

# Inferential Statistics

## Lecture 11: Two-sample inference

Jérémy Hervelin

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

30 November 2026 13:45–15:15  
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.

# Two samples: start with the design

- Distinguish paired observations from independent groups.
- Apply a t test to within-pair differences.
- Compare independent means using Welch or a justified pooled model.
- Understand the normal-model F test for two variances.

# Paired or independent?

- Before and after outcomes for the same people are paired.
- Two outcomes for each matched pair are paired.
- Different independently sampled people in two groups are independent samples.
- Equal group sizes do not make a design paired; pairing is a feature of data collection.

# Classroom check – Q1

## Exam-style

1–2 minutes

Scores before and after training for the same people should usually use. . .

- A. A paired comparison;
- B. An independent-groups test;
- C. An F test of means

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

Lecture 11, classroom quiz Q1. Same question number as the printed quiz.

# Original paired example

For ten matched observations, define the difference in the prespecified direction:

$$d = (60, 32, -1, 79, 26, 28, 30, -7, 61, 35).$$

Analyse one sample of differences. Dependence within each pair is allowed; independence across pairs is required.

A before-after change may reflect time trends as well as a programme. Pairing alone does not establish causality.

# Paired t test

Test  $H_0 : \mu_D = 0$  using

$$T = \frac{\bar{D}}{S_D/\sqrt{n}} \sim t_{n-1}$$

under iid normal differences and the null.

For the original ten differences,  $\bar{d} = 34.3$ ,  $t = 4.0498$ ,  $df = 9$ . The prespecified upper-tail p-value is 0.001443; the one-sided 95% lower bound is 18.774.

# A paired sign test asks a different question

Eight of the ten original differences are positive. With no zero differences, the one-sided sign-test p-value is

$$P\{\text{Bin}(10, 0.5) \geq 8\} = 0.0546875.$$

It tests a sign/median question under the appropriate model, not the mean difference.

Discuss why this result differs from the paired t test. The sign test ignores the magnitudes of positive and negative differences.

# Independent group means

Let group 1 have  $(n_1, \bar{x}_1, s_1^2)$  and group 2 have  $(n_2, \bar{x}_2, s_2^2)$ . Estimate

$$\Delta = \mu_1 - \mu_2 \quad \text{by} \quad \hat{\Delta} = \bar{x}_1 - \bar{x}_2.$$

For independent groups,

$$\widehat{SE}(\hat{\Delta}) = \sqrt{s_1^2/n_1 + s_2^2/n_2}.$$

Keep the subtraction direction consistent in the hypotheses, estimate and interpretation.

# Welch two-sample t test

For  $H_0 : \Delta = \Delta_0$ ,

$$T = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{\sqrt{s_1^2/n_1 + s_2^2/n_2}}.$$

An approximate  $t$  reference distribution uses

$$\nu = \frac{(s_1^2/n_1 + s_2^2/n_2)^2}{(s_1^2/n_1)^2/(n_1 - 1) + (s_2^2/n_2)^2/(n_2 - 1)}.$$

Welch allows unequal variances. It is a sensible default for independent means; it does not fix dependence or extreme non-normality.

# Classroom check – Q3

## Concept check

1–2 minutes

Welch's independent-means test. . .

- A. Requires equal population variances;
- B. Allows unequal variances;
- C. Makes any observational difference causal

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

Lecture 11, classroom quiz Q3. Same question number as the printed quiz.

# The equal-variance pooled model

If independent normal populations share a common variance,

$$s_p^2 = \frac{(n_1 - 1)s_1^2 + (n_2 - 1)s_2^2}{n_1 + n_2 - 2},$$

$$T = \frac{\bar{x}_1 - \bar{x}_2 - \Delta_0}{s_p \sqrt{1/n_1 + 1/n_2}} \sim t_{n_1 + n_2 - 2}.$$

Use this when the common-variance assumption is defensible. A non-significant preliminary variance test does not prove it.

# Original PISA comparison

Original groups of 8 and 10 scores have means 500.125 and 480.5. The estimated difference is 19.625 points.

Welch:  $t = 1.3377$ ,  $df \approx 15.051$ ,  $p = 0.2009$ , 95% CI  $[-11.64, 50.89]$ .

Pooled:  $t = 1.3398$ ,  $df = 16$ ,  $p = 0.1990$ , 95% CI  $[-11.43, 50.68]$ .

The similar answers here do not make the assumptions interchangeable.

# Comparing employment rates

Synthetic independent groups: 58% of 100 programme participants and 49% of 100 comparison participants are employed.

$$\hat{p}_1 - \hat{p}_2 = 0.09 = 9 \text{ percentage points.}$$

An approximate unpooled SE is

$$\sqrt{0.58(0.42)/100 + 0.49(0.51)/100} \approx 0.0702.$$

Distinguish 9 percentage points from a 9% relative change. Causal interpretation depends on assignment and follow-up, not just the test.

# Classroom check – Q2

## Exam-style

1–2 minutes

Programme employment is 58%, comparison employment is 49%. What is the gap in percentage points?

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

Lecture 11, classroom quiz Q2. Same question number as the printed quiz.

# The F test compares two variances

For independent normal samples under  $H_0 : \sigma_1^2 = \sigma_2^2$ ,

$$F = S_1^2/S_2^2 \sim F_{n_1-1, n_2-1}.$$

For a two-sided level- $\alpha$  test, reject below  $F_{n_1-1, n_2-1; \alpha/2}$  or above  $F_{n_1-1, n_2-1; 1-\alpha/2}$ .

Keep group 1 in the numerator and preserve the degree-of-freedom order. The F test is sensitive to non-normality.

# Classroom check – Q4

## Concept check

1–2 minutes

For an F ratio with fixed numerator group and a two-sided level- $\alpha$  test, the upper quantile is. . .

- A.  $1-\alpha$ ;
- B.  $1-\alpha/2$ ;
- C.  $\alpha/2$

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

Lecture 11, classroom quiz Q4. Same question number as the printed quiz.

# Original variance comparison

For the PISA groups,  $s_1^2/s_2^2 = 1.026885$  with degrees of freedom 7 and 9.

The two-sided F-test p-value is 0.9472. The 95% confidence interval for the variance ratio is approximately  $[0.2447, 4.9529]$ .

This wide interval permits substantial variance differences. Failure to reject equality is not evidence of precise equality.

# Random assignment and interpretation

- Random assignment balances pre-treatment characteristics in expectation; realised groups can differ.
- Follow the assignment and outcome collection design when calculating uncertainty.
- A before-after comparison can be confounded by time; an observational group comparison can be confounded by selection.
- Statistical significance does not repair a weak causal design.

# Classroom check – Q5

## Concept check

1–2 minutes

Random assignment guarantees identical observed baseline group means.

- A. True;
- B. False

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

Lecture 11, classroom quiz Q5. Same question number as the printed quiz.

# Design workshop

- Same workers before and after training: identify the unit and paired differences.
- Independent customers in two advertisement groups: justify an independent comparison.
- Scores recorded for classmates in the same schools: discuss dependence within schools.
- For each design, specify a target, direction, method, assumptions and a limitation.

# 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.