

Exam info:

The exam is based on all material that was discussed in class. So I recommend to use your class notes (the slides and the notes you took from what I wrote on the board) as the basis for your exam preparation. The exam will be a mix of practical and theoretical problems. There will be at least one question that is taken from the exercises. You should be able to understand and explain the R-code that was used in the class. You don't have to produce R-code yourself during the exam.

Below is a list of topics that are likely to come up at the exam. This list is not complete, but is meant to give you a better idea of what type of questions you can expect. You should be able to:

- Study design
 - Identify between different types of studies (e.g., observational study, double/single-blind randomized controlled experiment, etc)
 - Discuss the elements of the 'optimal' (=double-blind randomized controlled experiment) study design, and explain the role of each of the elements.
 - Discuss situations where the 'optimal' design is not possible.
 - Explain the concept *non-adherence*, why this is a problem, and how we can deal with it.
 - Explain the concept *confounding*, give examples of it in practical situations, describe solutions for dealing with it.
 - Describe the distinction between association and causation, give examples where association does not imply causation.
 - Describe Simpson's paradox, and give an example of it.
- Sampling
 - Describe ways to get a bad sample (convenience sample)
 - Describe the method *quota sampling*, including its pros and cons.
 - Describe *probability sampling methods*, including *simple random sampling* and *cluster sampling*. Explain why these methods give representative samples, and give examples of situations in which each of the methods can be used.
- Graphics and summary statistics for multivariate data
 - Interpret a *pairs plot* (scatter plot matrix), *star plot*, *segment plot*, *conditioning plot*, and describe their pros and cons.
 - Describe summary statistics used for multivariate data.
- Principal component analysis (PCA)
 - Describe the method: what is it used for, what is the algorithm, what are limitations.
 - Give the properties of PCA (for population version and sample version).
 - Discuss ways to choose the number of principal components.

- Interpret R-output of principal component analysis.
- Reproduce all proofs we discussed in class.
- Factor analysis
 - Describe possible uses of factor analysis, and limitations of this method.
 - Describe the factor analysis model, including its assumptions.
 - Describe two methods for factor analysis: principal factor analysis and maximum likelihood estimation.
 - Explain what *Heywood cases* are.
 - Discuss why factor loadings are non-unique, how this non-uniqueness can be resolved.
 - Describe methods for factor rotation (orthogonal: varimax; oblique: promax)
 - Describe how factor scores can be computed (Bartlett's method; Thompson's method)
 - Discuss commonalities and differences between PCA and factor analysis.
 - Interpret R-output of factor analysis.
 - Reproduce all proofs we discussed in class.
- Multivariate normal distribution and related distributions
 - Give definitions of the multivariate normal, the Wishart, and Hotelling's T^2 distribution.
 - Give reasons for the importance of the multivariate normal distribution in multivariate statistics.
 - Give an example of variables with univariate normal distributions that do not have a multivariate normal distribution.
 - State relationship between the T^2 distribution and the F distribution.
 - Give the definition of Mahalanobis distance.
 - Reproduce all proofs we discussed in class.
- One- and two-sample test for means
 - Explain general idea of a statistical test, including the definition of a p-value.
 - Describe one-sample T^2 test, including its assumptions.
 - Describe two-sample T^2 test, including its assumptions.
 - Discuss difference between performing one multivariate test and several univariate tests.
- Classification: Linear discriminant analysis
 - Describe Fisher's linear discriminant analysis method.
 - Reproduce the proof of the formula for Fisher's linear discriminant analysis (handout).
 - Interpret R-output of discriminant analysis.
 - Discuss how classification methods can be evaluated.

- Describe maximum likelihood rule for linear discriminant analysis.
- Clustering:
 - Identify situations in which clustering methods can be used, give possible applications
 - Describe the three methods we discussed: hierarchical clustering, k -means clustering, model based clustering
 - For hierarchical clustering: describe various ways to measure distances between points and clusters
 - For k -means clustering: describe methods to choose k