Department of Statistics
Autumn, 2024

Course Description

Selected statistical methods with applications, ST3201, 7.5 hp

1. Course content

The course aims to broaden students' knowledge in statistics and give inspiration when choosing a topic for the bachelor thesis.

The course comprises three selected topics: survival analysis, handling missing data with imputations and text analysis. For each topic, the course relates theory to practice by combining lectures, practical exercises and programming. A central element is the students' independent work with a number of problems (case studies).

2. Intended learning outcomes

After completing the course, students should be able to:

Knowledge and comprehension:

- describe definitions and central concepts,
- explain relevant theory,
- discuss the pros and cons of different methodological considerations.

Intellectual and practical skills:

- solve method-specific theoretical problems,
- apply the selected methods to various problems,
- perform calculations and analyses in R,
- present the method area and applicable problem solutions in writing and orally.

Transferable skills and critical and independent thinking:

- interpret, evaluate and critically review results with regard to relevant scientific aspects.

3. Teaching format

Teaching consists of lectures (F1-F12) and computer exercises (C1-C3).

Attendance at the seminars is compulsory. Absence from one or more seminars can be compensated by attending a recapitulation seminar. The course examiner will inform about the date for this seminar via Athena.

Submission of home assignments before seminars and oral presentations at seminars are mandatory elements of the course.

Medium of instruction is English. The course is given on campus, but there may be elements of digital teaching (this will be announced via Athena).

Part 1: **Analysis of survival data**

This part provides introduction to models and methods used in the analysis of survival (duration) data - with applications in the social sciences. Relevant R-procedures will be also covered in the computer session. Preliminary topics that will be covered include: censoring and other special features of survival data; functions of survival time; Kaplan-Meier and life-table estimation of survival functions; log-rank test for comparison of survival functions; Cox proportional hazards models; parametric survival models.
For successful completion of this part students should be able to: describe and explain basic concepts, functions, and distributions for survival data; compute and compare survival functions for different groups; and model associations between survival functions and explanatory variables using R.

Teaching consists of 4 lectures (L1 – L4) and one computer-session (C1) according to the schedule. Lecture notes and other relevant material can be handed-out (or made available in Athena) in connection with the lectures and computer session.

Literature: Moore, D. F. (2016), *Applied Survival Analysis Using R*. Springer. Available online via SU-library.
Responsible (examiner) for part 1: Gebrenegus Ghilagaber, e-mail: Gebre@stat.su.se.
Consultation hours: in connection with lectures or by appointment.

Part 2: **Handling missing data with imputations**

This part provides an introduction to handling missing data with imputations. The following topics will be considered: reasons for missing data, types of missing data, and methods to dealing with missing data (simple and multiple imputation methods). Additionally, in the computer session students will learn how to implement the imputation techniques using R.
For successful completion of this part students should be able to: describe and explain basic concepts related to missing data and imputation techniques, and be able to handle missing data of different types using software R.

Teaching consists of 4 lectures (L5 – L8) and one computer-session (C2) according to the schedule. Lecture notes and other relevant material will be available in Athena in connection with the lectures and computer session.

Literature: research papers and other materials provided by instructor, the material will be available via Athena at course start.
Responsible (examiner) for part 2: Tatjana von Rosen, e-mail: tatjana.vonrosen@stat.su.se.
Consultation hours: in connection with lectures or by appointment.

Part 3: **Text mining**

Nowadays, there are vast quantities of unstructured textual information available, for example, from emails, medical journals, social media activities, movie recommendations and web server logs. It is of utmost importance e.g. for companies to be able to analyze such information and make it quantifiable, in order to see trends and remain competitive.

This part provides an introduction to quantitative methods for analyzing text. Students will learn how to retrieve the text from the original source, process the text data, analyze and summarize the results of text mining experiments.
For successful completion of this part students should be able to: describe and explain basic concepts related to text mining, apply text mining methods to practical problems using software R.

Teaching consists of 4 lectures (L9 – L12) and one computer-session (C3) according to the schedule. Lecture notes and other relevant materials will be available in Athena in connection with the lectures and computer session.

Literature: research papers and other material provided by instructor, the material will be available via Athena at course start.
Responsible (examiner) for part 3: Tatjana von Rosen, e-mail: tatjana.vonrosen@stat.su.se.
Consultation hours: in connection with lectures or by appointment.

4. Examination and assessment criteria

a) The course is examined through three individual assignments, each comprising a written report and an oral presentation of this at the mandatory seminar.
To pass the course, a minimum of 17 points must be achieved on each individual assignment: 14 points must be achieved on the written report and 3 points on the oral presentation.
The sum of the points from the three individual assignments defines the final grade (possible maximum 100 points, Part 1: 33 points, Part 2: 34 points, Part 3: 33 points).
b) Grading of the course is done according to a seven-point scale related to the specified learning outcomes: A = Excellent, B = Very good, C = Good, D = Satisfactory, E = Sufficient, F = Insufficient.
c) For grades E-A on the course, besides the minimum of 17 points for each individual assignment, it is required the attendance at all three seminars.
Grade F requires re-examination.
d) No points from the assignments achieved at HT24 can be transferred to the next time the course will be given.
Students who get less than 17 points on an assignment or who fail to submit it before hand-in date are given one opportunity to re-submit it until a second hand-in date but can then get a maximum of only 20 points for this assignment.

The grading requirements are as follows.

A: Excellent (90-100 points). The student has in a well-structured and correct manner solved the pre-specified statistical problem that reflect the course material using software R. Furthermore, the student has also demonstrated the ability to solve problems that have not explicitly been explored in the course material. The student was able to draw correct conclusions from the statistical analysis and clearly present the obtained results (in a written report and orally).
B: Very good (80-89 points). The student has in a well-structured and correct manner solved the pre-specified statistical problem that reflect the course material using software R. The student has demonstrated the ability to solve problems that were partly explored in the course material. The student was able to draw correct conclusions from the statistical analysis, and clearly present the obtained results (in a written report and orally.
C: Good (70-79 points). The student has correctly solved the pre-specified statistical problem that reflect the course material and that was directly explored in the course material. The student was able to use software R for performing statistical analysis, to draw correct conclusions, interpret and discuss the obtained results (in a written report and orally).
D: Satisfactory (60-69 points). The student has mostly correctly solved the pre-specified statistical problem that reflect the course material and that was directly explored in the course material. The

student was able to use software R for conducting the statistical analysis, to draw mostly correct conclusions from this and to interpret the obtained results (in a written report and orally).

E: Sufficient (51-59 points). The student could, for the most part, correctly solve the pre-specified statistical problem that reflect the course material and that was directly explored in the course material. The student was able to use software R for conducting the statistical analysis, to draw in most cases correct conclusions and to interpret the obtained results (in a written report and orally).

F: Insufficient (0-50 points). The student cannot correctly apply statistical methods that have been considered in the course. The student fulfils some but not all requirements for an E grade.