INSTITUTO POLITÉCNICO DE BRAGANÇA

Course Unit	Exploratory Data Analysis			Field of study	Mathematics		
Bachelor in	Electrical and Computers Engineering			School	School of Technology and Management		
Academic Year	2022/2023	Year of study	2	Level	1-2	ECTS credits	6.0
Type	Semestral	Semester	1	Code	9112-742-2101-00-22		

Workload (hours)	162	Contact hours	T	TP	PL	TC	S	E	OT	O
			-	30	30	-	-	-	-	-

T - Lectures; TP - Lectures and problem-solving; PL - Problem-solving, project or laboratory; TC - Fieldwork; S - Seminar; E - Placement; OT - Tutorial; O - Other

Name(s) of lecturer(s): Maria Clara Rodrigues Bento Vaz Fernandes

Learning outcomes and competences

At the end of the course unit the learner is expected to be able to:
1. Understand the probability concept and its axioms and compute simple probability using the basic probability and combinatorics laws;
2. Understand the basic concepts related to random variables and perform simple calculations based on them;
3. Know and to manipulate, at a basic level, the most common random variables and use them to model simple situations;
4. Apply and to interpret the most common ways of representing and synthesize the information in a dataset;
5. Group data according to different similarity measures (cluster analysis);
6. Compute point and interval estimates for the most common population parameters;
7. Understand and to apply the hypothesis test methodology on the most common population parameters;
8. Apply the non-parametric Chi-Square test in the analysis of qualitative data.

Prerequisites

Before the course unit the learner is expected to be able to:
1. Manipulate basic mathematical concepts
2. Use a high-level programming language

Course contents

Introduction. Elementary Probability Theory. Random variables and probability distributions. Random sampling and sample distributions. Pre-processing, organization and representation of data. Classification and data grouping (Cluster Analysis). Point estimation and Interval estimation. Parametric and non-parametric hypothesis tests.

Course contents (extended version)

1. Introduction
   - The statistical object;
   - Descriptive statistics and statistical inference;
   - Populations and samples.
2. Basic Probability Theory
   - Random experiments, sampling spaces and events;
   - Combinatorics;
   - The probability concept;
   - Conditional probability;
   - Independent events;
   - Bayes' theorem.
3. Random Variables and Probability Distributions
   - Definition of random variable;
   - Discrete variables: probability function and cumulative probability function;
   - Continuous variables: probability density function and cumulative density function;
   - Distribution parameters;
   - Variable transformations: linear and non linear transformations.
4. Joint Probability Distributions
   - Definition of joint distributions;
   - Marginal distributions;
   - Conditional distributions;
   - Independence;
   - Covariance and correlation;
   - Variable combinations;
   - Calculation of a combination expected value and variance: linear and non linear combinations.
5. Characterization of Some Discrete Distributions
   - Binomial distribution;
   - Negative Binomial distribution;
   - Hypergeometric distribution;
   - Relations between the Binomial and the Hypergeometric distributions;
   - Poisson distribution;
   - Relations between the Poisson and the Binomial and Hypergeometric distributions.
6. Characterization of Some Continuous Distributions
   - Uniform distribution;
   - Exponential distribution;
   - Relations between the Poisson and Exponential distributions;
   - Normal distribution;
   - Standard Normal distribution;
   - Linear combination of independent Normal variables;
   - Relations between the Normal and Binomial distributions;
   - Chi-squared, Student's t and F distributions.
7. Random Sampling and Sampling Distributions
   - Distribution of sample mean;
   - Expected value and variance for sample mean;
   - Sample mean distribution shape for Normal populations;
   - Central limit theorem;
   - Implications on sampling distributions;
   - Random sample generations using Monte Carlo method.
8. Descriptive Statistics
   - Data classification;
   - Qualitative and quantitative data;
   - Univariate samples characterization;
   - Location statistics (average, median and mode) e dispersion (variance);

Course contents (extended version)

- Skew and Kurtosis coefficients;
- Bivariate samples characterization;
- Calculation of a linear relation coefficients using least squares;
- Correlation and determination coefficients.

9. Cluster Analysis
   - Objective of cluster analysis
   - Similarity measures;
   - Hierarchical clustering;
   - Non-hierarchical clustering.
10. Point Estimation
   - Estimators and estimates;
   - Desirable estimator properties: unbiased, efficient and consistent;
   - Estimation methods: maximum likelihood and least squares;
   - Estimator selection.
11. Interval Estimation
   - Confidence interval concept;
   - Confidence interval specification;
   - Confidence intervals for the continuous populations mean;
   - Confidence intervals for the Binomial proportion: small and large samples;
   - Confidence intervals for a Normal population variance;
   - Confidence intervals for the mean difference between two continuous populations;
   - Confidence intervals for the proportion difference with large samples;
   - Confidence intervals for the variance ratio of Normal populations;
   - Sample size determination.
12. Hypothesis Tests
   - Basic hypothesis test procedure;
   - Hypothesis definition;
   - Test statistic characterization;
   - Decision rule definition;
   - Type I error and significance;
   - Calculation of test statistics and decision making;
   - P-value;
   - Type II error and statistical power;
   - Relations between hypothesis tests and confidence intervals;
   - Reference to the most common tests.
13. Non Parametric Tests
   - Classification of the most frequently used tests;
   - Chi-Square test for categorical data.

Recommended reading

1. Rui Guimarães, Sarsfield Cabral - Estatística - 2ª edição, Verlag Dashofer, 2010.
2. William Navidi - Statistics for Engineers & Scientists - 5th edition, McGraw-Hill Education, 2020.
3. Alberto Leon-Garcia - Probability, Statistics, and Random Processes for Electrical Engineering - 3rd edition, Pearson Prentice Hall, 2008.
4. Ronald K. Pearson - Exploratory Data Analysis Using R - Chapman and Hall/CRC, 2020.
5. Subhash Sharma - Applied Multivariate Techniques - John Wiley & Sons, 1996.

Teaching and learning methods

The contents will be presented in the classroom, in a theoretical-practical regime, accompanied by the resolution of exercises. In non-classroom hours the topics will be explored through application exercises using computer tools, to be developed in R. Individual or group tutorial sessions may be held in non-classroom hours, whenever necessary.

Assessment methods

1. Final Exam - (Regular, Student Worker) (Final, Supplementary, Special)
   - Final Written Exam - 100%
2. Alternative 2 - (Regular, Student Worker) (Final)
   - Intermediate Written Test - 40% (To be held during classes.)
   - Final Written Exam - 40% (To be held at the end of the semester.)
   - Practical Work - 20%

Language of instruction

Portuguese, with additional English support for foreign students.

Electronic validation

Maria Clara Rodrigues Bento Vaz Fernandes	António Jorge da Silva Trindade Duarte	Orlando Manuel de Castro Ferreira Soares	Paulo Alexandre Vara Alves
07-10-2022	11-10-2022	21-10-2022	03-11-2022