Lesson - Statistical sampling and data

Year: Year 12

Statistics asks how to collect data fairly and how to describe it. Year 12 covers sampling, measures of centre and spread, histograms, box plots, outliers and bivariate data. Always answer in context, using the words of the question.

Starter activity

You are asked to find out how long students in your school spend on homework each night. Write down three different ways you could choose which students to ask. Which of your methods might give biased results? Suppose the school has 1 000 students and you can only ask 50. How could you make sure that the 50 are a fair mix of all year groups? Explain why asking only your friends may give a poor picture of the whole school.

Population and samples

Definition: population, sample and census

The **population** is everything you want to know about; a **sample** is a part of it. A **census** measures the whole population: accurate but slow, costly and sometimes destructive. A **sampling frame** is the list of all members.

Key fact: sampling methods

- **Simple random:** every member has an equal chance (random numbers). Unbiased, but needs a full list.
- **Systematic:** every kth member after a random start, $k = \dfrac{\text{population}}{\text{sample size}}$. Easy, but may clash with a pattern.
- **Stratified:** split into strata and sample each in proportion: $\dfrac{\text{stratum}}{\text{population}} \times \text{sample size}$.
- **Quota:** interviewers fill set quotas. Quick, but not random.
- **Opportunity:** take whoever is available. Easy but biased.

Worked example 1

A company has 120 men and 80 women. Describe a stratified sample of 20.

Men: $\dfrac{120}{200} \times 20 = 12$. Women: $\dfrac{80}{200} \times 20 = 8$.

Watch out

Do not call a sample "random" when it is quota or opportunity.

Averages and spread

Rule: mean, variance and IQR

Mean $\bar{x} = \dfrac{\sum x}{n}$; for a frequency table $\bar{x} = \dfrac{\sum fx}{\sum f}$. Variance $\sigma^2 = \dfrac{\sum x^2}{n} - \bar{x}^2$, standard deviation $\sigma = \sqrt{\text{variance}}$. Interquartile range $IQR = Q_3 - Q_1$.

Worked example 2

$n = 10$, $\sum x = 150$, $\sum x^2 = 2500$. **Find the mean and standard deviation.**

Mean 15, variance $\dfrac{2500}{10} - 15^2 = 25$, standard deviation 5.

Note: coding

If $y = \dfrac{x - a}{b}$ then $\bar{x} = b\bar{y} + a$ and $\sigma_x = b\sigma_y$. Adding a constant changes the mean but not the spread.

Histograms

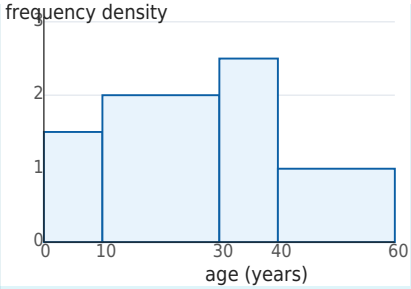
Definition: frequency density

For unequal class widths the *area* of each bar represents frequency. The height is the **frequency density** $= \dfrac{\text{frequency}}{\text{class width}}$.

Worked example 3

Ages 0–10 (15), 10–30 (40), 30–40 (25), 40–60 (20). Find the frequency densities.

$\dfrac{15}{10} = 1.5$, $\dfrac{40}{20} = 2$, $\dfrac{25}{10} = 2.5$, $\dfrac{20}{20} = 1$.



Method: median and mean from grouped data

1. Find the cumulative frequencies and locate the class containing the $\frac{n}{2}$th value.
2. Interpolate inside that class.
3. For the mean use class mid-points: $\bar{x} \approx \dfrac{\sum fm}{\sum f}$.

Worked example 4

Using the data above ($n = 100$), estimate the median.

The 50th value lies in 10–30 (cumulative 15 up to 10 and 55 up to 30): median $= 10 + \dfrac{35}{40} \times 20 = 27.5$.

Quartiles, box plots and outliers

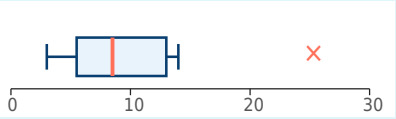
Definition: outlier

A value is an **outlier** if it is more than $1.5 \times IQR$ beyond a quartile: below $Q_1 - 1.5\,IQR$ or above $Q_3 + 1.5\,IQR$. A box plot shows the minimum, Q_1, median, Q_3 and maximum, with outliers marked by a cross.

Worked example 5

Data: 3, 5, 6, 8, 9, 12, 14, 25. Is 25 an outlier?

Median 8.5, $Q_1 = 5.5$, $Q_3 = 13$, $IQR = 7.5$. Upper fence $13 + 1.5 \times 7.5 = 24.25$. Since $25 > 24.25$ the value 25 is an outlier.



Note: skewness and cleaning

If the median is closer to Q_1 than Q_3 the distribution has positive skew. Investigate outliers before removing them: never remove a value only because it is inconvenient. Cleaning data means correcting or removing errors and missing values.

Scatter diagrams and regression

Definition: regression line

A scatter diagram shows two variables, with positive, negative or no correlation. A **regression line** $y = a + bx$ predicts y from x; the gradient b is the change in y for one unit of x.

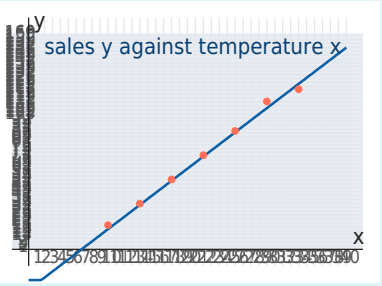
Rule: using a regression line

Prediction **within** the range of the data is interpolation (reliable). Prediction outside the range is extrapolation (unreliable). **Correlation does not imply causation.**

Worked example 6

Sales $y = -30 + 4.5x$, with temperature x in °C. Interpret the gradient and predict sales at 22°C and 50°C.

Each extra degree gives 4.5 more sales. At 22°C: $-30 + 99 = 69$ sales. At 50°C the prediction is an extrapolation and not trustworthy.



Watch out

Use frequency density, not frequency, as the bar height. Remember to subtract the squared mean in the variance. Find Q_1 and Q_3 before using the outlier rule. Never claim causation from correlation.

Worked example 7: Choosing between mean and median

House prices (£000s) in a street: 210, 225, 230, 240, 245, 250, 1 200. Which average should an estate agent quote and why?

Mean $= \dfrac{2600}{7} = 371$, median $= 240$. The £1.2 million house pulls the mean far above the typical price. The median 240 is much closer to what most people pay. Using the 1.5 IQR rule: $Q_1 = 225$, $Q_3 = 250$, IQR 25, upper limit $250 + 37.5 = 287.5$, so 1 200 is an outlier. A sensible agent quotes the median (and possibly the IQR) and mentions the outlier.

Worked example 8: Standard deviation by hand

Find the mean and standard deviation of 2, 4, 4, 5, 10.

Mean $= \frac{25}{5} = 5$. $\sum x^2 = 4 + 16 + 16 + 25 + 100 = 161$. Variance $= \frac{161}{5} - 25 = 7.2$. Standard deviation $= \sqrt{7.2} = 2.68$. Check using deviations: $(2-5)^2 + (4-5)^2 + (4-5)^2 + 0 + (10-5)^2 = 9 + 1 + 1 + 0 + 25 = 36$, and $\frac{36}{5} = 7.2$ ✓. The two formulas $\sigma^2 = \frac{\sum (x - \bar{x})^2}{n} = \frac{\sum x^2}{n} - \bar{x}^2$ are equivalent.

Worked example 9: Frequency density for unequal class widths

Classes 0-5 (frequency 10), 5-15 (frequency 40), 15-40 (frequency 25). Find the frequency densities and the frequency in 10-20.

Frequency densities: $\frac{10}{5} = 2$, $\frac{40}{10} = 4$, $\frac{25}{25} = 1$. The class 10-20 consists of half of the 5-15 class (10-15, which is 5 of the 10 units) and one fifth of the 15-40 class (15-20, 5 of the 25 units): $5 \times 4 + 5 \times 1 = 25$. So about 25 observations are in 10-20.

Worked example 10: Describing correlation and cause

A study finds a strong positive correlation between ice-cream sales and the number of people swimming. Does one cause the other?

No. Both depend on a third variable, the temperature, which is called a lurking (or confounding) variable. Strong correlation shows a linear association but does not prove cause and effect. To investigate a causal link one would need a controlled experiment, where only the factor of interest is changed.

Worked example 11: Using regression carefully

The line $y = 12 + 3.5x$ was fitted for x between 10 and 40. Estimate y at $x = 25$ and at $x = 60$ and say which is more reliable.

At $x = 25$: $y = 12 + 87.5 = 99.5$ (interpolation, inside the range, reliable). At $x = 60$: $y = 12 + 210 = 222$ (extrapolation, outside the range). The second estimate is less reliable because the relationship may change outside the observed range. The closer the correlation coefficient is to ± 1, the more reliable the interpolated estimate.

Worked example 12: Interquartile range from grouped data

Use the cumulative frequencies below to estimate the median and quartiles: classes 0-10 (8), 10-20 (14), 20-30 (18), 30-40 (10); total 50.

Cumulative frequencies: 8, 22, 40, 50. The median is the 25th value, in the class

20-30: $20 + \dfrac{25 - 22}{18} \times 10 = 21.7$. The lower quartile is the 12.5th value, in the class 10-20: $10 + \dfrac{12.5 - 8}{14} \times 10 = 13.2$. The upper quartile is the 37.5th value, in 20-30: $20 + \dfrac{37.5 - 22}{18} \times 10 = 28.6$. The IQR is about $28.6 - 13.2 = 15.4$. These are estimates because we assume the values are spread evenly inside each class.

Worked example 13: Conditions for a good sample

A radio station asks listeners to text in their opinion on a new road. Give three reasons why the result might not represent the town.

First, only listeners can respond, and they may differ from the general population in age or interests. Second, people with strong views are more likely to bother texting, which produces self-selection bias. Third, texting costs money or effort and excludes people without phones. A better design is a random sample taken from the electoral register or a stratified sample using age groups and areas of the town, with follow-up for non-respondents.

Method: comparing two data sets

1. Compare a measure of location: the median (or the mean if there are no outliers).
2. Compare a measure of spread: the IQR (or the standard deviation).
3. Check for outliers and comment on skewness by comparing mean and median or the quartiles.
4. Write a sentence in context, for example "Class B has a similar median score but a smaller IQR, so its results are more consistent".

Watch out: common statistics mistakes

- Calculating the mean of grouped data from the class boundaries instead of the mid-points.
- Plotting frequency instead of frequency density on a histogram with unequal class widths.
- Using the sample standard deviation formula (dividing by $n - 1$) where the course uses the population formula, or the reverse; follow the formulae given.
- Extrapolating a regression line far outside the data range.
- Concluding cause from correlation.

Note: the large data set

The A level course uses a large data set (weather data for several UK and overseas stations). You are expected to know the context: which variables are recorded, how daily values are summarised, what units are used (millimetres of rainfall, degrees

Celsius, knots for wind, hours of sunshine) and which entries are missing or recorded as "tr" for a trace. In an exam you might be given an extract and asked to calculate averages, comment on outliers, draw a diagram or interpret a regression line. Practise using real data on a spreadsheet to build familiarity with its quirks.

Did you know?

In 1936 a magazine, the *Literary Digest*, predicted that Alf Landon would win the US presidential election by a landslide after receiving 2.4 million replies. Franklin Roosevelt won easily. The magazine had sampled from telephone directories and car registrations, which in the Depression meant mainly wealthy voters. A much smaller but properly random sample by George Gallup predicted the correct result. The lesson is that sample quality matters far more than sample size.

Otto's tip

"Area is frequency." In a histogram with unequal widths, always work out frequency density first.