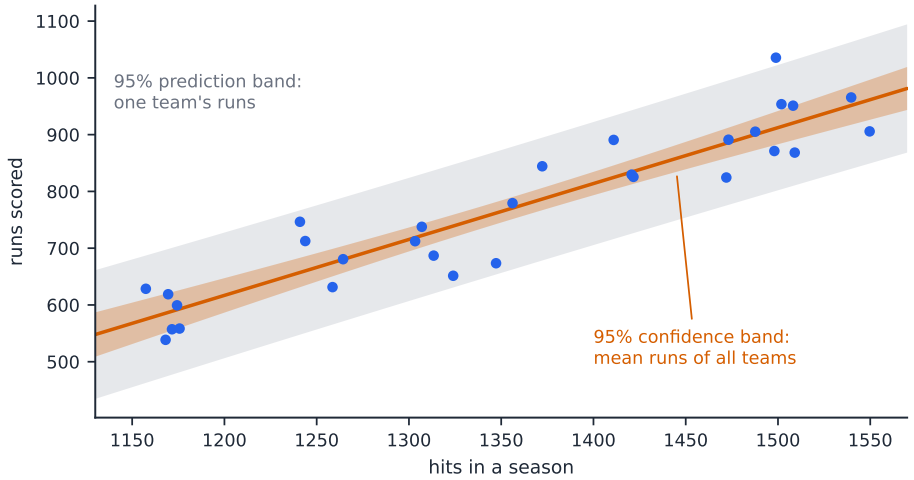
Two other intervals. Regression software can also give an interval for a **predicted value**. They answer different questions:

- a **confidence interval for the mean response** estimates the mean y of **all** individuals with a given x;
- a **prediction interval** estimates the y of **one** new individual with that x, so it must also allow for that individual's scatter about the line, and it is always wider.

exercises.lol exercises.com.cn

Worked example. For 30 teams, the line predicts 683 runs for a team with 1250 hits. Software gives $SE = 17.48$ for the mean response and $SE = 56.78$ for a single prediction; $df = 28$ gives $t^* = 2.048$.

- Mean runs for **all** teams with 1250 hits: $683 \pm 2.048(17.48) = 683 \pm 35.8$, so $(647,\ 719)$ runs.
- Runs for **one** team with 1250 hits: $683 \pm 2.048(56.78) = 683 \pm 116.3$, so $(567,\ 799)$ runs.

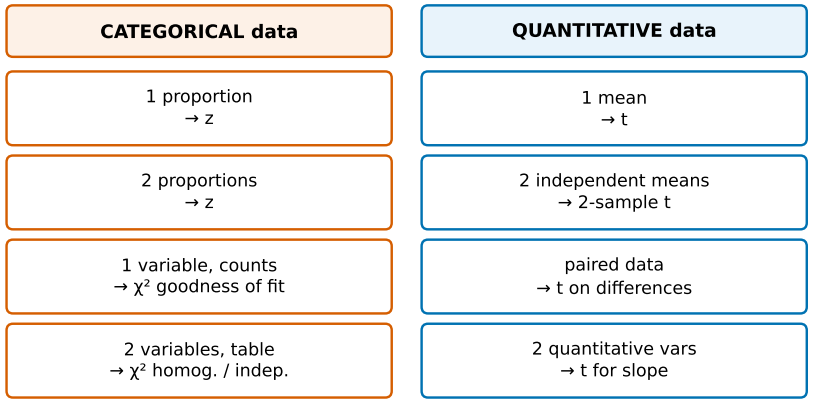


Predicting one team is far less certain than estimating the mean of all such teams; both intervals widen away from the mean of x

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** 配对; sample or experiment). Then name the **procedure** 检验方法, 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 was central to the investigative-task question on exams through 2026.

exercises.lol exercises.com.cn



then ask: interval (estimate) or test (claim)?

Choosing any inference procedure: type of data, number of groups, then estimate or test

Worked example. Name the procedure for each study.

Study	What is measured	Procedure
Does the proportion of left-handed students differ between two schools?	categorical, two independent samples	two-sample z test for $p_1 - p_2$
Estimate the mean resting heart rate of all teachers in a city	quantitative, one sample	one-sample t interval for μ
Each runner is timed in old shoes and in new shoes	quantitative, two values per runner	paired t test on the differences, μ_d
Are blood type and region of birth associated, in one sample of adults?	two categorical variables, one sample	chi-square test of independence
Does fertiliser amount (g) predict plant height (cm)?	two quantitative variables	t test or interval for the slope β

The clue is in the variables: a **mean** needs a quantitative response, a **proportion** a yes/no outcome, and a **slope** two quantitative variables measured on the same individuals. Two measurements on the **same** individual means paired, not two-sample.

Practice

Answer each before reading the answers below. Use the output in "Reading regression output".

1. Write the equation of the least-squares line in context.
2. Predict the score of a student who revises for 6 hours.
3. Another sample of 15 points gives slope $b = -1.8$ with $SE_b = 0.6$. Find the **test statistic** 检验统计量 and **degrees of freedom** 自由度 for $H_0 : \beta = 0$.
4. A residual plot shows a clear U-shaped curve. Which condition fails, and what does that mean for inference?
5. A student concludes from $p = 0.003$: "Revising longer causes higher test scores." Correct the student.
6. Output for $n = 25$ gives $b = 1.20$ with $SE_b = 0.40$. Test $H_0 : \beta = 2$ against $H_a : \beta < 2$.
7. A regression of house price on floor area is used to (a) estimate the mean price of all 120 m^2 houses and (b) predict the price of one particular 120 m^2 house. Which interval is wider, and why?

Answers

1. Predicted score $= 48.2 + 3.05$ (hours of revision).
2. $48.2 + 3.05(6) = 66.5$ points.
3. $t = \dfrac{-1.8 - 0}{0.6} = -3.0$, with $df = 15 - 2 = 13$.
4. The **Linear** condition fails: the true relationship is curved, so a slope interval or test for a straight line is not valid. A transformation or a different model is needed first.
5. The students were sampled, not randomly assigned to revise for different times, so this is observational data. The small p-value is convincing evidence of an association only. A confounding variable such as motivation could explain both.
6. $t = \dfrac{1.20 - 2}{0.40} = -2.0$ with $df = 23$; one-sided $p \approx 0.029$. At $\alpha = 0.05$, reject H_0: convincing evidence that the true slope is less than 2.
7. (b), the prediction interval. It must cover the price of one house, which varies around the line by about s, as well as the uncertainty in the line itself.

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.

exercises.lol exercises.com.cn

- Read the slope, standard error, t-statistic, and p-value from the regression output. For a one-sided alternative, halve the usual **two-sided** p-value only when the observed slope points in the alternative's direction; otherwise use one minus half that p-value.
- Check linearity and equal spread via the residual plot, normality via a histogram or dotplot of residuals, and independence via the study design.
- Interpret the interval and test in context, tied to the true slope.
- The printed t and p test $\beta = 0$; to test another value β_0, compute $t = \dfrac{b - \beta_0}{SE_b}$ yourself.
- Check each regression condition **with its evidence**: the residual plot (linear, equal spread), a histogram or dotplot of residuals (normal), and the design (random, 10%).
- A prediction interval (one individual) is always wider than a confidence interval for the mean response at the same x.