

Exploring One-Variable Data

AP Statistics ■ Unit 1

AP Statistics



What we will cover

- 1 Variables and data
- 2 Displaying categorical data
- 3 Displaying quantitative data
- 4 Describing a distribution
- 5 Summary statistics
- 6 The normal distribution



1

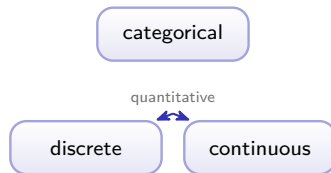
Variables

Variables and data

Statistics 统计学 learns from **data** 数据 through **variability** 变异性. Each **variable** 变量 is one of two types:

- **categorical** 分类变量: a group (eye colour)
- **quantitative** 定量变量: a number (height)

Descriptive 描述统计学 statistics summarize a sample; **inference** 推断 generalizes to a population.



The vocabulary the whole course rests on

statistics 统计学	the science of learning from data 数据 – numbers or labels collected from the real world
a parameter 参数	a numerical summary of a whole population
a statistic 统计量	a numerical summary of a sample – and it varies from sample to sample
a variable 变量	a characteristic that can differ between individuals
categorical	places an individual in a group – shown with bar charts 条形图 and pie charts
quantitative	a numerical measurement – shown with dotplots, histograms and boxplots

Descriptive statistics summarises the sample; **inferential statistics** 推断统计 uses it to say something about the population. The whole course is the second.

2

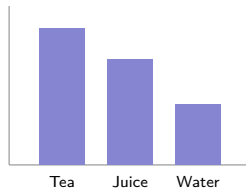
Categorical

Displaying categorical data

Summarize with a **frequency table** 频数表 and **relative frequency** 相对频率 (which sums to 1).

- a **bar chart** 条形图 has **gaps** between bars
- a **two-way table** 双向表 records two variables

Watch the axis: one not starting at 0 exaggerates tiny differences.



Bars separated by
gaps.

3

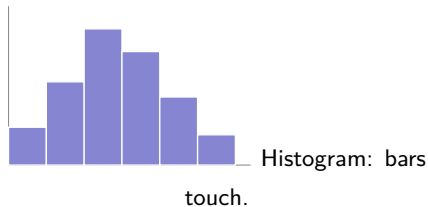
Quantitative

Displaying quantitative data

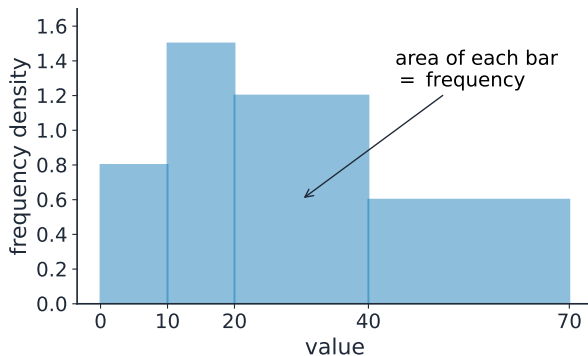
Show *where numbers fall on a scale*:

- a **dotplot** 点图 or **stem-and-leaf** 茎叶图 keeps every value
- a **histogram** 直方图 groups values into **bins** 组距
- an **ogive** 累积相对频率图 reads off **percentiles** 百分位数

Histogram bars **touch** (a number scale) – the visual difference from a bar chart.



Representing a Quantitative Variable with Graphs



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

4

Describing

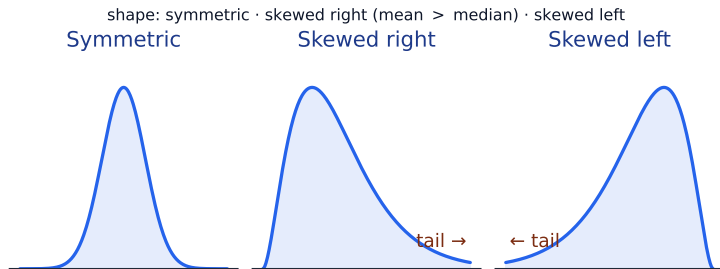
Describing a distribution

Hit four features – **S**hape,
Outliers, **C**enter, **S**pread

– in context.

- **symmetric** 对称 or **skewed** 偏
- skew is named for where the *tail* points

Always mention **outliers** 离群值: they may be errors – or the whole story.



Describing a distribution, in detail

shape	symmetric, or skewed 偏斜 left or right – the tail names the direction
peaks	one main peak is unimodal; two is bimodal 双峰, and usually means two groups are mixed
centre	the mean for a symmetric distribution, the median when it is skewed
spread	the standard deviation with a mean; the interquartile range 四分位距 with a median
outliers	more than $1.5 \times \text{IQR}$ beyond a quartile
a cumulative relative frequency graph 累积相对频率图	reads off percentiles directly

Answer all four, **in context**, every time. Naming the shape without the context is the commonest way to lose a description mark.

5

Summaries

Summary statistics

Center: mean 均值 $\bar{x}$ vs **median** 中位数.

The median is **resistant** 稳健 to outliers.

$$\bar{x} = \frac{1}{n} \sum x_i \quad s_x = \sqrt{\frac{1}{n-1} \sum (x_i - \bar{x})^2}$$

Spread: range 极差, **IQR** 四分位距
= $Q_3 - Q_1$ (resistant), **standard deviation** 标准差.

Symmetric $\Rightarrow$ mean & SD. Skewed $\Rightarrow$ median & IQR.

The z-score

$$z = \frac{x - \bar{x}}{s_x}$$

How many **standard deviations** 标准差 a value sits from the mean – so we can compare different scales.

Describing a distribution, and one more display

Shape words

Skewed left or right, symmetric, **uniform**, or bimodal – name the shape before any number.

Variance

The **variance** is the mean squared deviation, s^2 ; the standard deviation is its square root, and it carries the data's own units.

Variation

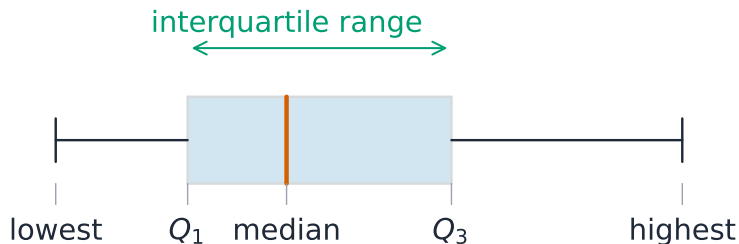
Talking about spread *is* talking about **variation** – the range, the IQR and the standard deviation all measure it differently.

Stem-and-leaf plot

A **stem-and-leaf plot** keeps every original value while showing the shape, so the median is read directly off it.

Shape, centre, spread, and unusual features – in context. Four things, every time a distribution is described.

Graphical Representations of Summary Statistics



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

6

Boxplots

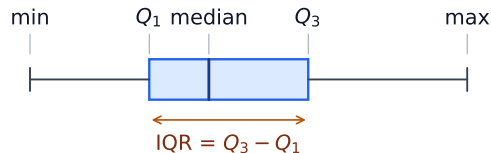
Boxplots and the five-number summary

The **five-number summary** 五数概括:

min, Q_1 , median, Q_3 , max.

- the box length *is* the IQR
- outlier if beyond $Q_{1,3} \mp 1.5 \text{ IQR}$

A **boxplot** 箱线图 hides peaks
– pair it with a histogram when
shape matters.



five-number summary: min, Q_1 , median, Q_3 , max · $\text{IQR} = Q_3 - Q_1$

Worked example – is it an outlier?

Salaries (\$000s):

28, 31, 34, 36, 39, 42, 47, 96. Is 96 an outlier?

$$Q_1 = \frac{31+34}{2} = 32.5, \quad Q_3 = \frac{42+47}{2} = 44.5$$

$$\text{IQR} = 44.5 - 32.5 = 12, \quad 1.5 \times 12 = 18$$

$$\text{upper fence} = 44.5 + 18 = 62.5$$

$96 > 62.5$, so 96 is an outlier.

The fence comes off Q_3 , **not** the median or the mean. And an outlier is not an error – here it is a real salary that makes the **mean** (44.1) a worse summary than the **median** (37.5).

7

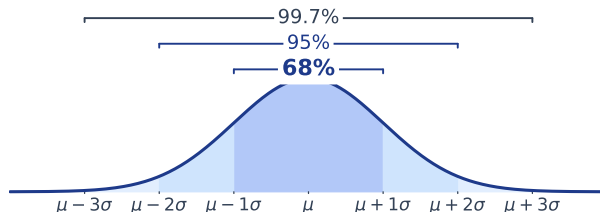
Normal

The normal distribution

The **normal distribution** 正态分布 is set by μ and σ . The **empirical rule** 经验法则 (68–95–99.7):

- 68% within $\mu \pm \sigma$
- 95% within $\mu \pm 2\sigma$; 99.7% within 3σ

The **proportion** 比例 below a value = the **area** 面积 to its left, found from z .



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

Worked example – the empirical rule and z

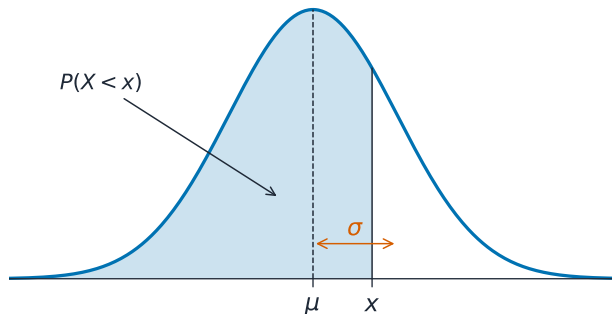
Heights are normal with $\mu = 170$ cm, $\sigma = 8$ cm. (a) What percent are between 154 and 186? (b) Where does 182 cm sit?

(a) $154 = \mu - 2\sigma$ and $186 = \mu + 2\sigma$, so by the empirical rule **95%**.

(b) $z = \frac{182 - 170}{8} = 1.5$ – one and a half SDs **above** the mean.

The empirical rule only works at **whole** multiples of σ . 182 is not one, so it needs z – and z answers “**how many SDs from the mean**”, which is what every later unit builds on.

Probability as area under the normal curve



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

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.

8

Recap

Unit 1 – key takeaways

1

Variables

categorical vs quantitative decides every graph

2

Describe

Shape, Outliers, Center, Spread – in context

3

Summaries

median/IQR resistant; mean/SD for symmetric; z-score

4

Normal

68–95–99.7 rule; area to the left = proportion

Check yourself

- 1 Is a zip code categorical or quantitative?
- 2 Which center is resistant to outliers?
- 3 A right-skewed tail points toward high or low values?
- 4 What percent of a normal lies within 2σ ?



Answers 1. categorical (a label) 2. median 3. high values 4. about 95%