SUMMARY · AT EXAM LEVEL

Module 1

Diagnostic analytics and causality

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	Measures of centre and dispersion	2
02	Probability distributions and the normal distribution	2
03	Randomness and the central limit theorem	3
04	Confidence interval	3
05	The empirical cycle and hypotheses in advance	3
06	Causal relationship scheme and roles of variables	4
07	Formulating hypotheses	5
08	Model and variables	5
09	Validity and research design	5
10	Example from the School data	5
11	Pitfalls at a glance	6
→	How to write the answer	6
→	Checklist: can I explain this?	6
→	Test yourself	7

EXAM LIKELIHOOD

Mediator and moderator CERTAIN	94%
Correlation is not causation CERTAIN	93%
Causal relationship scheme CERTAIN	92%
Formulating a testable hypothesis and null hypothesis CERTAIN	92%
Measurement levels: nominal, ordinal, continuous, discrete CERTAIN	92%
Formulating hypotheses before data collection (data dredging) CERTAIN	90%
Normal distribution and computing probabilities (z-score) CERTAIN	90%

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

KEY CONCEPTS

standard deviation · standard error · confidence interval · central limit theorem · causal relationship scheme · mediator · moderator · confounder

ON THE EXAM

Block 1

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

Block 1 of the exam (4 MC, 10 points) and question 7.3 (7 points, max 150 words) come from this module. The rules for the CI and correlation are on the formula sheet (say so in your answer); the percentages of the normal distribution (68-95-99.7) are not on it, you have to know them.

01 Measures of centre and dispersion

- **Mean** = sum / n. **Median** = middle value after sorting. **Mode** = most frequent value.
- **Skewness**: symmetric gives mean = median; right-skewed gives mean > median; left-skewed gives mean < median. The mean is sensitive to outliers, the median is not.
- **Dispersion**: minimum, maximum, range (max – min), MAD (mean absolute deviation), variance, **standard deviation** (square root of the variance), coefficient of variation = sd / mean, percentiles and IQR (75th minus 25th percentile).
- **R**: mean(x), median(x), sd(x), var(x), range(x), quantile(x, 0.25), IQR(x), summary(x). In the output of summary() you read min, 1st Qu., Median, Mean, 3rd Qu., Max.

02 Probability distributions and the normal distribution

- A **probability mass function** belongs to discrete variables (all probabilities add up to 1); a **probability density function** to continuous ones (the area under the curve is 1). The probability of one exact value of a continuous variable is 0, a probability is an area between two values.
- Normal distribution: symmetric, mean = median = mode, determined by mu and sigma. 68% within 1 sigma, 95% within 2 sigma, 99.7% within 3 sigma.
- **Z-score**: Z = (x – mu) / sigma. Example: waiting time mu = 7.1 and sigma = 2.3 weeks: P(> 11.7) has Z = 2, so about 2.5%. P(4.8 < waiting time < 9.4) has Z between -1 and 1, so about 68%.
- **R**: pnorm(x, mu, sigma) gives the probability to the left of x (left tail is the default), so 1 – pnorm(11.7, 7.1, 2.3) for the right tail. qnorm(p) gives the critical value, for example qnorm(0.975) = 1.96.

03 Randomness and the central limit theorem

- A **random variable** is a quantity whose value depends on chance. The key question of statistics: is this **signal** or **noise**?
- Notation: mu (population mean), sigma (population sd), x-bar (sample mean), s (sample sd), n (sample size).
- **Central limit theorem**: the mean of a simple random sample of size n is approximately normally distributed with expectation mu and standard deviation sigma / sqrt(n), even if the population itself is not normal. In practice this works from n of about 20 to 30.
- **Standard error** = s / sqrt(n) (with sigma if known). The SE falls with a larger n and rises with a larger s. To halve the SE you need four times as many observations.

04 Confidence interval

- Formula (on the formula sheet): CI = x-bar +/- t* x s / sqrt(n), or z* x sigma / sqrt(n) for known sigma or large n.
- Critical values: 90% gives 1.64, 95% gives 1.96 (rule of thumb 2), 99% gives 2.58. In R: qnorm(1 – alpha/2), for unknown sigma qt(0.975, df = n – 1).
- Example (Donations, simulated donations): x-bar = 36.9, s = 14.3, n = 100, so SE = 1.43 and CI = [34.1; 39.7].
- **Correct interpretation**: if you draw samples repeatedly, about 95% of the intervals contain the true mean ("we are 95% confident"). **Incorrect**: the probability that mu is in this one interval is 95%; 95% of the data lie in the interval; 95% chance that the sample mean of a repeat falls inside.
- A larger n or a smaller s gives a narrower interval; a higher confidence level gives a wider interval.

05 The empirical cycle and hypotheses in advance

- Cycle: research question, ideas (theory, logic, earlier results), hypotheses, model, data, analysis, conclusion. Three ingredients between question and data: **causal relationship scheme**, hypotheses, model.
- **Why hypotheses in advance?** (1) Focus and discipline; (2) they steer data collection, so you do not collect the wrong data or answer a different question; (3) they prevent **data dredging**: whoever keeps searching finds relationships by chance; (4) they limit confirmation bias afterwards.
- Deductive (confirmatory) research starts with hypotheses, inductive (exploratory) research looks for patterns.

06 Causal relationship scheme and roles of variables

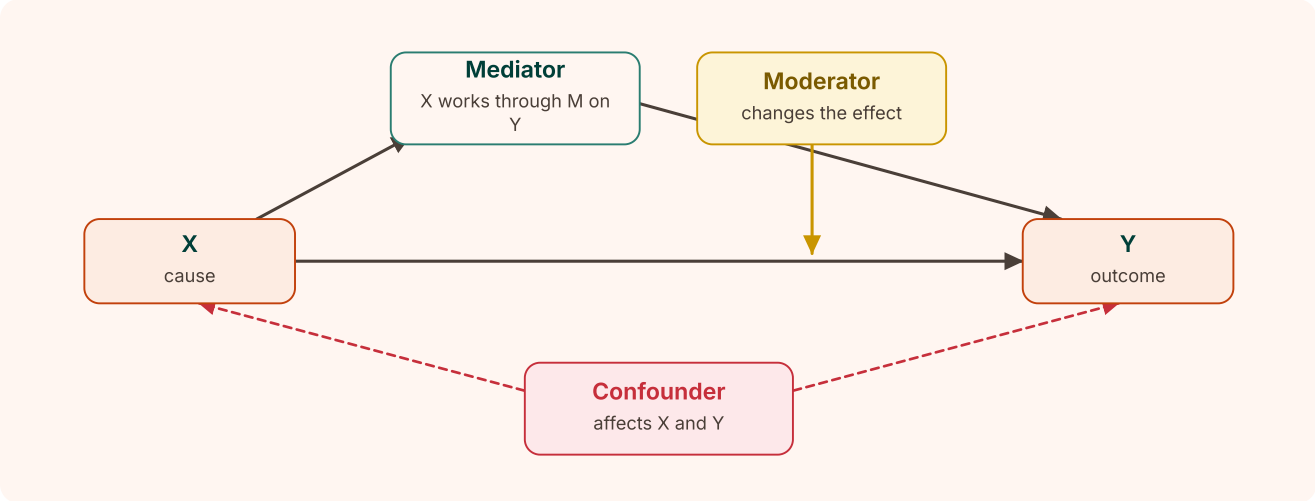


FIGURE 1.1 **Causal relationship diagram.** X affects Y directly and via a mediator; a moderator changes the strength; a confounder creates a spurious link.

Role	Meaning	Example
Direct effect	X acts directly on Y	advertising, revenue
Mediator	sits between X and Y and explains the indirect effect	advertising, store visits, revenue
Moderator	determines the strength (or sign) of the effect of X on Y: interaction effect	season (Christmas or not)
Confounder	influences both X and Y and can create a spurious relationship	temperature for ice cream and drownings
Reverse causality	Y also influences X (simultaneous effects)	revenue, budget

- Arrow symbols: + expected positive, – expected negative, * categorical variable (moderator as a nominal variable), ? competing hypotheses.
- Recognising them: "explains how or through which channel" is a mediator. "Stronger for group A than for group B" is a moderator. "Influences both" is a confounder.
- A mediator sits in the causal path, a confounder stands outside it and causes X and Y. A moderator changes the strength of an arrow, a mediator adds an extra link.
- Only the relationships you ask about go into the scheme. You usually leave control variables out.

07 Formulating hypotheses

- **Associative**: "Students who spend more hours on social media are less satisfied with their studies." **Causal**: "Limiting social media to 30 minutes a day increases satisfaction."
- Choose causal only if your design allows causal inference (experiment: you intervene and see what changes). With a survey, only an associative hypothesis.
- A good hypothesis: testable, with a direction, variables named clearly. The **H0** is the counterpart: no effect (for a directional H1: no positive influence).
- One hypothesis per arrow: H1 direct effect, H2 indirect effect via mediator ("positive indirect influence via store visits"), H3 moderator ("larger in the Christmas period than in other seasons").

08 Model and variables

- A **model** is a formal, abstract and subjective representation of reality: no vagueness, only the essential features, assumed relationships. It is the framework within which you test hypotheses. An abstract concept (construct) does not have to be directly observable; you make it measurable through operationalisation.
- **Measurement levels**: nominal (no order), ordinal (order, distances meaningless), continuous and discrete (both quantitative). In R: character or factor, ordered factor, double, integer.
- **Dependent variable** (Y, outcome) and **independent variable** (X, explanatory variable).

09 Validity and research design

- **Internal validity**: how trustworthy is the answer for the source population itself? Depends on statistics, sampling method, biases and confounders.
- **External validity**: how well can you generalise to the target population, other groups or other times? Pitfall: a source population that is too specific.
- **Experiment** versus **observational data (survey)**: an experiment (preferably with random assignment) gives strong evidence for causality; observational data gives associations. That is why causal conclusions from a survey are wrong.
- Ceteris paribus: causality means that a change in X leads to a change in Y while all other influences stay constant.

10 Example from the School data

cor(School$Sat, School$Debts) gives 0.166 and cor(School$CostTot, School$Debts) gives 0.424. That is association, not proof that higher costs cause debts: the differences between groups (tapply, aggregate) only show that averages differ.

11 Pitfalls at a glance

- Confusing SE with sd (SE is about the mean, sd about the data).
- Mixing up moderator and mediator.
- A causal conclusion from a survey.
- A hypothesis that is not testable ("social media is important") or has no direction.
- Formulating hypotheses only after looking at the data.

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 researcher sees that ice cream sales and drownings rise together. Explain whether ice cream causes drownings.

TERM A confounder is at work: a variable that affects both X and Y. DEFINITION A confounder can create a spurious relationship: X and Y move together without X causing Y. CASE FACT Warm weather raises both ice cream sales and the number of people who go swimming. CONCLUSION Correlation is not causation: the link says nothing about ice cream as a cause unless you control for the weather.

64 / 75 words

KEY CONCEPTS

✓ Checklist: can I explain this?

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

- ☐ standard deviation
- ☐ standard error
- ☐ confidence interval
- ☐ central limit theorem
- ☐ causal relationship scheme
- ☐ mediator
- ☐ moderator
- ☐ confounder
- ☐ reverse causality
- ☐ internal validity
- ☐ external validity

FLASHCARDS

Test yourself

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

Question	Answer
1 How do you calculate the mean?	Add up all values and divide by the number of observations n.
2 How do you calculate the IQR (interquartile range)?	The 75th percentile (3rd Qu.) minus the 25th (1st Qu.).
3 What does pnorm(x, mu, sigma) give in R?	Probability left of x (left tail); right tail = 1 – pnorm().
4 What is a random variable?	A quantity whose outcome depends on chance, such as the mean of a sample.
5 How many more observations do you need to halve the standard error?	Four times as many, since SE = s / sqrt(n) and sqrt(4n) = 2 x sqrt(n).
6 How do you read a 95% confidence interval correctly?	If you draw samples repeatedly, about 95% of the intervals contain the true mean.
7 What is data dredging?	Searching data for a chance relationship, with no hypotheses in advance.
8 What is a confounder?	Third variable that drives both X and Y.
9 What makes a hypothesis good?	Testable, with a direction, and with clearly named variables.
10 What is a nominal variable?	Categories without an order, such as country or gender (in R: character or factor).
11 What is internal validity?	Reliability of the answer for the source population itself.
12 Which relationships do you put in a causal relationship scheme?	Only relationships from your research question, no controls.