SUMMARY · AT EXAM LEVEL

Module 2

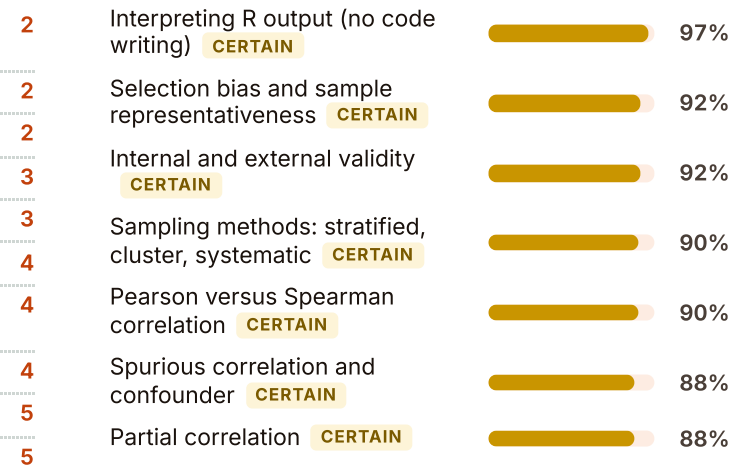
Descriptive statistics, sampling and correlation

Everything you must be able to define, explain, apply and assess for the open questions. With the models as figures, the comparisons as tables and the pitfalls marked.

WHAT IS INSIDE

01	Descriptive statistics and pitfalls of shallow statistics	2
02	Visualisation	2
03	Population, sample and validity	2
04	Sampling methods	3
05	Sampling errors and biases	3
06	Inference: point and interval estimators	4
07	Central limit theorem, standard error and confidence interval	4
08	Covariance	4
09	Correlation	5
10	Correlation is not causation	5
11	Statistical dependency	5
12	Reading R output	6
13	Worked example: from standard error to interval	6
14	Common mistakes	6
→	How to write the answer	6
→	Checklist: can I explain this?	7
→	Test yourself	7

EXAM LIKELIHOOD



Estimate per topic, based on the sample exam and the emphasis in the material.

KEY CONCEPTS

descriptive statistics · population · validity · stratified sampling · selection bias · central limit theorem · confidence interval · covariance · Pearson · Spearman · partial correlation · statistical independence

ON THE EXAM

Block 2

10 points · 180 min in total

MC	4 multiple-choice questions	2.5 points per question
7	Block 7 (25 points) consists of three open questions about R output.	200 / 250 / 150 words

Exam: Mon, 7 Dec 2026 · 180 min · Technical terms stay in English

01 Descriptive statistics and pitfalls of shallow statistics

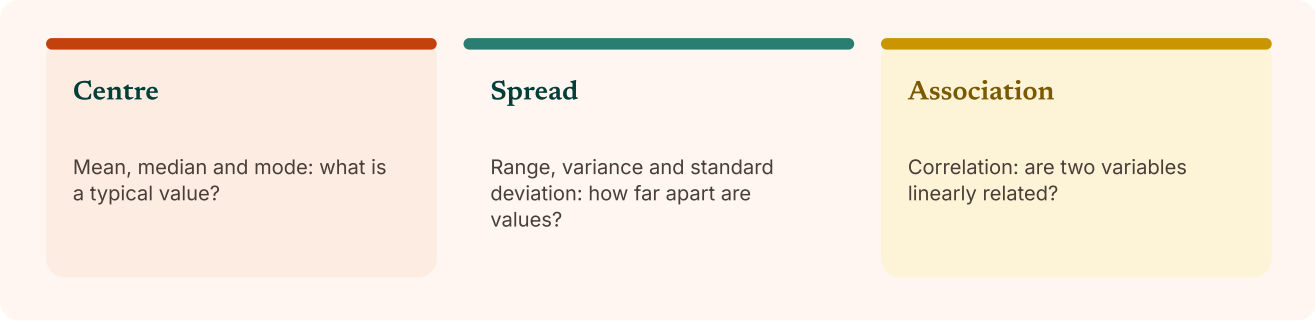


FIGURE 2.1 **Three kinds of measures.** Descriptive statistics summarise a dataset.

Descriptive statistics are summary measures for quantitative data: mean, median, standard deviation, interquartile range. A mean larger than the median points to a tail to the right. A Z-score is (x – mean) / sd and counts the number of standard deviations from the mean; a value with $|Z|$ larger than 2 stands out. The outlier fences of the box plot are Q3 + 1.5 x IQR and Q1 – 1.5 x IQR.

A summary is never neutral. The retail newspaper example made three choices: one observation instead of the trend, the previous month as reference instead of a year earlier, and an axis that does not start at zero. For every chart ask which choice is behind it.

Qualitative data are described with a frequency table, bar chart or pie chart. A mean of category codes has no interpretation.

02 Visualisation

- Histogram: the distribution of one quantitative variable (bins).
- Scatter plot: relationship between two quantitative variables, possibly a third as colour.
- Box plot: a quantitative variable per category (median, quartiles, outliers).

03 Population, sample and validity

You draw a sample from a population to say something about it. The three levels are target population (what you want to say something about), source population (what you really draw from) and study sample.

- **Internal validity**: how trustworthy is the answer for the source population? Depends on statistics, sampling design and biases and confounders.
- **External validity**: to what extent do you generalise to other populations or times?

Pitfall in a thesis: a source population that is too specific costs external validity. **Representativity** means that mean, spread and shape of the sample resemble those of the population.

04 Sampling methods

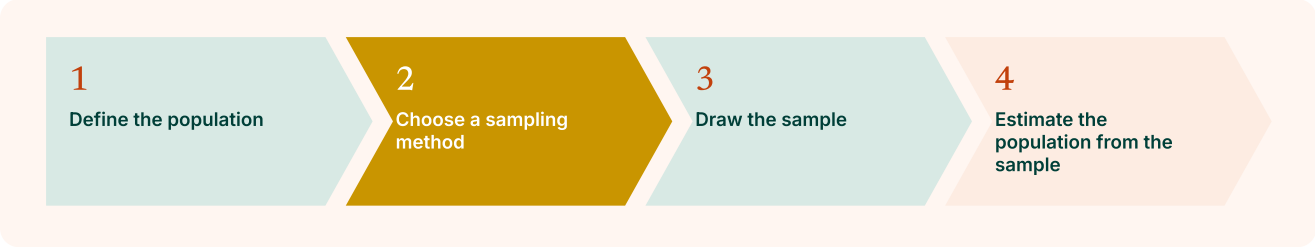


FIGURE 2.2 **From population to sample.** A sample is only useful if it is representative.

- **Simple random sampling**: every element has the same probability. Easy in theory, hard in practice (the Literary Digest example of 1936: large sample, wrong selection).
- **Stratified sampling**: divide into strata (age, gender, department) and draw from each stratum. This way the sample is balanced on those characteristics.
- **Systematic sampling**: ordered list, fixed step size. With N = 2000 and n = 50 you take every 40th.
- **Cluster sampling**: randomly choose clusters (schools, cities, companies) and collect data inside them.

The advantage of the alternatives: less effort in data collection, often a better estimate.

05 Sampling errors and biases



FIGURE 2.3 **From population to sample.** A sample is only useful if it is representative.

Sampling error is the difference between a sample statistic and the population value (x-bar versus mu). Non-systematic errors come from chance and have no direction. Systematic errors are biases:

- **Selection bias**: the sampling design over- or under-represents groups.
- **Non-response bias**: some groups disproportionately often do not take part.
- **Response bias**: answers are systematically not in line with reality. Example: the same two product pictures in reversed order gave 56% for the first picture in sample 1 and 58% for the first (other) picture in sample 2.

A larger n reduces the chance error, but does not remove bias.

06 Inference: point and interval estimators

Inference is a conclusion about the population based on a sample. Two sources of uncertainty: sampling variation and unknown population characteristics. A **point estimator** gives one number; an **interval estimator** gives an interval that contains the parameter with a certain probability.

07 Central limit theorem, standard error and confidence interval

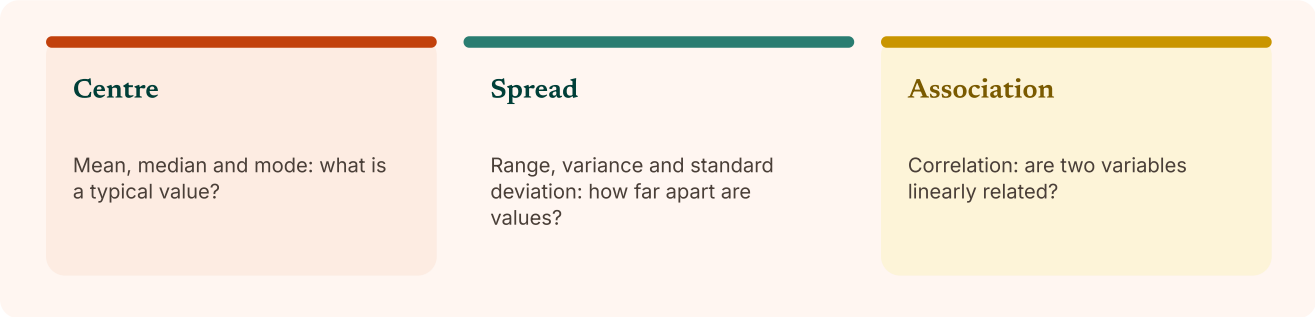


FIGURE 2.4 **Three kinds of measures.** Descriptive statistics summarise a dataset.

For a random variable with expected value mu and finite variance sigma squared, the mean of a simple random sample of size n is approximately normally distributed with mean mu and standard deviation sigma / sqrt(n): the **standard error** (SE). In practice this works from n = 20 to 30.

- 95% confidence interval: x-bar plus or minus 1.96 x SE; 90%: 1.64; 99%: 2.58.
- Unknown sigma: use s and t* from the t-distribution: CI = x-bar plus or minus t* x s / sqrt(n).
- A higher confidence level means a wider interval; four times as many observations halves the SE.
- Interpretation: the 95% belongs to the method, not to one interval.

In R: qnorm(0.975) = 1.96; pnorm gives the left tail, the right tail is 1 – pnorm(); qt(0.975, df) gives t*.

08 Covariance

Cov(X, Y) = sum of (x_i – x-bar)(y_i – y-bar), divided by n – 1. It is the average rectangle area around the centre of gravity. Positive means x and y rise together, negative that one rises when the other falls. Drawback: the value depends on the units and can be anything, so it is hard to interpret.

09 Correlation

Pearson: r = Cov(X, Y) / (sd(X) x sd(Y)). Always between -1 and +1, unit-free. Example: Cov = 54, sd's 12 and 9 gives r = 54 / 108 = 0.5. What is strong or weak depends on the context.

Three pitfalls:

1. An **outlier** can pull r strongly up or down. A conclusion based on r alone is wrong; always look at the scatter plot.
2. An r around 0 does not mean there is no relationship; Pearson only measures linear association (y = x squared gives r = 0).
3. For a non-linear but rising relationship Pearson gives a value below 1.

Spearman is the correlation of the ranks and fits monotone non-linear relationships and outliers.

10 Correlation is not causation

A correlation between A and B can arise from: A causes B, B causes A (**reverse causality**), chance (mozzarella and doctorates), feed-forward (police and crime; cycling and BMI), or a third variable C (**confounder**: ice cream sales and drownings via temperature).

A **spurious correlation** is a statistical relationship without a causal one, due to chance or a confounder. The **partial correlation** measures the relationship between X and Y after controlling for Z:

r_XY.Z = (r_XY – r_XZ x r_YZ) / sqrt((1 – r_XZ squared)(1 – r_YZ squared)).

A partial correlation much smaller than r_XY points to a confounder or to a mediator; the design (the causal order in the scheme) decides which of the two it is.

11 Statistical dependency

Causal relationship: a change in X leads to a change in Y while all other influences stay constant (ceteris paribus). Statistical relationship: the joint distribution shows a systematic pattern. Causality is not the same as statistical dependence; dependence can be determined objectively from data, causality is harder.

Two variables are statistically independent if the marginal distribution equals the conditional distribution: P(buy) = P(buy | blue website). Information about X then has no predictive value for Y.

12 Reading R output

The exam does not ask for R code, but it does ask you to interpret output. In summary() you compare mean and median; in cor.test() you read r, t, df, the p-value and the 95% confidence interval. If the interval contains 0, the correlation is not significant at 5%. You read a correlation matrix per pair; a single outlier can make a significance disappear or appear.

13 Worked example: from standard error to interval

Suppose n = 100, mean 50 and s = 20 (simulated data). Then SE = 20 / sqrt(100) = 2. The 95% confidence interval is 50 +/- 1.96 × 2, so [46.08; 53.92]. To halve the SE to 1 you need n = 400: four times as many, because the SE falls with the square root of n. A higher confidence level (99%: 2.58) makes the interval wider, a larger n makes it narrower.

WATCH OUT

14 Common mistakes

- ✖ Saying that a strong correlation proves causality.
- ✖ Thinking that the 95% is a probability for this one interval.
- ✖ Seeing a larger sample as a cure for selection bias.
- ✖ Interpreting a correlation without the scatter plot.

MODEL ANSWER

→ How to write the answer

Four building blocks, in this order. Each one earns points; leave none out.

1. Term
Name the model or concept with the English term.

2. Definition
Give the definition or features, precisely.

3. Case fact
Link it to a concrete fact from the case.

4. Conclusion
End with a judgement, advice or consequence.

QUESTION 1 · WORD LIMIT 75 WORDS (WORD LIMIT ASSUMED)

A dataset contains a few extremely high salaries. Which measure of centre fits best, and why?

TERM The median fits best. DEFINITION The median is the middle value of the ordered data and is not affected by outliers; the mean is. CASE FACT A few extremely high salaries pull the mean up, so it is no longer typical for most employees. CONCLUSION For a skewed distribution report the median, possibly next to the mean and the standard deviation.

57 / 75 words

KEY CONCEPTS

✓ Checklist: can I explain this?

Tick what you can explain from memory in one sentence, using the English term.

- ☐ descriptive statistics
- ☐ population
- ☐ validity
- ☐ stratified sampling
- ☐ selection bias
- ☐ central limit theorem
- ☐ confidence interval
- ☐ covariance
- ☐ Pearson
- ☐ Spearman
- ☐ partial correlation
- ☐ statistical independence

FLASHCARDS

? Test yourself

Cover the right-hand column and say the answer out loud.

Question	Answer
1 Why can "sales fell in April" be a misleading headline?	One month, a comparison with the previous month and a chosen axis shape the story.
2 What does a Z-score of 2.5 mean?	The value lies 2.5 standard deviations above the mean.
3 What risk do you run with a source population that is too specific?	You lose external validity: the outcome is hard to generalise.
4 Why are the alternatives to random sampling interesting?	They reduce the effort of data collection and often improve the estimate.
5 What went wrong in the Literary Digest poll of 1936?	Car owners, phone users and subscribers were wealthier: selection bias, despite 2.3 million answers.
6 What does the central limit theorem say?	The sample mean is approximately normally distributed with mean mu and sd sigma / sqrt(n).

Question	Answer
7 What is the difference between standard deviation and standard error?	Standard deviation: spread of the data; standard error: spread of the sample mean.
8 When do you use the t-distribution instead of the normal distribution?	When sigma is unknown and you use s; df = n – 1.
9 What does the Pearson correlation measure?	Only the linear association between two quantitative variables, between -1 and 1.
10 When do you use the Spearman correlation?	For a rising or falling relationship that is not straight, or with outliers and ordinal data: it works on ranks.
11 What does a partial correlation do?	It measures the association between X and Y after you have removed the effect of a third variable Z.
12 Is statistical dependence the same as causality?	No: dependence says variables go together; causality also asks for direction and ceteris paribus.