Course guide
270210 - PIE2 - Probability and Statistics 2

Last modified: 10/07/2025

Unit in charge: Barcelona School of Informatics
Teaching unit: 715 - EIO - Department of Statistics and Operations Research.

Degree: BACHELOR'S DEGREE IN DATA SCIENCE AND ENGINEERING (Syllabus 2017). (Compulsory subject).

Academic year: 2025 **ECTS Credits:** 6.0 **Languages:** Catalan

LECTURER

Coordinating lecturer: MARTA PÉREZ CASANY

Others:

PRIOR SKILLS

To follow this subject, the student needs to have a good understanding of the previous subjects entitled: PiE1 and Calcul.

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.

Generical:
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.

Transversal:
CT5. Solvent use of information resources. Manage the acquisition, structuring, analysis and visualization of data and information in the field of specialty and critically evaluate the results of such management.
CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.

Basic:
CB1. That students have demonstrated to possess and understand knowledge in an area of ??study that starts from the base of general secondary education, and is usually found at a level that, although supported by advanced textbooks, also includes some aspects that imply Knowledge from the vanguard of their field of study.
CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy

TEACHING METHODOLOGY

One half of the sessions will consist on the exposition of new concepts and contents. The other half will be devoted to solve practical exercises. At the end of each practical session, some exercises will be proposed to the students in order that they can work i autonomously.

LEARNING OBJECTIVES OF THE SUBJECT

1.To learn how to contruct statistical models in order to sinthesize information, explain a response variable as a function of some explanatory variables, and do forecasting.
2.To understand the basic concepts and the philosophy behind Bayesian statistics.
3.To learn software statistics and how to use it to analyze real data
4.To learn model validation techniques.
5.To learn how to do a report containign the results of a data analysis
6.To understand the difference between the Bayesian and frequentist statistics
7.To know which is the most suitable modalization technique for each problem.
8.To learn how to interpret the results of a fitted model
9.To understand the concept of crossvalidation and the ones of overfitting and underfitting
10.To use the model fitted for predictions
11.To understand the difference between parameter and parameter estimation, to solve inference problems in linear and generalized linear models.
12.To learn how to include cathegorical variables in linear and generalized linear models.
13.To analyze with critical sense, data and topics relevant for the society
14.To perform estimation usign confidence intervals
15.To understand the importance of the hypothesis testing. To know how to perform the classical hypothesis tests and to know techniques to face new hypothesis test that can appear doing research.

STUDY LOAD

Type	Hours	Percentage
Hours large group	30,0	20.00
Self study	90,0	60.00
Hours small group	30,0	20.00

Total learning time: 150 h

CONTENTS

Distributions related to the Normal distribution. Confidence Interval Estimation.

Description:
Distribucions Chiquadrat, t-d'Student i F-Fisher-Snedecor. Definició d'intèrval de confiança. IC per un valor esperat, per una variància, per una probabilitat i per la diferència de dos valors esperats i dues probabilitats. Quantitats pivotals.

Hypothesis testing

Description:
Conceptes generals en l'entorn dels test d'hipòtesis. Comparcions d'una esperança i una variància amb un valor concret. Comparació de dos valors esperats, comparació de dues variàncies. Comparació d'una probabilitat amb un valor concret. Comparació de dues probabilitats.

Linear model

Description:
Linear model definition. Parameter estimation. Anova Table and goodness-of-fit measures. Inference. Prediction. Model validation. Model selection. Model interpretation. Bias, colliniarity. causality. Use of cathegorical explanatory variables.

Generalized linear model

Description:
Definition of generalized linear model. Models for binary response. Parameter estimation. Inference. Model validation. Model selection. prediction. Model interpretation.

Introduction to Bayessian statistics.

Description:
Bayes theorem. Bayessian model. Predictive distribution a priori and a posteriori. Selection of the a priori distribution.

ACTIVITIES

Linear models

Description:
Linear model definition. Estimation and inference in a linear model. Model validation. Model selection. Model interpretation. Linear models with categorical variables. Non linear models with gaussian response.

Specific objectives:
1, 3, 4, 5, 7, 8, 10, 11, 12, 13

Related competencies :
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy
CB1. That students have demonstrated to possess and understand knowledge in an area of ??study that starts from the base of general secondary education, and is usually found at a level that, although supported by advanced textbooks, also includes some aspects that imply Knowledge from the vanguard of their field of study.
CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.
CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CT5. Solvent use of information resources. Manage the acquisition, structuring, analysis and visualization of data and information in the field of specialty and critically evaluate the results of such management.
CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.

Full-or-part-time: 59h
Self study: 35h
Theory classes: 12h
Laboratory classes: 12h

Generalized linear model.

Description:
Generalized linear model definition. Models for binary response. Estimation and inference in a generalized linear model. Prediction, interpretation and model selection.

Specific objectives:
1, 3, 4, 5, 7, 8, 9, 10, 11, 12, 13

Related competencies :
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy
CB1. That students have demonstrated to possess and understand knowledge in an area of ??study that starts from the base of general secondary education, and is usually found at a level that, although supported by advanced textbooks, also includes some aspects that imply Knowledge from the vanguard of their field of study.
CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.
CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CT5. Solvent use of information resources. Manage the acquisition, structuring, analysis and visualization of data and information in the field of specialty and critically evaluate the results of such management.
CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.

Full-or-part-time: 27h
Self study: 15h
Theory classes: 6h
Laboratory classes: 6h

Bayesian statistics

Description:
Statistical model. Inference based on liekelihood. Bayesian model. "A posteriori" distribution. Predictive distribution "a priori" and "a posteriori". How to select the "a priori distribution". Bayesian inference. Model validation.

Specific objectives:
2, 6

Related competencies :
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy
CB1. That students have demonstrated to possess and understand knowledge in an area of ??study that starts from the base of general secondary education, and is usually found at a level that, although supported by advanced textbooks, also includes some aspects that imply Knowledge from the vanguard of their field of study.
CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CT5. Solvent use of information resources. Manage the acquisition, structuring, analysis and visualization of data and information in the field of specialty and critically evaluate the results of such management.
CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.

Full-or-part-time: 14h
Self study: 10h
Theory classes: 2h
Laboratory classes: 2h

Distributions related to the Normal distribution. Confidence interval estimation

Description:
The distributions chi-square, Student-t and Fisher are defined. The concept of Confidence interval and pivotal quantity are introduced. The most important and useful confidence intervals are computed.

Specific objectives:
13, 14

Related competencies :
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy
CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CT5. Solvent use of information resources. Manage the acquisition, structuring, analysis and visualization of data and information in the field of specialty and critically evaluate the results of such management.
CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.

Full-or-part-time: 23h
Self study: 15h
Theory classes: 4h
Laboratory classes: 4h

Hypothesis test

Description:
The basic concepts related to hypothesis test are introduced. The hypothesis for comparing one mean and one variance to a given value, to compare two means, two variances and to probabilities are shown.

Specific objectives:
5, 13, 15

Related competencies :
CB5. That the students have developed those learning skills necessary to undertake later studies with a high degree of autonomy
CB3. That students have the ability to gather and interpret relevant data (usually within their area of ??study) to make judgments that include a reflection on relevant social, scientific or ethical issues.
CE3. Analyze complex phenomena through probability and statistics, and propose models of these types in specific situations. Formulate and solve mathematical optimization problems.
CG2. Choose and apply the most appropriate methods and techniques to a problem defined by data that represents a challenge for its volume, speed, variety or heterogeneity, including computer, mathematical, statistical and signal processing methods.
CG1. To design computer systems that integrate data of provenances and very diverse forms, create with them mathematical models, reason on these models and act accordingly, learning from experience.
CT5. Solvent use of information resources. Manage the acquisition, structuring, analysis and visualization of data and information in the field of specialty and critically evaluate the results of such management.
CT6. Autonomous Learning. Detect deficiencies in one's own knowledge and overcome them through critical reflection and the choice of the best action to extend this knowledge.

Full-or-part-time: 27h
Self study: 15h
Theory classes: 6h
Laboratory classes: 6h

GRADING SYSTEM

There will be a partial exam and a final exam, as well as exercises of data analysis assigned during the course.

The partical exam will correspond to the confidence intervals and hypothesis tests.

The final exam will correspond to the rest of the subject contents.

The course mark will be the sample mean of the exercices realized during the course.

The final mark will be computed as:

Subject Mark = 0.25 * Cours+ 0.25 * Partial+ 0.5 *FinalExam

In the case of the students that go to the reevaluation, the final mark will be computed like this:

Subject Mark=max(Reevaluation, 0,25Cours+0,75Reevaluation)

BIBLIOGRAPHY

Basic:
- James, G.; Witten, D.; Hastie, T.; Tibshirani, R. An introduction to statistical learning. Springer, 2013. ISBN 97-1461471370.
- Gelman, A.; Carlin, J.B.; Stern, H.S.; Dunson, D.B.; Vehtari, A.; Rubin, D.. Bayesian data analysis. 3rd ed. Chapman & Hall, 2014. ISBN 978-1439840955.
- Weisberg, S. Applied linear regression [on line]. 4th ed. John Wiley and Sons, 2014 [Consultation: 15/09/2025]. Available on: https://ebookcentral-proquest-com.recursos.biblioteca.upc.edu/lib/upcatalunya-ebooks/detail.action?pq-origsite=primo&docID=1574352. ISBN 9781118594858.
- Dobson, A.J.; Barnett, A.G. An introduction to generalized linear models. 4th ed. Chapman & Hall, 2018. ISBN 978-1138741515.
- Dobrow, R.P. Introduction to stochastic processes with R. John Wiley and Sons, 2016. ISBN 978-1118740651.