SUMMARY · AT EXAM LEVEL

Module 3

Hypothesis testing

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	Hypotheses and the step plan	2
02	Decision rules	2
03	Type I error, Type II error and power	2
04	The four tests	3
05	Interpreting R output	4
06	Likert scales and Cronbach's alpha	4
07	Common mistakes	4
08	At the exam	5
→	How to write the answer	5
→	Checklist: can I explain this?	5
→	Test yourself	6

EXAM LIKELIHOOD

p-value CERTAIN	97%
t-test (one-sample, two-sample, paired) CERTAIN	95%
Analysis of variance (ANOVA) and F-test CERTAIN	93%
Chi-square test: goodness of fit and independence CERTAIN	92%
Type I and Type II error CERTAIN	92%
Likert item versus Likert scale (composite score) CERTAIN	90%
Rejecting or maintaining H0 (never proving H0) CERTAIN	90%

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

KEY CONCEPTS

H0 · H1 · alpha · p-value · critical value · rejection region · Type I error · Type II error · power · one-sided · two-sided · t-test · paired t-test · ANOVA

ON THE EXAM

Block 3

15 points · 180 min in total

MC	6 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

01 Hypotheses and the step plan

A **hypothesis** is a testable statement about a population. **H1** (alternative hypothesis) is the researcher's expectation; **H0** (null hypothesis) is the opposite and is taken as true as the starting point. You reject H0 only given strong evidence; that protects against confirmation bias. Always formulate H1 first and derive H0 from it. Hypotheses are about **population parameters** (mu, p), never about sample means.

The direction of H1 determines the test: **right-tail** (H1: greater than), **left-tail** (H1: smaller than) or **two-sided** (H1: not equal, rejection region in both tails). In a two-sided test you split alpha over two tails (alpha/2 per tail), so the critical value lies further from the centre than in a one-sided test.

The seven-step plan (Neyman-Pearson): (1) state H0 and H1, (2) choose the test statistic, (3) derive the distribution under H0, (4) set the intuitive rejection region, (5) choose alpha, (6) find the critical value(s), (7) perform the test and conclude in the language of the problem. The idea: compare the sample outcome with the value under H0, divided by the standard error.

02 Decision rules

- **Critical value**: reject H0 if the test statistic falls in the rejection region (two-sided: |t| larger than the critical value).
- **p-value**: the probability of an outcome as extreme as or more extreme than the observed one, assuming H0 is true. Reject H0 if **p <= alpha**. The lower p, the stronger the evidence against H0.
- **Confidence interval**: in a two-sided test with alpha = 0.05 you reject H0 if the value under H0 lies outside the 95% interval.

Conclusion sentences: "reject H0" or "maintain H0" (never "H0 is proven" or "H1 is proven"). A p-value is not the probability that H0 is true, and not the probability that you make an error.

03 Type I error, Type II error and power

	H0 true	H0 false
Reject H0	Type I error (probability alpha)	correct (power = 1 – beta)
Maintain H0	correct	Type II error (probability beta)

A Type I error is a false positive (seeing an effect that is not there), a Type II error a false negative (missing an existing effect). The **trade-off**: a lower alpha reduces Type I but increases Type II. Both probabilities fall only with a **larger sample**. A common requirement for power is 80%. Running several tests in a row makes the total probability of at least one Type I error grow; that is a reason to use ANOVA instead of several t-tests.

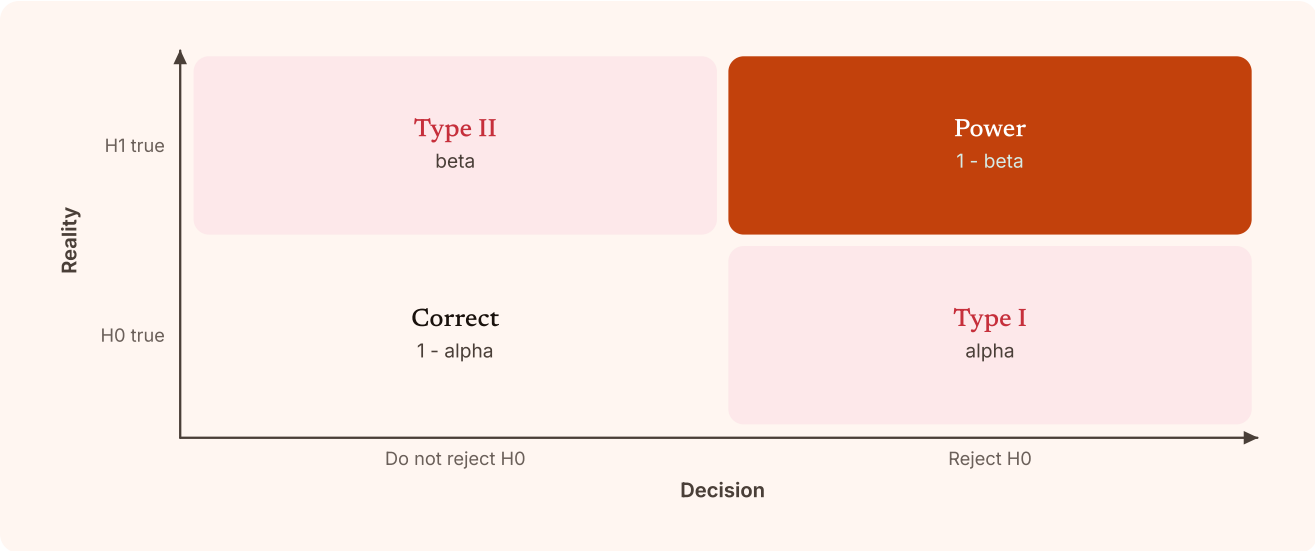


FIGURE 3.1 **Type I and Type II errors.** Alpha is the probability of a Type I error; power is 1 minus the probability of a Type II error.

04 The four tests

t-test: tests means with unknown population standard deviation; the test statistic is (mean minus value under H0) divided by the standard error, with a t distribution and df = n – 1 (one-sample). Variants:

- one-sample: mean against a fixed value;
- two-sample: two independent groups; **var.equal = TRUE** assumes equal variances, otherwise R uses the **Welch** test (fractional df, for example 628.47);
- paired: linked observations (before/after, same day, same person); you test the mean of the differences.

ANOVA: H0 is that all group means are equal; H1 that at least one group differs. The F statistic compares variation between the groups with variation within the groups. A significant F does not say which groups differ; that needs post-hoc tests (such as Tukey, which gives an adjusted p-value per pair). In the output of **summary(aov(...))** you read Df, Sum Sq, Mean Sq, F value and **Pr(>F)**.

z-test: known population standard deviation, or a hypothesis about a proportion. The test statistic is z = (p-hat – p0) / SE, with SE = sqrt(p0 (1 – p0) / n) under H0.

chi-square test: for two categorical variables. H0: the variables are independent. The expected frequency per cell is (row total x column total) / total; the statistic is the sum of (observed – expected)^2 / expected, with df = (rows – 1) x (columns – 1). Also usable as a goodness-of-fit test against an expected distribution. A low X-squared (and high p) means: no evidence of association.

Situation	Test	R
one mean against a value	one-sample t	t.test(x, mu = ...)
two groups	two-sample t	t.test(y ~ g)
linked pairs	paired t	t.test(a, b, paired = TRUE)
three or more groups	ANOVA	aov(y ~ g)
two categorical variables	chi-square	chisq.test(table)

05 Interpreting R output

Look in the output for: the test statistic (t, F, X-squared), the df, the p-value, the interval and the estimate (sample estimates). Note the **alternative hypothesis** that R prints ("true mean is not equal to ..."): that is H1. First check whether the test fits the question, then whether p <= alpha, then the direction of the effect in the means. For the chi-square test: check whether the association is practically relevant, not only whether it is significant.

06 Likert scales and Cronbach's alpha

A **Likert scale** measures an abstract construct with several items, a rating scale (for example 1 to 5) and an instruction. A **reverse coded** item is phrased negatively; you flip it (new score = max + min – score, so 6 – score for 1 to 5) before you compute the scale score. Such items keep respondents alert. Other quality tools: **attention checks** and removing **flatliners** and respondents who are too fast.

Cronbach's alpha measures internal consistency, usually between 0 and 1; 0.7 is the usual lower bound. Alpha rises with more items and with higher inter-item correlation. A negative alpha points to a forgotten reversal. Example from the course data: alpha was negative until the three reversed items were flipped, afterwards about 0.80 for eight items. The column "alpha if item dropped" shows which item weakens the scale.

07 WATCH OUT Common mistakes

- ✖ Using a two-sample t-test on linked data, or several t-tests for three groups.
- ✖ Forgetting that in a one-sided test the p-value is half of the two-sided one (if the effect lies in the right direction).
- ✖ Confusing significance with practical importance in large samples.

08 At the exam

About 6 multiple-choice questions on module 3 (the questions come from the practice exams). The formula sheet gives the formulas for the confidence interval, the t and z statistics, chi-square and Cronbach's alpha; the F formula and the p-value calculation are not on it. You must be able to draw a conclusion from R output of **t.test**, **chisq.test** or **aov**, but you do not write code.

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 test gives p = 0.03 at alpha = 0.05. What is your conclusion?

TERM The p-value is smaller than alpha. DEFINITION The p-value is the probability of a result at least as extreme as the one found, if H0 is true. CASE FACT Here 0.03 < 0.05, so the data are unlikely under H0. CONCLUSION Reject H0. This is not the probability that H0 is true, and it says nothing about the size of the effect.

58 / 75 words

KEY CONCEPTS

✓ Checklist: can I explain this?

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

- ☐ H0
- ☐ H1
- ☐ alpha
- ☐ p-value
- ☐ critical value
- ☐ rejection region
- ☐ Type I error
- ☐ Type II error
- ☐ power
- ☐ one-sided
- ☐ two-sided
- ☐ t-test
- ☐ paired t-test
- ☐ ANOVA
- ☐ z-test
- ☐ chi-square test
- ☐ Likert scale
- ☐ reverse coded

- [ ] Cronbach's alpha

FLASHCARDS

Test yourself

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

Question	Answer
1 What is a hypothesis in the context of this course?	A testable, provisional statement about an expected relationship between variables or about a property of a population, such as its mean.
2 Which hypothesis do you best start with when formulating a pair?	With H1, the expectation you want to show; H0 then follows as its exact opposite.
3 What is the critical value?	The boundary of the rejection region; it follows from alpha and the distribution of the test statistic under H0.
4 What is the p-value?	The probability of a test result as extreme as or more extreme than the one found, assuming H0 is true.
5 What is a Type I error?	Rejecting H0 while H0 is true (false alarm); the probability of this is alpha.
6 What does an alpha of 0.05 say about the probability of a Type I error?	In the long run you wrongly reject H0 in 5% of cases when H0 is true.
7 What does the argument var.equal do in t.test()?	It switches the assumption on or off that the variance is equal in both groups; this determines how the standard error is computed.
8 What does ANOVA test?	Whether the means of three or more groups are equal; H0 is that all group means are equal.
9 When do you use a z-test?	When the population standard deviation is known or for a hypothesis about a proportion.
10 What is H0 in a chi-square test of independence?	The two categorical variables are independent; there is no association between them.
11 What are reverse coded items?	Negatively worded items that you must flip (on a 5-point scale 6 minus the score) before you take a total score.
12 What does alpha if item dropped show?	The alpha you get if you drop that one item; a value above the total alpha points to a weak item.