Universitat Autònoma de Barcelona

2022/2023

Exploratory Data Analysis

Code: 104853
ECTS Credits: 6

Degree	Type	Year	Semester
2503852 Applied Statistics	FB	1	1

Contact

Name: Rosa Camps Camprubi
Email: rosa.camps@uab.cat

Use of Languages

Principal working language: catalan (cat)
Some groups entirely in English: No
Some groups entirely in Catalan: Yes
Some groups entirely in Spanish: No

Teachers

Montserrat Ferre Delgado

Prerequisites

None.

Objectives and Contextualisation

Learn the descriptive and exploratory techniques applied to summarize the information contained in theexperimental datasets. Also the interpretation of the results and the diagrams in the context of the data,

Finally it is important that the students learn to use statistical software to manipulate data, perform descriptive analysis and graphs.

Competences

- Analyse data using statistical methods and techniques, working with data of different types.
- Correctly use a wide range of statistical software and programming languages, choosing the best one for each analysis, and adapting it to new necessities.
- Make efficient use of the literature and digital resources to obtain information.
- Select the sources and techniques for acquiring and managing data for statistical processing purposes.
- Students must be capable of applying their knowledge to their work or vocation in a professional way and they should have building arguments and problem resolution skills within their area of study.
- Students must be capable of communicating information, ideas, problems and solutions to both specialised and non-specialised audiences.
- Students must have and understand knowledge of an area of study built on the basis of general secondary education, and while it relies on some advanced textbooks it also includes some aspects coming from the forefront of its field of study.
- Summarise and discover behaviour patterns in data exploration.
- Use quality criteria to critically assess the work done.

- Work cooperatively in a multidisciplinary context, respecting the roles of the different members of the team.

Learning Outcomes

1. Critically assess the work done on the basis of quality criteria.
2. Describe, using suitable graphical and analytic methods, qualitative data related to one or more variables.
3. Describe, using suitable graphical and analytic methods, quantitative data related to one or more variables.
4. Design syntax modifications to programmes in order to conduct new processes.
5. Explore behaviour patterns of bivariant data.
6. Explore behaviour patterns of univariant data.
7. Identify and select the most important information sources for the descriptive analysis of data of different types: social, environmental, medical, economic, etc.
8. Make effective use of references and electronic resources to obtain information.
9. Purge data: lost data, transformation of variables, anomalous data, selection of cases, and other techniques preceding statistical analysis.
10. Students must be capable of applying their knowledge to their work or vocation in a professional way and they should have building arguments and problem resolution skills within their area of study.
11. Students must be capable of communicating information, ideas, problems and solutions to both specialised and non-specialised audiences.
12. Students must have and understand knowledge of an area of study built on the basis of general secondary education, and while it relies on some advanced textbooks it also includes some aspects coming from the forefront of its field of study.
13. Use specific statistical software for the descriptive analysis of data.
14. Work cooperatively in a multidisciplinary context, accepting and respecting the roles of the different team members.

Content

1. Preliminaries.

1.1.The goal of Exploratory Data Analysis.
1.2. Types of variables and measurement scales.
1.3. Rounding and scientific notation.

2. Summary of statistical data.

2.1. Frequency distributions: tables.
2.2. Grouping data into intervals.
2.3. Graphical representation.

3. Numerical measures of a variable.

3.1. Central position measures: mean, median, mode.
3.2. Other position measures: quartiles, deciles and percentiles.
3.3. Measures of dispersion: variance and standard deviation (sample and population), range, interquartilerange.
3.4. Measures of relative dispersion
3.5. Standard scores.
3.6. Measures of form: symmetry and curtosis

4. Extra tools for the study of a variable.

4.1. Exploratory analysis: boxplot and other diagrams.
4.2. Transformation of variables.
4.3. Other means: geometric, harmonic, quadratic.
4.4. Chebyshev's inequality.

5. Comparison of a variable in two or more groups: Exploratory analysis.

5.1. Situation of independent samples.
5.2. Situation of paired samples.

6. Tabulation and representation of the joint distribution of two categorical variables.

6.1. Contingency tables (joint, marginal and conditional frequency distributions).
6.2. Descriptive analysis of the dependence between two categorical variables.

7. Numeric description of the joint distribution of two statistical variables.

7.1. Marginal and conditionañ measures.
7.2. Regression curves and correlation coefficient.
7.3. Linear fitting and prediction.

8. Introduction to time series.

8.1. The classical decomposition.
8.2. Smoothing series: application of filters.

*Unless the requirements enforced by the health authorities demand a prioritization or reduction of these contents.dexes.

Methodology

Classroom work, theory and problems will be complemented by computer practices where the R-band packages will be used.

In the Moodle space of the course, students will find the subject's planning, the sets of exercices and computer lab sessions, as well as possible changes of classrooms, schedules, etc. It is important to keep in mind that CampusVirtual is not a static website
but will be updated throughout the course.

In the most practical part of the course, if possible and through the analysis and comparison of statistical data, we will comment on the causes and the social and cultural mechanisms that can sustain the observed inequalities.

* The proposed teaching methodology may experience some modifications depending on the restrictions to face-to-face activities enforced by health authorities.

Annotation: Within the schedule set by the centre or degree programme, 15 minutes of one class will be reserved for students to evaluate their lecturers and their courses or modules through questionnaires.

Activities

Title	Hours	ECTS	Learning Outcomes
Type: Directed			
Computer lab	30	1.2	1, 9, 2, 3, 4, 5, 6, 7, 12, 11, 10, 14, 8, 13

Lectures	18	0.72	1, 9, 2, 3, 5, 6, 7, 12, 11, 10, 8
Problem sessions	8	0.32	9, 2, 3, 5, 6, 7, 12, 11, 10
Studying theoretical concepts, solving problems by hand and using R	84	3.36	1, 9, 2, 3, 4, 5, 6, 7, 12, 10, 14, 8, 13

Assessment

The final grade of the subject F will be obtained from:

1) The notes of the two partial exams of theory and problems, TP1 and TP2, with respective weights 20% and 25%.

2) The notes of the two computer tests, O1 and O2, with respective weights 15% and 25%.

3) The attendance to the practical sessions with computer and deliveries that are proposed, PC, with a weight of 15%. This part is not recoverable.

The final grade of the subject is obtained by making the weighted average F = 0.15 TP1 + 0.2 O1 + 0.25 TP2 + 0.25 O2 + 0.15 PC.

A requirement to pass the subject with the previous formula is that the marks TP1, TP2 and O2 must be greater than or equal to 4 and O1 must be greater or equal to 3.5.

In case F<5 or the requirement (marks O2, TP1 and TP2 >=4 and mark O1 >= 3.5) is not satisfied the students will have the opportunity to take a resit test,

There will be two resit tests:

- STP a global exam oftheory and problems teoria i problemes, for students with marks of TP1 or TP2 less than 4 or whose bad grades in these exams cause F<5.

- SO a global exam with computer, for students with marks O1 less than 3.5 or O2 less than 4, or whose bad grades in these exams give F<5.

the final grade will be F=0,45 STP + 0,40 SO + 0,15 PC

(if only one of the tests are taken (STP or SO) then the other grade will be taken from the weighted mean of the passed partial exams grades).

Students who do not attend exam will get the qualification of "Not Evaluable".

"Without prejudice to other disciplinary measures deemed appropriate, and in accordance with current Academic regulations the irregularities committed by the student that may lead to a variation of the grade of an evaluation act will be scored with a zero. so, plagiarizing, copying or letting copy a practice or any other evaluation activity will involve suspending with a zero and cannot be recovered in the academic mateixcurs.If this activity has a minimum associated score, then the subject will be suspended. "

After the second partial tests the honors qualitfications will be considered and might be given, even before the resit exam.

*Student's assessment may experience some modifications depending on the restrictions to face-to-face activities enforced by health authorities.

Assessment Activities

Title	Weighting	Hours	ECTS	Learning Outcomes
End of term written exam	25%	2	0.08	1, 2, 3, 4, 5, 6, 7, 12, 11, 10, 14, 8, 13
First computer lab exam	20%	2	0.08	1, 9, 2, 3, 4, 5, 6, 7, 11, 10, 8, 13
Mid-term written exam	20%	2	0.08	2, 3, 5, 6, 7, 12, 11, 10
Second computer lab exam	25%	2	0.08	1, 9, 2, 3, 4, 5, 6, 7, 11, 10, 8, 13
Submission of exercise sets done with computer	10%	2	0.08	1, 9, 2, 3, 4, 5, 6, 7, 11, 10, 14, 8, 13

Bibliography

Course lecture notes

X. BARDINA, M. FARRÉ, Estadística descriptiva, Manuals, 54 Servei de Publicacions, UAB

Bibliography

A.J.B. ANDERSON, Interpreting Data. A first cours in Statistics, Ed Chapman and Hall, 1989.
R Tutorial. An R introduction to statistics. (2016). www.r-tutor.com
E. CASA ARUTA, Problemas de Estadística Descriptiva, Ed. Vicens Vives.
R. JOHNSON, P. KUBY, Estadística elemental: Lo esencial, Ed Thomson, 1999.
B. PY, Statistique Descriptive, Ed Económica, 1988.
M. SPIEGEL, Estadística, Teoría y 875 problemas resueltos, Schaum-McGraw-Hill, 1990.
V. ZAIATS, M.L. CALLE i R. PRESAS, Probabilitat i Estadística. Exercicis I, Eumo Ed, 1998.

Complementary Bibliography

G. CALOT, Curso de Estadística Descriptiva. Ed Paraninfo, 1988.
FERNÁNDEZ, J.M. CORDERO, A. C\'ORDOBA, Estadística Descriptiva, ed ESIC 1996.
L.C HAMMILTON, Modern Data Analysis, Brooks/Cole Publishing Company, 1990.
P.G. HOEL i R.J. JESSEN, Estadística básica para negocios y economía, Compañía Editorial Continental,Mexico, 1993.
R.K. PEARSON, Exploratory Data Analysis using R. Data Mining and Knowledge Discovery Series, Chapman & Hall/CRC, 2018.
D. PEÑA SÁNCHEZ DE RIVERA, Estadística. Modelos y métodos. 1. Fundamentos i 2. Modelos lineales yseries temporales, Alianza Editorial 1995. (2 volums)

Software

R and RStudio