Statistics

Course leader: Marco Ortu
Other instructors:

Aims of the course

Provide PhD students in Economics and Business with a compact but rigorous introduction to core inferential statistics (hypothesis testing and regression-based inference) and to Principal Component Analysis (PCA), including a hands-on laboratory session in Python.

Learning outcomes and competences

- Formulate research questions as statistical hypotheses and select appropriate tests.
- Compute and interpret p-values, confidence intervals, effect sizes, and statistical power.
- Apply classical tests and regression-based inference, including basic diagnostic checks.
- Understand PCA (scores, loadings, explained variance) and interpret components in economic/business contexts.
- Implement a reproducible PCA workflow in Python using standard pipelines and diagnostics.

Pre-requisites

Basic probability and statistics (descriptive statistics, distributions), and familiarity with linear algebra (vectors, matrices). Prior exposure to linear regression is recommended. Basic Python proficiency is useful for the laboratory session.

Course contents and syllabus

Lesson 1 (3h) - Foundations of inference and hypothesis testing

- Statistical models, parameters, estimates; sampling vs population; i.i.d. vs dependent data.
- Estimation and uncertainty: standard errors, confidence intervals, asymptotic approximations (CLT).
- Hypothesis testing framework: null/alternative, test statistics, rejection regions.
- Type I/II errors, power, effect size; pre-specification and research integrity issues.

Lesson 2 (3h) - Classical tests and regression-based inference

- Two-sample comparisons (independent/paired), unequal variances; nonparametric alternatives (overview).
- Chi-square tests (goodness-of-fit, independence); ANOVA as model comparison.
- Regression inference: linear model assumptions, t/F tests, interpretation of coefficients, omitted-variable risk (conceptual), robust standard errors (overview), model diagnostics.

Lesson 3 (3h) - Principal Component Analysis

- Motivation: dimension reduction, multicollinearity, latent factors; exploratory vs confirmatory uses.
- Preprocessing: centering, scaling/standardization; effect of units; missingness and outliers.
- PCA definitions: variance maximization and minimum reconstruction error; equivalence of formulations.
- SVD/eigendecomposition view: covariance vs correlation; scores and loadings; explained variance.

- Choosing the number of components: scree plot, cumulative explained variance, parallel analysis.
- Interpretation: variable contributions, score plots and biplot.
- Limitations and pitfalls: interpretability, sign indeterminacy, linearity, influence of outliers.

Lesson 4 (3h) - PCA applications with Python and real data examples (hands-on lab)

- Reproducible workflow: data import, cleaning, type handling (numeric vs categorical), missing-data strategies.
- Standardization and pipelines: StandardScaler + PCA with sklearn.pipeline.
- Diagnostics and reporting: explained variance ratio and cumulative curves; scree plot; loadings table; score plots.
- Interpretation in practice: component meaning, variable contributions; biplot-style visualization.
- Sensitivity checks: with/without scaling; influence of outliers; alternative scalers.

Organization of the course

The course consists of 12 hours of lectures and guided exercises, organized in 4 lessons of 3 hours each. Lesson 4 is a hands-on PCA laboratory in Python using real data examples.

The relevant materials can be found in the *Teams class* "PhD Programme in Economics and Business" following this path: Documenti/General/Class Materials/Statistics
First year PhD students are made members by using their UniCa email account. Students can access the *Teams* application by using the same account.

The timetable of the course can be found in the Team class calendar.

Assessment method

The examination consists of 30 single-choice questions. Each question has one correct answer. Scoring: 1 point for each correct answer; 0 points for each incorrect answer; no penalties for incorrect answers. The passing threshold is 18/30 (at least 18 correct answers out of 30).

Reading list

James, G., Witten, D., Hastie, T., Tibshirani, R. (latest edition). An Introduction to Statistical Learning (ISLP, Python edition): Linear Regression (Ch. 3, p. 69–), PCA theory (Ch. 12.2, pp. 504–515), PCA lab in Python (Ch. 12.5.1, p. 535–), Hypothesis Testing review (Ch. 13.1, pp. 558–562).

Other information