

Contents

Preface	9
1. Why statistics?	13
1.1. Introduction: the statistical method	13
1.2. Data collection	14
1.3. Populations and samples	15
1.3.1. <i>A brief overview of more advanced sampling schemes</i>	16
1.3.2. <i>Potential bias in sampling</i>	17
1.4. Preparing and organizing data	18
1.4.1. <i>Some key definitions</i>	18
1.5. Types of variables	20
1.6. Parameters and statistics: a fundamental distinction	21
1.7. Statistical methods: descriptive and inferential statistics	22
2. Univariate descriptive analysis: tables and graphical methods	23
2.1. The frequency distribution	23
2.2. Frequency distributions: tables and plots	24
2.2.1. <i>Qualitative variables</i>	24
2.2.2. <i>Quantitative variables</i>	29
2.3. Cumulative frequencies	39
2.4. Time series: some basic considerations	44
2.5. Misleading data representations	45
2.6. UBStats functions to tabulate and plot univariate distributions	49
Appendix 2.1. Details on the data used in this chapter	52
3. Summarising a single variable	53
3.1. Summary measures: an overview	53
3.2. Central tendency measures	54
3.2.1. <i>Central tendency measures for grouped and classified data</i>	59
3.2.2. <i>A comparison between median and mean</i>	61

3.2.3. <i>Applications</i>	65
3.3. Non-central tendency measures and the boxplot	69
3.3.1. <i>Quartiles</i>	69
3.3.2. <i>Five-number summary and the boxplot</i>	73
3.3.3. <i>Beyond five-number summary: quantiles and percentiles</i>	78
3.3.4. <i>Applications</i>	80
3.4. Dispersion measures	81
3.4.1. <i>Unit of measurement and coefficient of variation</i>	87
3.5. UBStats functions to calculate summary measures	90
Appendix 3.1. Properties of the mean and of the variance	92
Appendix 3.2. Details on the data used in this chapter	94
4. Bivariate descriptive analysis: tables, graphical methods, and summaries	95
4.1. The joint frequency distribution	95
4.2. Relationship between two variables with few distinct values	96
4.2.1. <i>Association and independence</i>	103
4.2.2. <i>Applications</i>	104
4.3. Comparing distributions across distinct groups	107
4.4. Studying the relation between two numerical variables	113
4.4.1. <i>Covariance</i>	115
4.4.2. <i>Covariance, correlation, and unit of measurement</i>	117
4.4.3. <i>The regression line</i>	123
4.4.4. <i>Correlation and regression line: some considerations</i>	125
4.4.5. <i>Correlation, causality, and the case for control variables</i>	128
4.4.6. <i>Some concluding remarks</i>	138
4.5. UBStats functions for bivariate analysis	138
Appendix 4.1. Properties of the covariance	140
Appendix 4.2. Determination of the coefficients of the regression line	141
Appendix 4.3. Details on the data used in this chapter	142
5. Probability and random variables	143
5.1. A (brief) formal introduction to random events and probability	143
5.1.1. <i>Probability: definition and interpretation</i>	145
5.1.2. <i>Conditional probability and independence</i>	149
5.1.3. <i>Bayes' theorem</i>	153
5.2. Random variables	155
5.2.1. <i>Discrete random variables</i>	156
5.2.2. <i>Relevant discrete random variables</i>	158
5.2.3. <i>Continuous random variables</i>	161
5.2.4. <i>Relevant continuous random variables</i>	164
5.2.5. <i>Linear transformations of random variables</i>	169

5.3. Joint distribution of random variables	173
5.3.1. <i>Joint distributions of discrete random variables</i>	173
5.3.2. <i>Joint density probability functions</i>	177
5.3.3. <i>Linear combinations of two random variables</i>	178
5.3.4. <i>Linear combinations of more random variables</i>	181
5.4. Sum and mean of i.i.d. variables and Central Limit theorem	181
5.4.1. <i>Central Limit theorem</i>	185
5.5. RStudio functions for probability and density distributions	190
Appendix 5.1. Basic probability results	191
Appendix 5.2. Theoretical results on expected value and variance	192
6. Estimation	193
6.1. Estimation: some introductory considerations	193
6.1.1. <i>Defining point estimators and estimates</i>	195
6.1.2. <i>Properties of point estimators</i>	196
6.2. Estimating the mean of a population	199
6.3. From point to interval estimation	203
6.4. Confidence intervals for the mean of a population	205
6.4.1. <i>Confidence interval for μ when the population variance σ^2 is known</i>	205
6.4.2. <i>Confidence interval for μ when the population variance σ^2 is unknown</i>	209
6.5. Estimating the population proportion	214
6.5.1. <i>Large-sample confidence interval for the proportion</i>	215
6.6. Estimating the difference between the means of two populations	219
6.6.1. <i>The case of independent samples</i>	221
6.6.2. <i>The case of paired samples</i>	230
6.7. Estimating the difference between two proportions	235
6.8. Why confidence intervals alone may not justify strong conclusions	242
6.9. RStudio functions for the Student's \mathcal{T} distribution	242
6.10. UBStats functions for confidence intervals	243
Appendix 6.1. Properties of the sample variance	246
Appendix 6.2. Explaining the Student's \mathcal{T} distribution	248
7. Hypothesis testing	249
7.1. Hypothesis test: some introductory considerations	249
7.1.1. <i>What is a statistical hypothesis?</i>	251
7.1.2. <i>Setup of a statistical test</i>	252
7.2. Tests on the population mean	256
7.2.1. <i>Tests on μ when the population variance σ^2 is known</i>	256
7.2.2. <i>Tests on μ when the population variance σ^2 is unknown</i>	273
7.3. Large sample tests on the proportion	279
7.4. Tests on the difference between the means of two populations	285

<i>7.4.1. Confidence intervals for the difference between means and test on variances</i>	298
7.5. Tests on the difference between population proportions	299
7.6. Goodness of fit and independence tests	306
<i>7.6.1. The Goodness of fit test</i>	307
<i>7.6.2. The Chi-square test of independence</i>	313
7.7. Concluding remarks	318
7.8. UBStats functions for hypothesis testing	319
7.9. R functions for Chi-square distribution and tests	323