

# Inferential Statistics

## Lecture 8: Principles of statistical testing

Jérémy Hervelin

Pforzheim University  
j.hervelin@lb.hs-pforzheim.de

9 November 2026 15:30–17:00  
Winter semester 2026/27

Adapted from the teaching materials of Prof. Dr. Thilo Klein.

# Our 90-minute route

- First 15 minutes: retrieve prerequisites and introduce the question.
- Next 35 minutes: worked examples, with short checks after each concept.
- Next 25 minutes: paired activity and interpretation.
- Final 15 minutes: synthesis, questions and an applied discussion.

# A test is a decision rule under uncertainty

- Formulate a population claim and a competing claim.
- Choose a test statistic and significance level before examining the result.
- Understand Type I error, Type II error, power and p-values.
- Report statistical evidence with effect sizes and uncertainty.

# Original five-coin experiment

Let  $X$  be the number of heads in five independent tosses.

$$H_0 : p = 0.5, \quad H_1 : p \neq 0.5.$$

Consider rejecting if  $X = 0$  or  $X = 5$ . Under  $H_0$ :

$$P(\text{reject}) = 2(0.5)^5 = 0.0625.$$

This rule has size 6.25%, so it is not a test at the 5% level. Discrete rejection probabilities cannot take arbitrary values.

# Classroom check – Q1

## Exam-style

1–2 minutes

Reject a fair-coin null after 5 tosses only for all heads or all tails. Find the null rejection probability.

*Think individually, then compare your reasoning with a neighbour.*

Lecture 8, classroom quiz Q1. Same question number as the printed quiz.

# State the hypotheses about a parameter

Illustrative employment benchmark:

$$H_0 : p = 0.50, \quad H_1 : p \neq 0.50.$$

A prespecified improvement question might instead be

$$H_0 : p \leq 0.50, \quad H_1 : p > 0.50.$$

The null concerns the population rate, not the observed sample share. A one-sided test calibrates its rejection probability over the composite null.

# A statistic translates data into evidence

A statistic is a function of the sample. A test statistic is chosen so that its null distribution can be evaluated.

For fixed  $n$ , the success count  $X$  and proportion  $X/n$  contain identical information. Either can define the same rejection rule.

Standardisation often makes a statistic comparable across sample sizes and measurement scales.

# Two possible errors

|                | Null true             | Alternative true  |
|----------------|-----------------------|-------------------|
| Reject null    | Type I error          | Correct rejection |
| Fail to reject | Correct non-rejection | Type II error     |

$\alpha$  bounds Type I error. At a specified alternative value  $\theta$ , Type II error is  $\beta(\theta)$  and power is  $1 - \beta(\theta)$ .

These are probabilities conditional on a data-generating parameter, not probabilities that a hypothesis is true.

# Classroom check – Q3

## Concept check

1–2 minutes

A Type I error is. . .

- A. Failing to reject a false null;
- B. Rejecting a true null;
- C. An inaccurate sample mean

*Think individually, then compare your reasoning with a neighbour.*

Lecture 8, classroom quiz Q3. Same question number as the printed quiz.

# Significance level and rejection region

A level- $\alpha$  test satisfies

$$P_{\theta}(\text{reject } H_0) \leq \alpha \quad \text{for every } \theta \in H_0.$$

The actual rejection probability may be below the chosen level, especially for discrete tests.

Smaller  $\alpha$  generally reduces power when the sample size and test family are fixed. Increasing sample size can improve power while retaining the same significance level.

# A p-value measures incompatibility under the null

A p-value is the probability, calculated under the null model, of a test statistic at least as incompatible with the null as the observed statistic, using the test's definition of extremeness.

It is not  $P(H_0 \mid \text{data})$ . It is not the probability that the result occurred “by chance”.

For a valid prespecified test, reject when  $p \leq \alpha$ .

# Classroom check – Q2

## Exam-style

1–2 minutes

A p-value of 0.03 means. . .

- A.  $P(H_0 \text{ is true})=0.03$ ;
- B. Under  $H_0$ , outcomes at least as incompatible occur with probability 0.03;
- C. The estimated effect is large

*Think individually, then compare your reasoning with a neighbour.*

Lecture 8, classroom quiz Q2. Same question number as the printed quiz.

# Normal test: a worked illustration

Synthetic earnings sample:  $n = 100$ ,  $\bar{x} = 1800$ , known  $\sigma = 400$ . Test  $H_0 : \mu = 1700$  against  $H_1 : \mu \neq 1700$ .

$$Z = \frac{1800 - 1700}{400/\sqrt{100}} = 2.5.$$

Two-sided p-value:

$$2[1 - \Phi(2.5)] \approx 0.0124.$$

At 5% we reject; at 1% we do not. Neither decision establishes the size or importance of a policy effect.

# Why failure to reject is not acceptance

A non-significant result can arise because the null is compatible with the data, because the sample is imprecise, or because the test has little power for the relevant alternative.

Report the estimate and confidence interval. Distinguish “no convincing evidence of a difference” from “convincing evidence of no meaningful difference”.

The latter requires a suitable equivalence question and design, beyond today's test.

# Classroom check – Q5

## Exam-style

1–2 minutes

After  $p=0.20$ , a sound conclusion is. . .

- A. The null is proved;
- B. Fail to reject at 5%; consider effect estimate and uncertainty;
- C. There is a 20% chance of no effect

*Think individually, then compare your reasoning with a neighbour.*

Lecture 8, classroom quiz Q5. Same question number as the printed quiz.

# Power depends on a specified alternative

For the normal one-sided mean test, with known  $\sigma$ , reject above  $\mu_0 + z_{1-\alpha}\sigma/\sqrt{n}$ .

At  $\mu_1 > \mu_0$ :

$$\text{power}(\mu_1) = 1 - \Phi\left(z_{1-\alpha} - \frac{(\mu_1 - \mu_0)\sqrt{n}}{\sigma}\right).$$

Power grows with sample size and effect magnitude, and falls with noise. There is no single power value without an alternative.

# Classroom check – Q4

## Exam-style

1–2 minutes

At a specified alternative, Type II error probability is 0.2. What is power?

*Think individually, then compare your reasoning with a neighbour.*

Lecture 8, classroom quiz Q4. Same question number as the printed quiz.

# A simple planning calculation

For this one-sided normal test, desired power  $1 - \beta$  at difference  $\delta > 0$  gives

$$n \geq \left( \frac{(z_{1-\alpha} + z_{1-\beta})\sigma}{\delta} \right)^2.$$

Round up. This assumes independent normal observations and known SD.

Planning for a two-group employment intervention requires a different variance calculation and attention to design and attrition.

# Confidence intervals and two-sided tests

For the same model and matching construction, a level-5% two-sided test rejects a parameter value exactly when it lies outside the 95% confidence interval.

A confidence interval displays the range of parameter values compatible with the data at that level. It adds information about magnitude and precision that a yes/no decision hides.

# Report an employment comparison responsibly

- State the population, sampling or assignment design, and target parameter.
- Report the estimated difference in percentage points with an interval.
- Describe the prespecified hypothesis, method and p-value.
- Discuss practical importance and limits: an association alone does not establish causation.

# Discussion: spot the misleading statement

- A p-value of 0.03 means a 3% probability that the null is true.
- A non-significant difference proves that the programme has no effect.
- A significant difference must be large enough to matter.
- Choose a one-sided test after seeing a positive difference.
- For each statement, explain the mistake and write a defensible replacement.

# Independent practice

- Practise the five online questions for this lecture; 80% provides access to the next practice module.
- Course materials remain accessible regardless of your practice progress.
- Explain your reasoning before comparing answers.
- Review the method, assumptions and interpretation as well as the calculation.

# Sources and teaching notes

Adapted from Prof. Dr. Thilo Klein's supplied Inferential Statistics materials.

Revised teaching edition: Jérémy Hervein, Pforzheim University, winter semester 2026/27.

New youth-employment, programme and interview-score examples are synthetic illustrations.