# Lesson - Sampling and scatter graphs

Year: Year 11

Statistics is about drawing sensible conclusions from data. This topic covers fair sampling, scatter graphs and lines of best fit. Capture-recapture questions are usually on the Higher paper (marked Higher).

**Starter activity**

Suppose you want to find out the favourite lesson in your school. You could ask your friends, ask the first ten students you meet in the corridor, or choose names at random from the register. Which method would you trust the most, and why? What could go wrong with each of them? Write down one way the result could be misleading for each method.

## Populations and samples

**Definition: population and sample**

The **population** is the whole group you want to know about. A **sample** is a smaller group chosen from it, because it is quicker and cheaper than asking everyone. A good sample is **representative** (it looks like the population) and **unbiased**. A larger sample is generally more reliable.

**Definition: bias**

**Bias** occurs when some members are more likely to be chosen than others. Asking only the pupils in the canteen queue about school lunches is biased, because it ignores pupils who bring a packed lunch.

## Types of sampling

**Key fact: three types of sampling**

- **Simple random sampling**: every member has an equal chance, for example numbering the 500 pupils and picking 30 numbers with a random number generator.
- **Systematic sampling**: choose every $k$th item from a list after a random start. For 400 names and a sample of 40, take every 10th name.
- **Stratified sampling**: split the population into groups (strata) and take from each group in proportion to its size.

**Rule: stratified sample**

$$\text{number from a group} = \frac{\text{size of group}}{\text{population size}} \times \text{sample size}$$

### Worked example 1: stratified sampling

**A school of 800 pupils has 200 in Year 7, 180 in Year 8, 220 in Year 9 and 200 in Year 10. A stratified sample of 80 is needed.**

Year 7: $\frac{200}{800} \times 80 = 20$. Year 8: $\frac{180}{800} \times 80 = 18$. Year 9: $\frac{220}{800} \times 80 = 22$. Year 10: $\frac{200}{800} \times 80 = 20$. Check that $20 + 18 + 22 + 20 = 80$.

### Worked example 2: estimating from a sample

**In a sample of 50 from 2000 pupils, 12 walk to school. Estimate how many in the school walk.**

If a sample is representative, the proportion in the sample estimates the proportion in the population: $\frac{12}{50} \times 2000 = 480$.

## (Higher) Capture-recapture

### Rule (Higher): capture-recapture

**Capture** some animals, mark them and release them. Later capture a second sample and count how many are marked. The proportion marked in the second sample is assumed to equal the proportion marked in the whole population:

$$N = \frac{\text{first} \times \text{second}}{\text{marked in second}}$$

### Worked example 3 (Higher): fish in a lake

**40 fish are tagged and released. Later 50 are caught and 8 have tags. Estimate the population.**

$N = \frac{40 \times 50}{8} = 250$. Assumptions: the population stays the same size (no births, deaths or migration), the tags do not fall off, and the tagged animals have mixed fully with the others.

## Scatter graphs

**Definition: scatter graph and correlation**

A scatter graph plots pairs of values to see if there is a link, called **correlation**.

- **Positive**: as one variable increases, the other increases (hours revised and test score).
- **Negative**: as one increases, the other decreases (age of a car and its value).
- **No correlation**: no pattern.

The correlation can be strong (points close to a line) or weak (points scattered).

**Watch out: correlation is not causation**

**Correlation does not mean causation**: ice cream sales and sunburn are both high in hot weather, but ice cream does not cause sunburn.

## Line of best fit

**Rule: line of best fit**

A line of best fit is a straight line through the middle of the points, with about the same number on each side. It should pass through the mean point $(\bar{x}, \bar{y})$ and need not go through the origin. To find its equation, choose two points *on the line* far apart, work out the gradient and then the intercept.

**Method: drawing and using a line of best fit**

1. Plot the points and plot the mean point if you can.
2. Draw a straight line through the middle with about the same number of points on each side.
3. To estimate, read from the line; to find the equation, use two points on the line far apart.

**Worked example 4: equation of the line**

**The line passes through $(2, 10)$ and $(8, 34)$. Find its equation and estimate $y$ when $x = 5$.**

Gradient $= \dfrac{34 - 10}{8 - 2} = 4$. Using $y = 4x + c$ and the point $(2, 10)$: $10 = 8 + c$, so $c = 2$.

The equation is $y = 4x + 2$. In context: each extra hour of revision adds about 4 marks. For $x = 5$ the estimate is $y = 22$.

**Key fact: interpolation, extrapolation and outliers**

**Interpolation** means estimating inside the range of the data, and it is usually

reliable. **Extrapolation** means estimating outside the range, which can be unreliable because the pattern may not continue. If the data run from $x = 1$ to $x = 9$, predicting at $x = 15$ is extrapolation: $y = 4x + 2$ gives 82 at $x = 20$, which may be unrealistic (a test mark cannot exceed the maximum). An **outlier** is a point that does not fit the pattern: check whether it is an error and ignore it in the line unless there is a reason to keep it.

## Questionnaires

**Note: good questions**

Good questions are clear, unbiased and give response boxes that do not overlap. "Do you agree that maths is brilliant?" is a leading question. A better one is "How would you rate maths on a scale of 1 to 5?"

**Watch out: common mistakes**

- Using the sample size instead of the population size in the stratified formula.
- Drawing a line of best fit that joins the first and last points.
- Writing "it is a good line" without a reason.
- Saying that correlation proves one thing causes the other.

**Worked example 5: Why a larger sample helps**

**A random sample of 10 pupils gives a mean height of 158 cm and a random sample of 200 pupils gives a mean of 162 cm. Which is better?**

The sample of 200 is better. A small sample can by chance contain mostly short or tall pupils, while a large sample averages out these differences. If the two means differ a lot, trust the larger sample. A larger sample does not remove bias, though: a biased sample of 10 000 is still biased.

**Worked example 6: Stratified sampling with uneven groups**

**A club has 45 juniors, 30 seniors and 25 veterans. Choose a stratified sample of 20 members.**

The club has 100 members, so the fraction is $\frac{20}{100} = \frac{1}{5}$. Juniors $45 \div 5 = 9$, seniors $30 \div 5 = 6$, veterans $25 \div 5 = 5$. Total $= 20$. If the answers do not come out as whole numbers, round and check that the total is still correct.

**Worked example 7: Using the capture-recapture formula**

**A zoologist marks 24 hares. In a second sample of 30 she finds 4 marked. Estimate the population.**

The proportion marked in the second sample, $\frac{4}{30}$, is taken to equal the proportion marked in the population, $\frac{24}{N}$. Hence $N = \frac{24 \times 30}{4} = 180$. The estimate depends on the second sample being random and on the population being the same at both times.

### Worked example 8: Interpolation and extrapolation

**A scatter graph of revision time $x$ (1 to 8 hours) and mark $y$ has line $y = 4x + 10$. Predict the mark for 6 hours and for 15 hours, and comment.**

For 6 hours, $y = 34$: this is interpolation, within the range of the data, so reasonably reliable. For 15 hours, $y = 70$ is extrapolation, outside the range. The marks may not keep increasing at the same rate (and might exceed the maximum possible mark), so this prediction is unreliable.

### Worked example 9: Describing correlation fully

**Describe the relationship shown when points go from top left to bottom right and lie close to a straight line.**

This is a **strong negative correlation**: as one variable increases the other decreases. The word strong refers to how close the points are to a line, while negative refers to the downward direction. If the points are widely scattered it is weak, and if there is no pattern then there is no correlation.

### Worked example 10: A good line of best fit

**List the features of a good line of best fit drawn by eye.**

It passes through the mean point of the data, roughly half of the points lie above it and half below it, and it follows the general direction of the points. It does not need to pass through the origin or through any particular point. Ignore obvious outliers when drawing it.

### Worked example 11: Questionnaire design

**Improve the question "Don\'t you agree that school lunches are terrible?"**

The question is leading because it suggests the answer. A better question is "How would you rate school lunches?" with response boxes: very good, good, average, poor, very poor. Response boxes should not overlap and should cover all possible answers. Keep questions short, clear and about one thing at a time.

### Note: types of sampling

A **random sample** gives every member an equal chance of being chosen, for example using random numbers. A **systematic sample** picks every $k$th item from a list. A **stratified sample** divides the population into groups (strata) and takes a

random sample from each group in proportion to the size of the group. Stratified sampling is best when the groups differ from each other. Larger samples give more reliable results, but only if the sampling method is fair.

## Method: drawing and using a line of best fit

1. Plot the points carefully and describe the correlation.
2. Draw a straight line that follows the trend, with about the same number of points on each side.
3. To predict, read from the line at the given value; to find the equation, choose two points on the line (not data points) that are far apart and use $y = mx + c$.
4. Comment on whether the prediction is interpolation (reliable) or extrapolation (unreliable).

## Watch out: errors in statistics

- Correlation does **not** mean causation. Ice cream sales and sunburn are both caused by sunshine.
- A bigger sample reduces random error but cannot fix a biased method.
- In capture-recapture the formula uses the **first** marked number, the size of the **second** sample and the number marked in it.
- Do not extend a line of best fit far beyond the data.

## Did you know?

In 1936 a magazine called the *Literary Digest* asked over two million people who they would vote for in the American presidential election and predicted that Alf Landon would win. It chose its sample from telephone books and car-ownership lists, which in those years left out many poorer voters. A much smaller random sample of only 50 000 people correctly predicted that Franklin Roosevelt would win. The story is now a classic warning that a large sample cannot make up for a biased method.

## Otto's tip

When you comment on a graph, mention the type of correlation, give the variables in words and say what it means in context, for example "the more hours revised, the higher the score".