

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>.

The Python solution codes in the appendix are published under the Creative Commons Attribution-ShareAlike 4.0 International License.

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

© Springer International Publishing Switzerland 2016

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, express or implied, with respect to the material contained herein or for any errors or omissions that may have been made.

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
December 2015

Thomas Haslwanter

Contents

Part I Python and Statistics

1	Why Statistics?	3
2	Python	5
2.1	Getting Started.....	5
2.1.1	Conventions	5
2.1.2	Distributions and Packages.....	6
2.1.3	Installation of Python.....	8
2.1.4	Installation of R and rpy2	10
2.1.5	Personalizing IPython/ <i>Jupyter</i>	11
2.1.6	Python Resources	14
2.1.7	First Python Programs	15
2.2	Python Data Structures	17
2.2.1	Python Datatypes	17
2.2.2	Indexing and Slicing	19
2.2.3	Vectors and Arrays	19
2.3	IPython/ <i>Jupyter</i> : An Interactive Programming Environment	21
2.3.1	First Session with the Qt Console	22
2.3.2	Notebook and rpy2	24
2.3.3	IPython Tips	26
2.4	Developing Python Programs	27
2.4.1	Converting Interactive Commands into a Python Program	27
2.4.2	Functions, Modules, and Packages	30
2.4.3	Python Tips	34
2.4.4	Code Versioning	34
2.5	Pandas: Data Structures for Statistics	35
2.5.1	Data Handling	35
2.5.2	Grouping	37
2.6	Statsmodels: Tools for Statistical Modeling	39
2.7	Seaborn: Data Visualization	40

- 2.8 General Routines 41
- 2.9 Exercises 42
- 3 Data Input 43**
 - 3.1 Input from Text Files 43
 - 3.1.1 Visual Inspection 43
 - 3.1.2 Reading ASCII-Data into Python 44
 - 3.2 Input from MS Excel 47
 - 3.3 Input from Other Formats 49
 - 3.3.1 Matlab 49
- 4 Display of Statistical Data 51**
 - 4.1 Datatypes 51
 - 4.1.1 Categorical 51
 - 4.1.2 Numerical 52
 - 4.2 Plotting in Python 52
 - 4.2.1 Functional and Object-Oriented Approaches
to Plotting 54
 - 4.2.2 Interactive Plots 55
 - 4.3 Displaying Statistical Datasets 59
 - 4.3.1 Univariate Data 59
 - 4.3.2 Bivariate and Multivariate Plots 69
 - 4.4 Exercises 71

Part II Distributions and Hypothesis Tests

- 5 Background 75**
 - 5.1 Populations and Samples 75
 - 5.2 Probability Distributions 76
 - 5.2.1 Discrete Distributions 77
 - 5.2.2 Continuous Distributions 77
 - 5.2.3 Expected Value and Variance 78
 - 5.3 Degrees of Freedom 79
 - 5.4 Study Design 79
 - 5.4.1 Terminology 79
 - 5.4.2 Overview 80
 - 5.4.3 Types of Studies 81
 - 5.4.4 Design of Experiments 82
 - 5.4.5 Personal Advice 86
 - 5.4.6 Clinical Investigation Plan 87
- 6 Distributions of One Variable 89**
 - 6.1 Characterizing a Distribution 89
 - 6.1.1 Distribution Center 89
 - 6.1.2 Quantifying Variability 91
 - 6.1.3 Parameters Describing the Form of a Distribution 96
 - 6.1.4 Important Presentations of Probability Densities 98

- 6.2 Discrete Distributions 99
 - 6.2.1 Bernoulli Distribution 100
 - 6.2.2 Binomial Distribution 100
 - 6.2.3 Poisson Distribution 103
- 6.3 Normal Distribution 104
 - 6.3.1 Examples of Normal Distributions 107
 - 6.3.2 Central Limit Theorem 107
 - 6.3.3 Distributions and Hypothesis Tests 108
- 6.4 Continuous Distributions Derived from the Normal Distribution 109
 - 6.4.1 t -Distribution 110
 - 6.4.2 Chi-Square Distribution 111
 - 6.4.3 F -Distribution 113
- 6.5 Other Continuous Distributions 115
 - 6.5.1 Lognormal Distribution 116
 - 6.5.2 Weibull Distribution 116
 - 6.5.3 Exponential Distribution 118
 - 6.5.4 Uniform Distribution 118
- 6.6 Exercises 119
- 7 Hypothesis Tests 121**
 - 7.1 Typical Analysis Procedure 121
 - 7.1.1 Data Screening and Outliers 122
 - 7.1.2 Normality Check 122
 - 7.1.3 Transformation 126
 - 7.2 Hypothesis Concept, Errors, p -Value, and Sample Size 126
 - 7.2.1 An Example 126
 - 7.2.2 Generalization and Applications 127
 - 7.2.3 The Interpretation of the p -Value 128
 - 7.2.4 Types of Error 129
 - 7.2.5 Sample Size 131
 - 7.3 Sensitivity and Specificity 134
 - 7.3.1 Related Calculations 136
 - 7.4 Receiver-Operating-Characteristic (ROC) Curve 136
- 8 Tests of Means of Numerical Data 139**
 - 8.1 Distribution of a Sample Mean 139
 - 8.1.1 One Sample t -Test for a Mean Value 139
 - 8.1.2 Wilcoxon Signed Rank Sum Test 141
 - 8.2 Comparison of Two Groups 142
 - 8.2.1 Paired t -Test 142
 - 8.2.2 t -Test between Independent Groups 143
 - 8.2.3 Nonparametric Comparison of Two Groups:
 - Mann–Whitney Test 144
 - 8.2.4 Statistical Hypothesis Tests vs Statistical Modeling 144

8.3	Comparison of Multiple Groups	146
8.3.1	Analysis of Variance (ANOVA)	146
8.3.2	Multiple Comparisons	150
8.3.3	Kruskal–Wallis Test	152
8.3.4	Two-Way ANOVA	152
8.3.5	Three-Way ANOVA	154
8.4	Summary: Selecting the Right Test for Comparing Groups	155
8.4.1	Typical Tests	155
8.4.2	Hypothetical Examples	156
8.5	Exercises	157
9	Tests on Categorical Data	159
9.1	One Proportion	160
9.1.1	Confidence Intervals	160
9.1.2	Explanation	160
9.1.3	Example	161
9.2	Frequency Tables	162
9.2.1	One-Way Chi-Square Test	162
9.2.2	Chi-Square Contingency Test	163
9.2.3	Fisher’s Exact Test	165
9.2.4	McNemar’s Test	169
9.2.5	Cochran’s Q Test	170
9.3	Exercises	171
10	Analysis of Survival Times	175
10.1	Survival Distributions	175
10.2	Survival Probabilities	176
10.2.1	Censorship	176
10.2.2	Kaplan–Meier Survival Curve	177
10.3	Comparing Survival Curves in Two Groups	180
Part III Statistical Modeling		
11	Linear Regression Models	183
11.1	Linear Correlation	184
11.1.1	Correlation Coefficient	184
11.1.2	Rank Correlation	184
11.2	General Linear Regression Model	185
11.2.1	Example 1: Simple Linear Regression	187
11.2.2	Example 2: Quadratic Fit	187
11.2.3	Coefficient of Determination	188
11.3	Patsy: The Formula Language	190
11.3.1	Design Matrix	190
11.4	Linear Regression Analysis with Python	193
11.4.1	Example 1: Line Fit with Confidence Intervals	193
11.4.2	Example 2: Noisy Quadratic Polynomial	194
11.5	Model Results of Linear Regression Models	198

11.5.1	Example: Tobacco and Alcohol in the UK	198
11.5.2	Definitions for Regression with Intercept	200
11.5.3	The R^2 Value	201
11.5.4	\bar{R}^2 : The <i>Adjusted</i> R^2 Value	201
11.5.5	Model Coefficients and Their Interpretation	205
11.5.6	Analysis of Residuals.....	209
11.5.7	Outliers.....	212
11.5.8	Regression Using Sklearn	212
11.5.9	Conclusion	214
11.6	Assumptions of Linear Regression Models	214
11.7	Interpreting the Results of Linear Regression Models	218
11.8	Bootstrapping	219
11.9	Exercises.....	220
12	Multivariate Data Analysis	221
12.1	Visualizing Multivariate Correlations	221
12.1.1	Scatterplot Matrix	221
12.1.2	Correlation Matrix	222
12.2	Multilinear Regression	223
13	Tests on Discrete Data	227
13.1	Comparing Groups of Ranked Data	227
13.2	Logistic Regression	228
13.2.1	Example: The Challenger Disaster	228
13.3	Generalized Linear Models	231
13.3.1	Exponential Family of Distributions.....	231
13.3.2	Linear Predictor and Link Function	232
13.4	Ordinal Logistic Regression	232
13.4.1	Problem Definition	232
13.4.2	Optimization	234
13.4.3	Code	235
13.4.4	Performance.....	235
14	Bayesian Statistics	237
14.1	Bayesian vs. Frequentist Interpretation	237
14.1.1	Bayesian Example	238
14.2	The Bayesian Approach in the Age of Computers.....	239
14.3	Example: Analysis of the Challenger Disaster with a Markov-Chain-Monte-Carlo Simulation	240
14.4	Summing Up	243
	Solutions.....	245
	Glossary	267
	References.....	273
	Index	275

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