Statistics and Computing

Series editor
W.K. Härdle

More information about this series at http://www.springer.com/series/3022

Thomas Haslwanter

An Introduction to Statistics with Python

With Applications in the Life Sciences

Springer

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

Series Editor:
W.K. Härdle
C.A.S.E. Centre for Applied
Statistics and Economics
School of Business and Economics
Humboldt-Universität zu Berlin
Unter den Linden 6
10099 Berlin
Germany

The Python code samples accompanying the book are available at www.quantlet.de. All Python programs and data sets can be found on GitHub: https://github.com/thomas-haslwanter/statsintro_python.git. Links to all material are available at http://www.springer.com/de/book/9783319283159.

ISSN 1431-8784 ISSN 2197-1706 (electronic)
Statistics and Computing
ISBN 978-3-319-28315-9 ISBN 978-3-319-28316-6 (eBook)
DOI 10.1007/978-3-319-28316-6

Library of Congress Control Number: 2016939946

Printed on acid-free paper

This Springer imprint is published by Springer Nature
The registered company is Springer International Publishing AG Switzerland

To my two, three, and four-legged household companions: my wife Jean, Felix, and his sister Jessica.

Preface

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 a degree, or a medical researcher analyzing the 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 a 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.

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.
- 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, plot labels, return values, 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 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. How to design a study, 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 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. The book ends with a presentation of the basic ideas of Bayesian statistics.

Additional Material

This book comes with many additional *Python* programs and sample data, which are available online. These programs include listings of the programs printed in the book, solutions to the examples given at the end of most chapters, and code samples

with a working example for each test presented in this book. They also include the code used to generate the pictures in this book, as well as the data used to run the programs.

The Python code samples accompanying the book are available at http://www.quantlet.de. All Python programs and data sets can be found on GitHub: https://github.com/thomas-haslwanter/statsintro_python.git. Links to all material are available at http://www.springer.com/de/book/9783319283159.

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:

- Paul E. Johnson read the whole manuscript and provided invaluable feedback on the general structure of the book, as well as on statistical details.
- 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 Carolyn Mayer for reading the manuscript and replacing colloquial expressions with professional English. And a special hug goes to my wife, who not only provided important suggestions for the structure of the book, but also helped with tips on how to teach programming, and provided support with all the tea-related aspects of the book.

If you have a suggestion or correction, please send an email to my work address `thomas.haslwanter@fh-linz.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 Thomas Haslwanter
December 2015

Contents

Acronyms

ANOVA	ANalysis Of VAriance
CDF	Cumulative distribution function
CI	Confidence interval
DF/DOF	Degrees of freedom
EOL	End of line
GLM	Generalized linear models
HTML	HyperText Markup Language
IDE	Integrated development environment
IQR	Inter quartile range
ISF	Inverse survival function
KDE	Kernel density estimation
MCMC	Markov chain Monte Carlo
NAN	Not a number
OLS	Ordinary least squares
PDF	Probability density function
PPF	Percentile point function
QQ-Plot	Quantile-quantile plot
ROC	Receiver operating characteristic
RVS	Random variate sample
SD	Standard deviation
SE/SEM	Standard error (of the mean)
SF	Survival function
SQL	Structured Query Language
SS	Sum of squares
Tukey HSD	Tukey honest significant difference test