BASICS OF STATISTICS
Óscar Rivero Salgado

Contents

1 Module description

One possible description for Statistics would be as the discipline whose purpose is to "use empirical data generated by a random phenomenon in order to make inferences about some deterministic characteristics of the phenomenon, while simultaneously quantifying the uncertainty inherent in these interferences". Statistics also tries to give an answer to concrete problems that occur in the real world, for instance: determine the (mean) cholesterol level of all the inhabitants of Barcelona; guess the efficiency of a vaccine against AIDS; determine the quality of a bunch of products; determine the opinion with respect to a certain politician... Moreover, statistics is nowadays a very important tool in new disciplines such as big data, machine learning and so on.

However, the study of statistics is unavoidable linked with the study of probability and basically, a good understanding of statistics needs first a good knowledge of probability theory. Roughly speaking, what probability does is to make deductions over a population to conclude some fact about a sample; for instance, given 75 green balls and 25 red balls, we want to determine the probability of having 3 green balls in a sample of 5 elements. Statistics makes the opposite process; given a sample of 5 balls in which 3 are green, we want to do an induction process to determine the total number of green balls in the population, and have of course a measure of the uncertainty.

This course tries to be an invitation to statistics, and for that, one third of the course will be devoted to probability, to have the basis one needs to carry on the complex study of statistics. We have divided the course into 4 curricular blocks. The first and the third are about probability and are shorter since the students have taken previous courses more related with this: we begin with a review of combinatorics to deal with the case of discrete variables; then, we move to the more sophisticated case of continuous variables, although it will be presented from an intuitive perspective without going into the more technical details. The second block will be devoted to descriptive statistics, analyzing different ways of representing data and also seeing how we can quantify the central tendency and the dispersion for a given sample. Finally, the last block is devoted to statistical inference, whose aim is to estimate a population parameter either with a single value computed from a sample (point estimation) or with a range of possible values also computed from the sample (interval estimation). We will finish this block with a study of hypothesis testing.

2 Module outline

1. FIRST BLOCK: DISCRETE PROBABILITY.

   - **Day 1: The art of counting.** We present the course with some motivational and classical examples. Then, we review some of the techniques the students may have learned about binomial numbers, permutations, variations, combinations (with and without repetition), multinomial numbers, double counting methods, inclusion-exclusion principle... We will focus on exercises analyzing different kinds of situations emphasizing the applications to statistics.
   - **Day 2: Basics of probability.** In this class we recall the language of probability theory, giving first the intuitive point of view and passing then to the most axiomatic approach. We will avoid a very abstract treatment going through examples from the real world. We introduce random the concept of independence and conditioned probability (Bayes formula)
   - **Day 3: Discrete random variables and discrete probability.** We begin by introducing several examples of random variables (geometric, binomial...). Then, we move to the introduction of parameters such as the expectation, the variance and in general higher order momenta. We will also see applications of the linearity of expectation. These concepts will be present in several moments of the course, so we will be working on them continuously.
   - **Day 4 and day 5** will be devoted to both solving problems and discussing these concepts, as well as doing the corresponding exam.

2. SECOND BLOCK: DESCRIPTIVE STATISTICS.

   - **Day 6: Univariate data analysis.** The aim of this first lecture in statistics is to make the students familiar with some technical vocabulary and also to present some motivations from the real world. Then, we present what is called descriptive analysis, the part of statistics that count, organize, summarize and graph data to describe their main features. We give examples of frequency tables and we focus on the graphical presentation of data: bar charts, pie charts, histograms...
   - **Day 7: Quantitative data analysis.** We will continue with the topics of the previous lecture introducing also quantitative measures of both the central tendency (mean, median, quantile) and the variation (range, standard deviation). We give examples with Excel and show how useful it can be for the treatment of data.

- **Day 8: Bivariate data analysis.** Following the approach we have done, now we move to bivariate data sets, collections of pairs of values that result of the jointly observation of two characteristics X, Y of a population. We speak about marginals, covariance and correlation, introducing the technique of least squares method. We will discuss several examples where this is used (also using Excel).
- **Day 9: Random variables and the normal distribution.** We will begin giving an intuitive view of the concepts of probability density function and distribution function, using the intuition they will have from the discrete case. Then, we will revisit some of the discrete random variables we have worked with, to motivate the presentation of the principal continuous random variables: uniform, exponential, normal... At the end, we state the central limit theorem and the law of large numbers, to explain some remarkable applications in statistics.
- **Day 10** will be devoted to both solving problems and discussing these concepts, as well as doing the corresponding exam.

3. THIRD BLOCK: STATISTICAL INFERENCE.

    - **Day 11: What is statistical inference?** We will introduce this process which consists on deducing properties of an underlying distribution by analysis of data, with the goal of getting information on the population. We discuss the techniques of sampling and talk about what parameter estimation represents, and what are the main mathematical ingredients involved in the process.
    - **Day 12: Interval estimation.** We mimic the discussion of point estimation in the setting where we must give a range of possible values for the parameters with a control over the error. We will introduce for that Student's t-distribution.
    - **Day 13: Hypothesis testing.** We will focus first on the parametric case, dealing with tests about a population means, or for the difference between means of two populations. We will discuss the concept of p-value and make a short analysis of variance.
    - **Day 14: Tests of goodness of fit and independence.** We now move to tests in which we want to compare proportions of two or more populations. Concepts such as contingency tables will arise. At the end, we will recover the least square method and apply some of the previous methods for a deeper study of it.

4. **Day 15: Summary of the course and final exam.**

3 ORGANIZATION OF THE COURSE

The lecturer will provide some notes of the course (although some of the proofs and examples that will be covered in the blackboard will not be included there with all detail) as well as an exercise list. Some of the exercises will be solved in class and some others will be left just for practice and self-assessment Further, there is the "Assignment list", that are the problems the students must deliver.

Some of the exercises will require some numerical computations and we will use Excel to deal with them, that will be presented as a very basic tool for data analysis, with many functionalities.
There will be two tests at the end of week 1 and one at the end of the week 2 (one hour duration). At the end of the course there will be a final exam of two hours. In addition, 4 or 5 short quizzes (20-30 minutes) will be proposed in random days.

4 GRADING

- There will be 9 homework assignments; the homework mark will be the average of all of them, dropping the two lowest scores. This will be a 30% of the final qualification.
- The two tests that will be done at the end of week 1 and week 2 will have a weight of a 10% each.
- Along the course, there will be 4 (very short) quizzes some of the days. The average of all of them, dropping the lowest score, will represent a 10% of the final qualification.
- The final exam represents the remaining 40% of the qualification. It will consist on a short questions and three problems.

5 TECHNICAL REQUIREMENTS

Some classes will be developed in the blackboard, as it usually occurs in mathematical courses. In some other, the more descriptive ones, I will use the screen and also the computer. Excel will be used along the course, both for the classes and for the assignments.

6 PREREQUISITES

The course will be adapted to the level of the students. Familiarity with basic counting procedures, as well as with the basis of probability would be nice, but not indispensable at all.

7 READING LIST

7.1 Basic texts

- Grimmett, G.R.; Stirzaker, D.R. Probability and randomm processes. Oxford University Press.
- Statistics for Mathematicians. Panaretos, Victor. Birkhäuser.
- De Groot, M.H.; Schervish, M.J. Probability and statistics. Pearson.
- More references will be added.

7.2 Complementary texts

- Dalgaard, P. Introductory statistics with R. Springer.
- Peck, R. et al. Statistics: a guide to the unknown. Duxbury Resource Center.
- Bartle, RG. The elements of integration and Lebesgue measure.
- Salsburg, D. The lady tasting tea (dissemination).