Umberto Michelucci

Statistics for Scientists

A Concise Guide for Data-driven Research

TEXTSEED

Springer

Umberto Michelucci

Statistics for Scientists
A Concise Guide for Data-driven Research

Umberto Michelucci
Lucerne University Applied Science, Dübendorf, Switzerland

ISBN 978-3-031-78146-9 e-ISBN 978-3-031-78147-6
https://doi.org/10.1007/978-3-031-78147-6

To Caterina and Francesca. To Francesca, the reason why all this is possible. To Caterina, my life, the person I am most proud of in my life.

Preface

This short book has originated from lectures I gave at various universities that have helped me identify the concepts in statistics that most scientists struggle with. Most of them have an intuitive understanding of what, for example, mean and median are, but few can tell when to use which, for example, or can explain how to interpret the coefficient R^2. This book is a summary of those concepts, described in a short and concise way, enriched with tips and rigorous explanations.

It is important to understand that the goal of this book is not to provide a complete course in statistics. There are many books that do this already. My goal is to give researchers and practitioners a short text that is easy to read to help them understand the most important concepts in statistics and especially how to use them properly. Readers are encouraged to continue the study of statistics, an incredibly interesting branch of mathematics that helps us understand the world around us.

This book contains enough material for a short introductory course for students at an undergraduate level. You will find some sections of this book marked with a star $\star$ at the beginning of the title. That means that the section is more mathematically challenging and can be safely skipped by those with less mathematical know-how or interest.

This book is structured to provide an accessible overview of statistics and data analysis for scientific research. It begins with basic concepts, including an explanation of random variables, outcome spaces, and the difference between descriptive and inferential statistics. It continues with data types, measures of central tendency, of dispersion, and of positions. The discussion continues with a discussion of outliers and various methods to define them. Then the book introduces more complex topics like distributions, hypothesis testing, and regression analysis. Each chapter builds on the previous one, introducing more

complex statistical techniques in a step-by-step manner, making it suitable for readers ranging from beginners to those needing a quick refresher. The choice of topics, as often, is somewhat subjective. But I selected the concepts that all those interested in using statistics should know about. Some topics (such as hypothesis testing) are not discussed at length, since a complete treatment of the topics would require not only more space but also the student having some more background in statistics and mathematics (e.g., in calculus). The goal of these chapters is to give the student enough understanding of the main idea and inspire them to study the concepts deeper.

In the book you will find four types of boxes: *definitions*, *tips*, *warnings*, and *examples*. The meaning should be clear. In *definitions* boxes, you will find definitions of concepts, so it is easier to find them. In *tips* you will find, as the name suggests, tips and suggestions that I hope will make specific concepts or applications clearer. In *warnings* I try to highlight tricky cases in the application or interpretation of methods or concepts. And finally, in *examples* I try to give some examples to make concepts clearer.

Two books have significantly influenced my journey in learning statistics. One is the book *Probability and Statistical Inference* by Hogg, Tanis, and Zimmerman now in its 10th edition [1]. In my opinion, it is the perfect book for (almost) *beginners*. A more advanced book, but beautiful in its rigor and choice of topics, is the one by Casella and Berger *Statistical Inference* [2]. A masterpiece that every scientist should have on his or her bookshelf. If you are looking to delve deeper into statistics after reading this short introduction, you cannot go wrong with those two books. If you are a beginner, after this book, I suggest you continue with the book by Hogg et al.

Umberto Michelucci
Dübendorf, Switzerland
February 2025

Declarations

Competing Interests The author has no competing interests to declare that are relevant to the content of this manuscript.

Acronyms and Symbols

ANOVA Analysis of Variance: a collection of statistical models to compare multiple group means.

AUC Area Under the Curve: used in classification analysis to determine which of the used models predicts the classes best.

CI Confidence Interval: a range of values used to estimate the true value of a population parameter.

IQR Interquartile Range: a measure of statistical dispersion, being equal to the difference between 75th and 25th percentiles, or between upper and lower quartiles.

LR Logistic Regression: a statistical model that in its basic form uses a logistic function to model a binary-dependent variable.

MLE Maximum Likelihood Estimation: a method of estimating the parameters of a statistical model.

N Sample Size: the number of observations in a sample.

p-value Probability Value: the probability of obtaining test results at least as extreme as the ones observed during the test, assuming that the null hypothesis is correct.

R R (Programming Language): a language and environment for statistical computing and graphics.

ROC

Receiver Operating Characteristic: a graphical plot that illustrates the diagnostic ability of a binary classifier system.

RSS Residual Sum of Squares: a measure of the discrepancy between the data and an estimation model.

SAS Statistical Analysis System: a software suite used for advanced analytics, multivariate analyses, business intelligence, and data management.

SD, σ Standard Deviation: a measure that is used to quantify the amount of variation or dispersion of a set of data values.

SEM, $\sigma_{\bar{X}}$ Standard Error of the Mean: an estimate of how far the sample mean is likely to be from the population mean.

SPSS Statistical Package for the Social Sciences: a software package used for statistical analysis.

TSS Total Sum of Squares: sum of squares between the data and an estimation model.

Contents

List of Definitions

2.5.1: Longitudinal Data

2.6.1: Cross-Sectional Data

2.7.1: Binary Data

2.7.2: Dichotomous Data

3.2.1: Research Question

3.2.2: Hypothesis

3.3.1: Survey Sampling

3.3.2: Non-probability Sampling

3.3.3: Probability Sampling

3.4.1: Stratified Population

3.4.2: Clustered Population

3.5.1: Random Sampling Without Replacement