Course guides  
200613 - ADD - Discrete Data Analysis

Unit in charge: School of Mathematics and Statistics
Teaching unit: 749 - MAT - Department of Mathematics.
Degree: MASTER'S DEGREE IN STATISTICS AND OPERATIONS RESEARCH (Syllabus 2013). (Optional subject).
Academic year: 2015  
ECTS Credits: 5.0  
Languages: English

LECTURER

Coordinating lecturer: MARTA PÉREZ CASANY
Others: MARTA PÉREZ CASANY - A

PRIOR SKILLS

Probability and Statistics: the student should know the basic concepts around probability models, the maximum likelihood and moment methods for performing the parameter estimation, and the basic concepts around hypothesis testing.

Mathematical Analysis: The student should know the concepts related to the real series of positive values and to the one-variable real functions.

Modelization: It is not mandatory but it is convinient that the student knows well the concept of linear model.

REQUIREMENTS

The students should have attend and pass, at least, a basic cours on Probability and Statistics and one in Mathematical Analysis.

DEGREE COMPETENCES TO WHICH THE SUBJECT CONTRIBUTES

Specific:
3. CE-1. Ability to design and manage the collection of information and coding, handling, storing and processing it.
4. CE-3. Ability to formulate, analyze and validate models applicable to practical problems. Ability to select the method and / or statistical or operations research technique more appropriate to apply this model to the situation or problem.
5. CE-4. Ability to use different inference procedures to answer questions, identifying the properties of different estimation methods and their advantages and disadvantages, tailored to a specific situation and a specific context.
6. CE-6. Ability to use appropriate software to perform the necessary calculations in solving a problem.
7. CE-7. Ability to understand statistical and operations research papers of an advanced level. Know the research procedures for both the production of new knowledge and its transmission.
8. CE-8. Ability to discuss the validity, scope and relevance of these solutions and be able to present and defend their conclusions.

Transversal:
1. TEAMWORK: Being able to work in an interdisciplinary team, whether as a member or as a leader, with the aim of contributing to projects pragmatically and responsibly and making commitments in view of the resources that are available.

2. FOREIGN LANGUAGE: Achieving a level of spoken and written proficiency in a foreign language, preferably English, that meets the needs of the profession and the labour market.
TEACHING METHODOLOGY

During the lectures, we will explain the theoretical concepts related to each unit, and propose several exercises that will be solved partly in class and at home. Some of the exercises will consist in searching additional information related to the concepts introduced in class. Every two weeks, the students will have to read a research paper related to the concepts studied. The paper will be later discussed in class. Some of the papers have a clear practical point of view since they are published in areas different from statistics and probability. Others are more methodological. During the course, the students will have three practical sessions using the statistical Package R.

LEARNING OBJECTIVES OF THE SUBJECT

The main objective of the course is to make sure that the student knows which are the main characteristics of discrete real data, and which are the models most used in practice to fit those data.

STUDY LOAD

<table>
<thead>
<tr>
<th>Type</th>
<th>Hours</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>Self study</td>
<td>80,0</td>
<td>64.00</td>
</tr>
<tr>
<td>Hours small group</td>
<td>15,0</td>
<td>12.00</td>
</tr>
<tr>
<td>Hours large group</td>
<td>30,0</td>
<td>24.00</td>
</tr>
</tbody>
</table>

Total learning time: 125 h

CONTENTS

Discrete probability distributions

Description:
This unit is devoted to the introduction of the most classical discrete probability distributions. These are: Bernoulli, Binomial, Multinomial, Geometric, Hypergeometric, Poisson, and logarithmic series distributions. Several hypothesis tests for the comparison of two probabilities in the binomial model will be explained, and the importance of the Hypergeometric distribution in the capture and recapture models will be established. The discrete Pareto or Zipfian distribution has gained importance recently, as a consequence of being the distribution of the degree of a node in a social, economic or protein networks, between others. The importance of this distribution for modelling ranked data as well as data corresponding to frequencies of frequencies will be set up.

Mixture or compound distributions (MD)

Description:
The need to consider a MD is a consequence of the excess of variance, due to the lack of homogeneity of the sampled population. Mixtures with a discrete as well as a continuous mixing distribution will be considered. The main results related to MD are explained. The zero-modified binomial and Poisson, the beta-binomial, the Negative Binomial, the IG-Poisson and the Sichel distribution are going to be considered. The areas of research where they have proved to be useful are established.

Poisson-stopped-sum distributions (PSS)

Description:
It will be seen how the PSS distributions come in practice. The more important discrete PSS will be introduced: Neyman A, Poisson-Pascal, Negative Binomial, Lagrangian Poisson, and Generalized-Inverse-Gaussian distribution. The conditions under which a PSS is a MD and a MD is a PSS are described.
**Discrete distributions in models will covariates**

**Description:**
The logit, probit and complementary log-log models for binary response, and the log-linear models for Poisson response will be studied. Parameter interpretation and hypothesis testing for the parameter significance will be seen. The relation between the Poisson and the multinomial distributions will be used to analyze a contingency table from two different points of view. Models with covariates with a response random variable that is Zipf distributed will also be considered.

**GRADING SYSTEM**

40% of the final mark will correspond to the qualification obtained in the final exam. Another 40% will come from the arithmetic mean of the two marks obtained in the report papers presented and the mark of the oral presentation. The 20% that remains will be decided by the teacher based on the degree of participation of the student in class, and the degree of accomplishment of the homework proposed.

**BIBLIOGRAPHY**

**Basic:**