

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



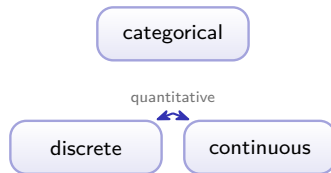


Variables and data

Statistics learns from **data** through **variability**. Each **variable** is one of two types:

- **categorical** : a group (eye colour)
- **quantitative** : a number (height)

Descriptive statistics summarize a sample; **inference** generalizes to a population.



The vocabulary the whole course rests on

statistics

the science of learning from **data** – numbers or labels collected from the real world

a parameter

a numerical summary of a whole **population**

a statistic

a numerical summary of a **sample** – and it varies from sample to sample

a variable

a characteristic that can differ between individuals

categorical

places an individual in a group – shown with **bar charts** and pie charts

quantitative

a numerical measurement – shown with dotplots, histograms and boxplots

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

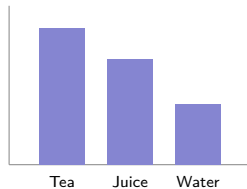


Displaying categorical data

Summarize with a **frequency table** and **relative frequency** (which sums to 1).

- a **bar chart** has **gaps** between bars
- a **two-way table** records two variables

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



Bars separated by
gaps.

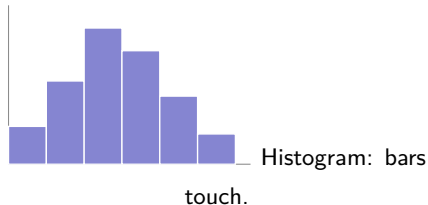


Displaying quantitative data

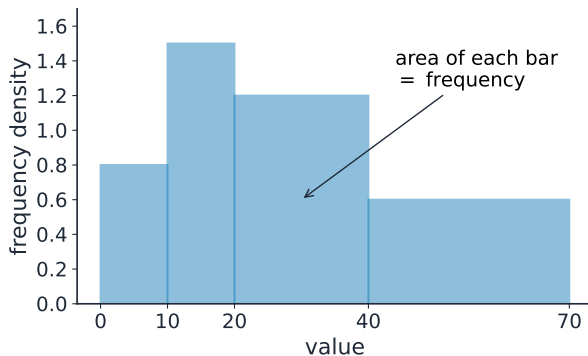
Show *where numbers fall on a scale*:

- a **dotplot** or **stem-and-leaf** keeps every value
- a **histogram** groups values into **bins**
- an **ogive** reads off **percentiles**

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

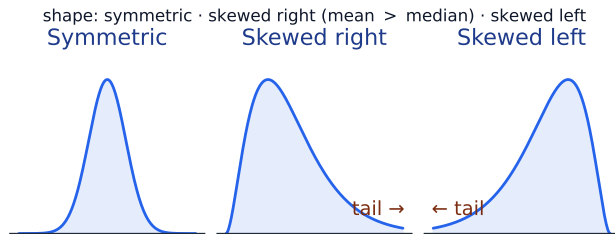


Describing a distribution

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

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

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



Describing a distribution, in detail

shape	symmetric, or skewed left or right – the tail names the direction
peaks	one main peak is unimodal; two is bimodal , 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 with a median
outliers	more than $1.5 \times \text{IQR}$ beyond a quartile
a cumulative relative frequency graph	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.



Summary statistics

Center: mean $\bar{x}$ vs **median** . The median is **resistant** 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 , **IQR** = $Q_3 - Q_1$
(resistant), **standard deviation** .

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

The z-score

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

How many **standard deviations** 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

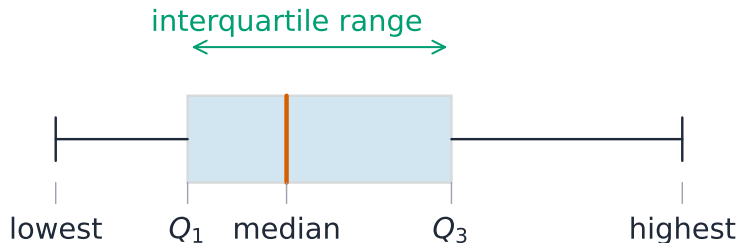
Talking about spread *is* talking about **variation** – the range, the IQR and the standard deviation all measure it differently.

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



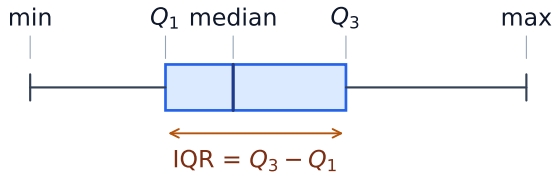
Boxplots and the five-number summary

The **five-number summary** : 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** hides peaks – pair it with a histogram when shape matters.

Boxplots and the five-number summary, in detail



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

The **five-number summary** : min, Q_1 , median, Q_3 , max.

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).

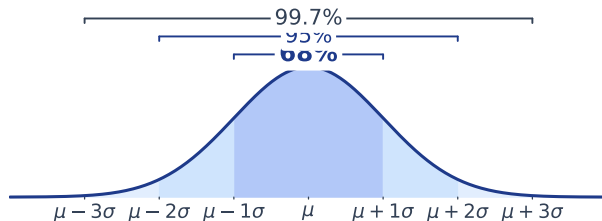


The normal distribution

The **normal distribution** is set by μ and σ . The **empirical rule** (68–95–99.7):

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

The **proportion** below a value = the **area** 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

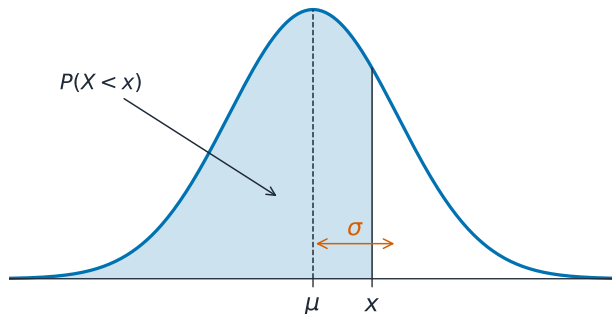
Heights are normal with $\mu = 170$ cm, $\sigma = 8$ cm. (a) What percent are between 154 and 186? (b) Where does 182 cm sit?

(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.

The Normal Distribution, continued



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.



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%