Statistics with R

Marco Bittelli and Martina Zappaterra
Department of Agriculture and Food Sciences, University of Bologna, Italy

Preface

The notes collected here have been written over a few years as a support material for classes in Statistics and Experimental Methodology that we teach at the University of Bologna. The notes and the R computer code is original material that we wrote, while the experimental data were collected during experiments that we performed in collaboration with colleagues and that have been published and cited in the notes. Some data have been given to us as teaching examples. We are grateful to Livia Antisari for sharing soil data from her experiments in the Emilia-Romagna Appennines, Aldo Gardini e Carlo Trivisano for code used to analyze the soil data, and Gabriele Antolini for Emilia-Romagna weather data. We also would like to thank Roberto Olmi for computer code and notes.

Contents

1 Introduction 1
- 1.1 Introduction 1

2 RStudio 3

3 Data types and operators 4
- 3.1 Vector 5
- 3.2 Matrix 11
- 3.3 Array 13
- 3.4 List 14
- 3.5 Data Frame 14

4 Conditional Statements and Loops 16
- 4.1 If statement 16
- 4.2 Else If statement 16
- 4.3 Else statement 17
- 4.4 For loop 17
- 4.5 While loop 19

5 Managing Data 20
- 5.1 Qualitative and quantitative data 20
- 5.2 Import Data 22
- 5.3 Open a Data File 22
- 5.4 Managing Dates 25
- 5.5 Subsetting 33
- 5.6 Merging 34

6 Data Visualization 36
- 6.1 Create graphs using plot 36
- 6.2 Create graphs using ggplot 39

7 Probability 47
- 7.1 Definition 47
- 7.2 Sample Space and Events 48
- 7.3 Conditional probability and Bayes Theorem 55
- 7.4 Probability and determinism: the Buffon's needle 68
- 7.5 Probability Distribution 71
- 7.6 Exercises 75

8 Distributions of Random Variables 76
- 8.1 Random variables 76

8.2 Distributions 76
8.3 Uniform distribution 77
8.4 Bernoulli distribution 79
8.5 Binomial distribution 81
8.6 Normal Distribution 83
8.7 t-student distribution 88
8.8 Poisson distribution 90

9 Descriptive statistics 96
9.1 Frequencies 96
9.2 Classes 101
9.3 Cumulative curves and frequencies 107
9.4 Measures of Central Tendency 113
9.5 Measures of Variability 116
9.6 Coefficient of variation 117
9.7 Quantiles 117
9.8 The box plot 120
9.9 Exercises 120

10 Inferential statistics 121
10.1 Population and sample 121
10.2 Deriving the mean and variance of different random variables 122
10.3 Confidence Intervals 128
10.4 Hyphotesis tests 135
10.5 Example 137

11 Linear Models 140
11.1 Fitting a line to data 140
11.2 Least squares regression 145
11.3 Linear models with multiple parameters 148

12 Experimental Designs 155
12.1 Completely randomized design 157
12.2 Randomized block design 158
12.3 Factorial experimental design 158
12.4 Latin square design 159
12.5 Nested design 159
12.6 Fixed or random variables: which model should I use? 161

13 Analysis of Variance 162
13.1 General description 162
13.2 Example 163
13.3 One way ANOVA 166
13.4 Test for normality 166
13.5 Example: Soil properties and Land Cover 166
13.6 Transformations 172
13.7 Kruskal-Wallis 173
13.8 Tukey test 175

13.9 Two way ANOVA 175
13.10 Tukey test 175

1
Introduction

1.1 Introduction

In the last four decades there have been an active development of analytical tools in statistics, based on the power of computer technology and computation. Today, the integration of classical theory and computational tools is an important component of current teaching curricula. It is very effective to teach statistics using programming languages that allows for direct application of concept to read world data, from the beginning of the learning process. There are a variety of good commercial programs for applications in statistics, including MatLab and others. However, the open source R programming language is the most popular one.

The goal of these notes is to help the student learn the most important tools in R that will allow for statistics and data analysis. The notes are divided in two main parts.

The first part of the notes will introduce the use of R and RStudio (Chapter 2) and fundamental concepts of data management programming that are necessary to understand the program and utilize it (Chapters 3-6). Concepts of *data types* and *operators* are provided, along with *conditional statements* and *loops*. In particular data requires *importing data*. Importing data may be more complicated than expected since the data may present *missing data*, *time series* may be organized with specific data formatting that should be properly read and so forth. After successfully importing the data, the second step is *visualization*. A good visualization will show the data in ways that the student may not expect or raise new questions about the data. These notes will explain the different visualization procedures, the type of graphs and other visualization options.

The second part is concerned with statistical analysis. This part presents an introduction to the most common one taught as preliminary classes in college statistics. Chapter 7 introduces basic notions of *probability*, including basic information on set theory, definition of probability, the law of large numbers and conditional probability. The chapter also introduces the Bayes theoream with examples of Bayesian statistics.

Chapter 8 describes the concept of continuous and discrete *random variables and distributions*. This chapter describes the bimodal, normal, t-student, geometric, chi-square, logistic and Poisson distributions.

Chapter 9 enters into statistical methods with principles of descriptive statistics. The chapter begins with descriptive statistics such as the construction of *classes*, *cumulative curves* and data visualization tools. Then it describes measures of *central tendency* and *variability* with examples in R on real data. Concepts of *quantiles* and *box plots* are provided.

Chapter 10 introduces inferential statistics with concepts of populations, samples and inferential methods. Chapter 11 introduces the concept of linear models, linear regression and correlation. Chapter 12 discusses the most common experimental designs, which are cornerstone for a good experiment and therefore a solid statistical analysis.

Chapter 13 describes the analysis of variance (ANOVA) with some simple applications and further examples on more complex data analysis.

Chapter 14 describes multivariate statistics including covariance matrices, correlation matrices, distances, principal component analysis, cluster analysis and other common techniques, while Chapters 14 non−linear optimization regression procedures and least squares concepts applied to non linear models.