Exploring One-Variable Data

AP Statistics

Introducing Statistics: What Can We Learn from Data?

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

The Language of Variation: Variables

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

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

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

Representing a Categorical Variable with Tables

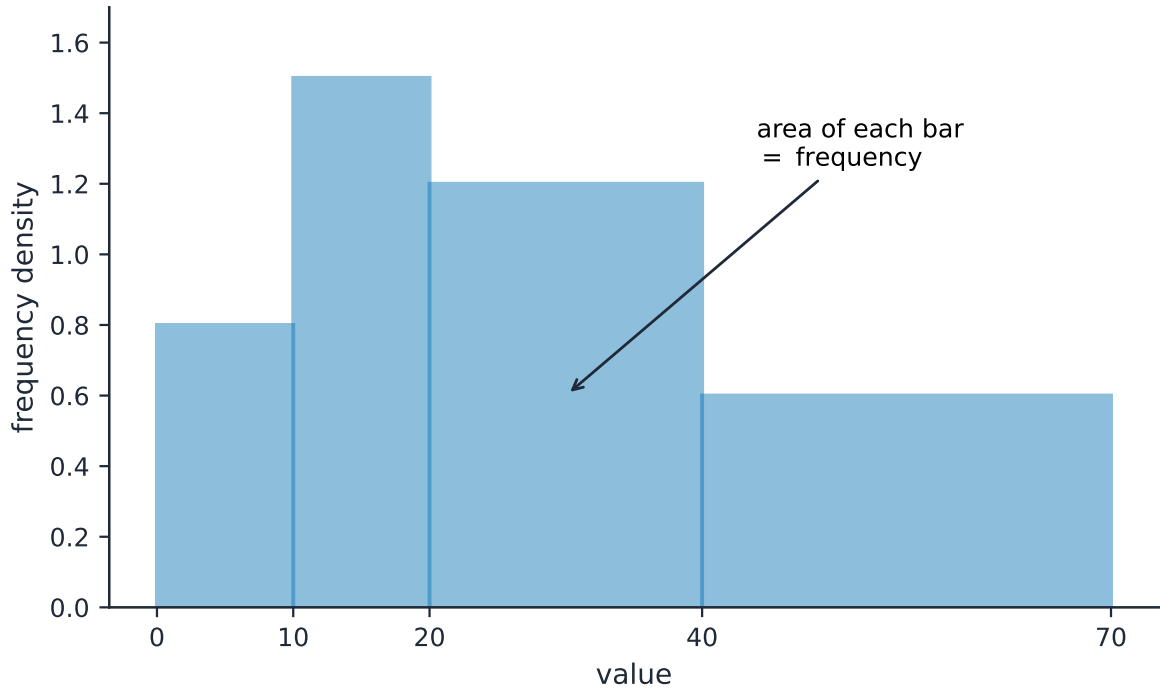
A **frequency table** 频数表 lists each category's **count** (frequency); a **relative frequency** 相对频率 table lists each category's **proportion** 比例 (count ÷ total). Relative frequencies let you compare groups of different sizes fairly.

Representing a Categorical Variable with Graphs

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

Representing a Quantitative Variable with Graphs

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



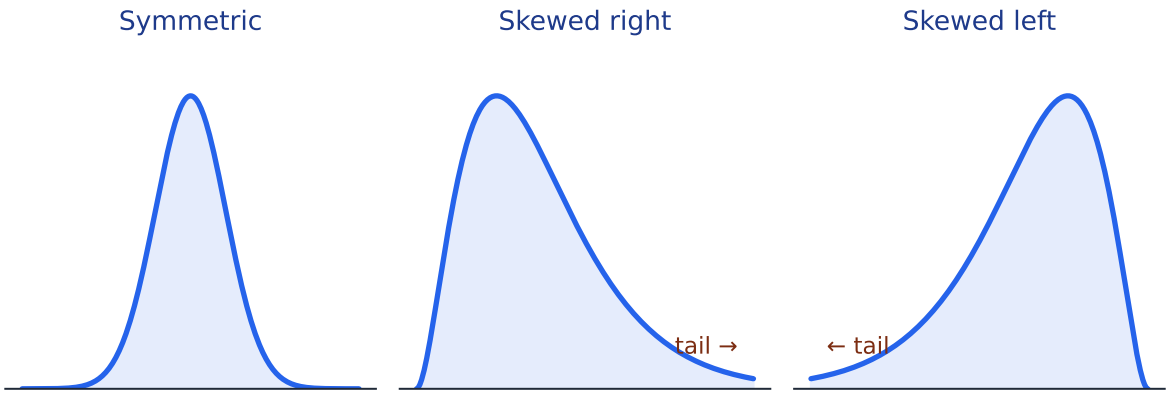
On a histogram with unequal class widths the bar area is the frequency

Describing the Distribution of a Quantitative Variable

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

- **Shape** 形状: symmetric, or **skewed** 偏斜 left/right (a long tail on that side), and how many peaks.
- **Outliers** 离群值: unusual values far from the rest.
- **Center**: a typical value (mean or median).
- **Spread**: how much the values vary (range, IQR, standard deviation).

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



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

Summary Statistics for a Quantitative Variable

- **Center:** the **mean** 均值 $\bar{x} = \dfrac{\sum x_i}{n}$ (average) and the **median** 中位数 (middle value). The median resists outliers; the mean is pulled toward a skew.
- **Spread:** the **range**, the **interquartile range** 四分位距 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.

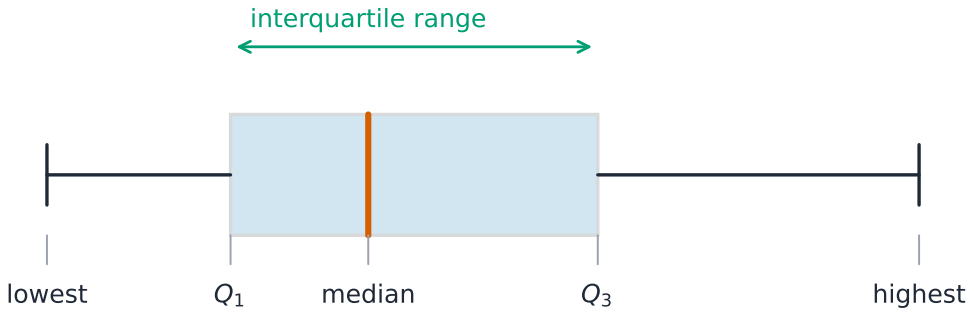
Use **resistant** measures (median, IQR) for skewed data; mean and standard deviation for roughly symmetric data.

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

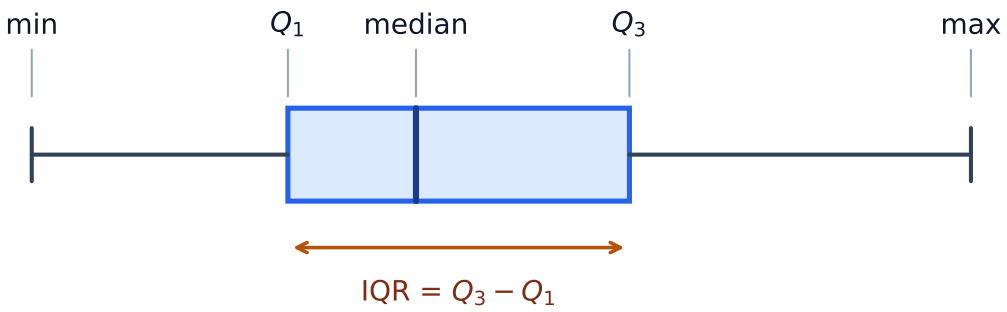
Graphical Representations of Summary Statistics

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

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



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



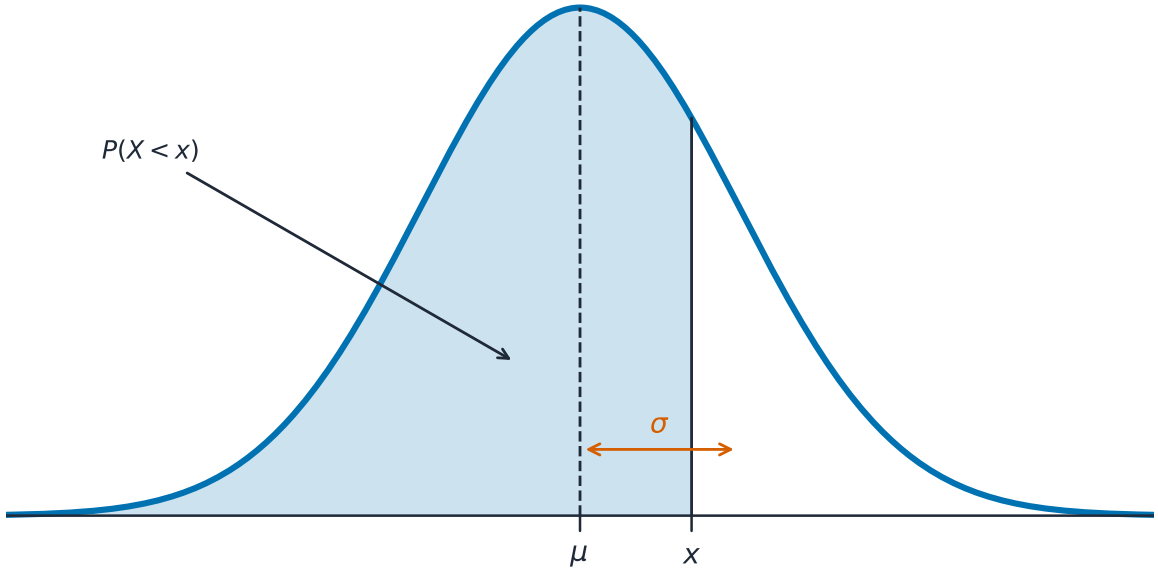
A boxplot draws the five-number summary; the box spans the IQR

Comparing Distributions of a Quantitative Variable

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

The Normal Distribution

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



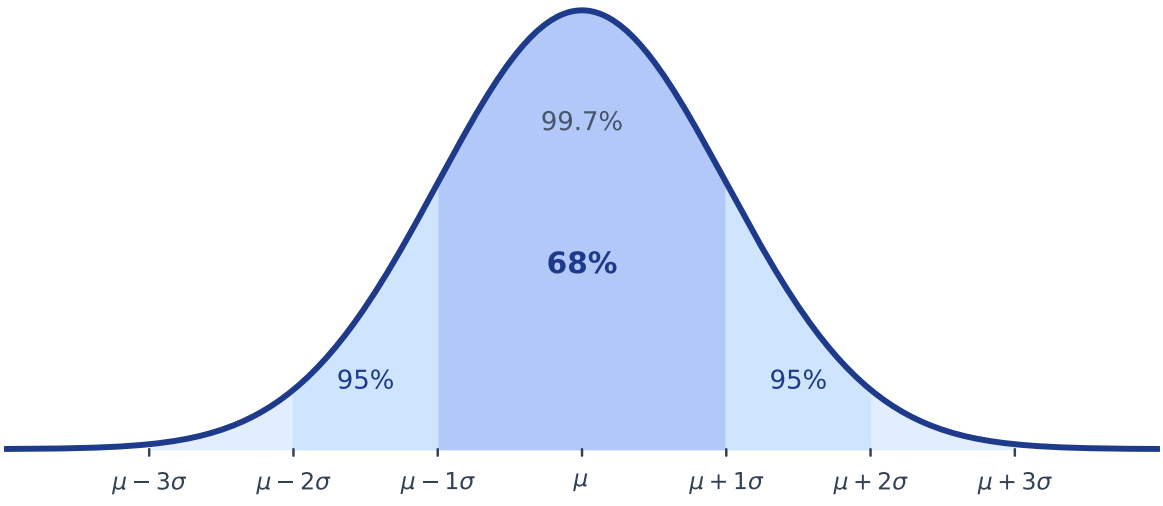
The normal curve: a probability is the area under it, centred on the mean

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

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

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

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



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

Exam tips

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