

Statistics and Computing

Thomas Haslwanter

An Introduction to Statistics with Python

With Applications in the Life Sciences

Second Edition



 Springer

Statistics and Computing

Series Editor

Wolfgang Karl Härdle, Humboldt-Universität zu Berlin, Berlin, Germany

Statistics and Computing (SC) includes monographs and advanced texts on statistical computing and statistical packages.

More information about this series at <https://link.springer.com/bookseries/3022>


Thomas Haslwanter

An Introduction to Statistics with Python

With Applications in the Life Sciences

Second Edition

 Springer

Thomas Haslwanter 
School of Medical Engineering and Applied
Social Sciences
University of Applied Sciences Upper Austria
Linz, Austria

ISSN 1431-8784

Statistics and Computing

ISBN 978-3-030-97370-4

<https://doi.org/10.1007/978-3-030-97371-1>

ISSN 2197-1706 (electronic)

ISBN 978-3-030-97371-1 (eBook)

1st edition: © Springer International Publishing Switzerland 2016

2nd edition: © Springer Nature Switzerland AG 2022

This work is subject to copyright. All rights are reserved by the Publisher, whether the whole or part of the material is concerned, specifically the rights of translation, reprinting, reuse of illustrations, recitation, broadcasting, reproduction on microfilms or in any other physical way, and transmission or information storage and retrieval, electronic adaptation, computer software, or by similar or dissimilar methodology now known or hereafter developed.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

*To my two-, three-, and four-legged
household companions: Jean, Felix, and his
sister Jessica: Thank you so much for all the
support you have provided over the years!*

Preface

Preface to the First Edition

In the data analysis for my own research work, I was often slowed down by two things: (1) I did not know enough statistics, and (2) the books available would provide a theoretical background, but no real practical help. The book you are holding in your hands (or on your tablet or laptop) is intended to be the book that will solve this very problem. It is designed to provide enough basic understanding so that you know what you are doing, and it should equip you with the tools you need. I believe that the *Python* solutions provided in this book for the most basic statistical problems address at least 90% of the problems that most physicists, biologists, and medical doctors encounter in their work. So if you are the typical graduate student working on your degree, or a medical researcher analyzing your latest experiments, chances are that you will find the tools you require here—explanation and source-code included.

This is the reason I have focused on statistical basics and hypothesis tests in this book, and refer only briefly to other statistical approaches. I am well aware that most of the tests presented in this book can also be carried out using statistical modeling. But in many cases, this is not the methodology used in many life science journals. Advanced statistical analysis goes beyond the scope of this book, and—to be frank—exceeds my own knowledge of statistics.

My motivation for providing the solutions in Python is based on two considerations. One is that I would like them to be available to everyone. While commercial solutions like *Matlab*, *SPSS*, *Minitab* etc. offer powerful tools, most can only use them legally in an academic setting. In contrast, Python is completely free (as in free beer is often heard in the Python community). The second reason is that Python is the most beautiful coding language that I have yet encountered; and around 2010 Python and its documentation matured to the point where one can use it without being an serious coder. Together, this book, Python, and the tools that the Python ecosystem offers today provide a beautiful, free package that covers all the statistics that most researchers will need in their lifetime.

Preface to the Second Edition

Since the publication of the first edition, Python has continuously gained popularity and become firmly established as one of the foremost programming languages for statistical data analysis. All the core packages have matured. And thanks to the stunning development of *Jupyter* as an interactive programming environment, Python has become even more accessible for people with little programming background. To reflect these developments, and to incorporate the suggestions I have received for improving the presentation of the material, Springer has given me the opportunity to bring out a new edition of *Introduction to Statistics with Python*.

Compared to the first edition, the following changes have been made:

- The package *pandas* and its *DataFrames* have become an integral part of scientific Python, as has the *Jupyter* framework for interactive data environments. Correspondingly, a bigger amount of space has been dedicated to their introduction.
- A new package, *pingouin*, is promising a simplified and more powerful interface for many common statistics function. This package is introduced, and many application examples have been added.
- The visualization of data has been expanded, including the preparation of publication-ready graphics.
- The design of experiments and power analyses are discussed in more detail.
- A new section has been added on the confidence intervals of frequently used statistical parameters.
- A new chapter has been added on finding patterns in data, including an introduction to the correlation coefficient, cross- and autocorrelation. For an application of these concepts, a short introduction is given to time series analysis.

As for the first edition, all examples and solutions from this book are again available online. This includes code samples and example programs, Jupyter Notebooks with additional or extended information, as well as the data and Python code used to generate most of the figures. They can be downloaded from <https://github.com/thomas-haslwanter/statsintro-python-2e>.

I hope this book will help you with the statistical analysis of your data, and convey some of the often really simple ideas behind the sometimes awkwardly named statistical analysis procedures.

For Whom This Book Is

This book assumes that

- you have some basic programming experience: If you have done no programming previously, you may want to start out with Python using some of the great links provided in the text. Starting programming *and* starting statistics may be a bit

much all at once. However, solutions provided to the exercises at the end of most chapters should help you to get up to speed with Python.

- you are not a statistics expert: If you have advanced statistics experience, the online help in Python and the Python packages may be sufficient to allow you to do most of your data analysis right away. This book may still help you to get started with Python. However, the book concentrates on the basic ideas of statistics and on hypothesis tests, and only the last part introduces linear regression modeling and Bayesian statistics.

This book is designed to give you all (or at least most of) the tools that you will need for statistical data analysis. I attempt to provide the background you need to understand what you are doing. I do not prove any theorems, and do not apply mathematics unless necessary. For all tests, a working Python program is provided. In principle, you just have to define your problem, select the corresponding program, and adapt it to your needs. This should allow you to get going quickly, even if you have little Python experience. This is also the reason why I have not provided the software as one single Python package; I expect that you will have to tailor each program to your specific setup (data format, etc.).

This book is organized into three parts

Part I gives an introduction to Python: how to set it up, simple programs to get started, and tips on how to avoid some common mistakes. It also shows how to read data from different sources into Python, and how to visualize statistical data.

Part II provides an introduction to statistical analysis; on how to design a study, power analysis, and how best to analyze data; probability distributions; and an overview of the most important hypothesis tests. Even though modern statistics is firmly based in statistical modeling, hypothesis tests still seem to dominate the life sciences. For each test, a Python program is provided that shows how the test can be implemented.

Part III provides an introduction to correlation and regression analysis, time series analysis, and statistical modeling, and a look at advanced statistical analysis procedures. I have also included tests on discrete data in this section, such as logistic regression, as they utilize “generalized linear models” which I regard as advanced. This part ends with a presentation of the basic ideas of Bayesian statistics.

To achieve all those goals as quickly as possible, the Appendix A of the book provides hints on how to efficiently develop correct and working code. This should get you to the point where you can get things done quickly.

Acknowledgments

Python is built on the contributions from the user community, and some of the sections in this book are based on some of the excellent information available on the web. (Permission has been granted by the authors to reprint their contributions here.)

I especially want to thank the following people:

- Christiane Takacs helped me enormously by polishing the introductory statistics sections.
- Connor Johnson wrote a very nice blog explaining the results of the statsmodels OLS command, which provided the basis for the section on *Statistical Models*.
- Cam Davidson Pilon wrote the excellent open-source e-book *Probabilistic-Programming-and-Bayesian-Methods-for-Hackers*. From there, I took the example of the Challenger disaster to demonstrate Bayesian statistics.
- Fabian Pedregosa's blog on ordinal logistic regression allowed me to include this topic, which otherwise would be admittedly beyond my own skills.

I also want to thank Springer Publishing for the chance to bring out the second edition of this book, and to base the three introductory chapters (Python, Data Import, and Data Display) to a significant part on the corresponding chapters of my book *Hands-on Signal Analysis with Python*.

If you have a suggestion or correction, please send an email to my work address thomas.haslwanter@fh-ooe.at. If I make a change based on your feedback, I will add you to the list of contributors unless advised otherwise. If you include at least part of the sentence the error appears in, that makes it easy for me to search. Page and section numbers are fine, too, but not as easy to work with. Thanks!

Linz, Austria
August 2022

Thomas Haslwanter

Contents

Part I Python and Statistics

1	Introduction	3
1.1	Why Statistics?	3
1.2	Conventions	5
1.3	Accompanying Material	5
2	Python	7
2.1	Getting Started	8
2.2	Elements of Scientific Python Programming	15
2.3	Interactive Programming—IPython/Jupyter	28
2.4	Statistics Packages for Python	39
2.5	Programming Tips	43
2.6	Exercises	45
3	Data Input	49
3.1	Text	49
3.2	Excel	54
3.3	Matlab	54
3.4	Binary Data: NPZ Format	55
3.5	Other Formats	56
3.6	Exercises	56
4	Data Display	59
4.1	Introductory Example	59
4.2	Plotting in Python	62
4.3	Saving a Figure	66
4.4	Preparing Figures for Presentation	67
4.5	Display of Statistical Data Sets	70
4.6	Exercises	82

Part II Distributions and Hypothesis Tests

- 5 Basic Statistical Concepts** 87
 - 5.1 Populations and Samples 87
 - 5.2 Data Types 89
 - 5.3 Probability Distributions 90
 - 5.4 Degrees of Freedom 94
 - 5.5 Study Design 94
- 6 Distributions of One Variable** 105
 - 6.1 Characterizing a Distribution 105
 - 6.2 Discrete Distributions 115
 - 6.3 Normal Distribution 120
 - 6.4 Continuous Distributions Derived from the Normal
Distribution 125
 - 6.5 Other Continuous Distributions 132
 - 6.6 Confidence Intervals of Selected Statistical Parameters 135
 - 6.7 Exercises 136
- 7 Hypothesis Tests** 139
 - 7.1 Typical Analysis Procedure 139
 - 7.2 Hypothesis Tests and Power Analyses 144
 - 7.3 Sensitivity and Specificity 152
 - 7.4 Receiver-Operating-Characteristic (ROC) Curve 155
 - 7.5 Exercises 157
- 8 Tests of Means of Numerical Data** 159
 - 8.1 Distribution of a Sample Mean 159
 - 8.2 Comparison of Two Groups 164
 - 8.3 Comparison of Multiple Groups 168
 - 8.4 Summary: Selecting the Right Test for Comparing Groups 176
 - 8.5 Exercises 178
- 9 Tests on Categorical Data** 181
 - 9.1 Proportions and Confidence Intervals 182
 - 9.2 Tests Using Frequency Tables 183
 - 9.3 Exercises 194
- 10 Analysis of Survival Times** 197
 - 10.1 Survival Distributions 197
 - 10.2 Survival Probabilities 198
 - 10.3 Comparing Survival Curves in Two Groups 202

Part III Statistical Modeling

- 11 Finding Patterns in Signals** 205
 - 11.1 Cross Correlation 205
 - 11.2 Correlation Coefficient 208
 - 11.3 Coefficient of Determination 211

- 11.4 Scatterplot Matrix 214
- 11.5 Correlation Matrix 214
- 11.6 Autocorrelation 217
- 11.7 Time-Series Analysis 218
- 12 Linear Regression Models 229**
 - 12.1 Simple Fits 230
 - 12.2 Design Matrix and Formulas 232
 - 12.3 Linear Regression Analysis with Python 237
 - 12.4 Model Results of Linear Regression Models 241
 - 12.5 Assumptions and Interpretations of Linear Regression 257
 - 12.6 Bootstrapping 262
 - 12.7 Exercises 262
- 13 Generalized Linear Models 265**
 - 13.1 Comparing and Modeling Ranked Data 265
 - 13.2 Elements of GLMs 266
 - 13.3 GLM 1: Logistic Regression 267
 - 13.4 GLM 2: Ordinal Logistic Regression 270
 - 13.5 Exercises 274
- 14 Bayesian Statistics 275**
 - 14.1 Bayesian Versus Frequentist Interpretation 275
 - 14.2 The Bayesian Approach in the Age of Computers 277
 - 14.3 Example: Markov-Chain-Monte-Carlo Simulation 278
 - 14.4 Summing Up 280
- Appendix A: Useful Programming Tools 283**
- Appendix B: Solutions 293**
- Appendix C: Equations for Confidence Intervals 321**
- Appendix D: Web Ressources 323**
- Glossary 325**
- Bibliography 331**
- Index 333**