

SECOND EDITION

Think Stats

Allen B. Downey

Beijing • Cambridge • Farnham • Köln • Sebastopol • Tokyo

O'REILLY[®]

Think Stats, Second Edition

by Allen B. Downey

Copyright © 2015 Allen B. Downey. All rights reserved.

Printed in the United States of America.

Published by O'Reilly Media, Inc., 1005 Gravenstein Highway North, Sebastopol, CA 95472.

O'Reilly books may be purchased for educational, business, or sales promotional use. Online editions are also available for most titles (<http://safaribooksonline.com>). For more information, contact our corporate/institutional sales department: 800-998-9938 or corporate@oreilly.com.

Editors: Mike Loukides and Meghan Blanchette

Indexer: Allen B. Downey

Production Editor: Melanie Yarbrough

Cover Designer: Karen Montgomery

Copyeditor: Marta Justak

Interior Designer: David Futato

Proofreader: Amanda Kersey

Illustrator: Rebecca Demarest

October 2014: Second Edition

Revision History for the Second Edition:

2014-10-09: First release

See <http://oreilly.com/catalog/errata.csp?isbn=9781491907337> for release details.

The O'Reilly logo is a registered trademarks of O'Reilly Media, Inc. *Think Stats*, second edition, the cover image of an archerfish, and related trade dress are trademarks of O'Reilly Media, Inc.

Many of the designations used by manufacturers and sellers to distinguish their products are claimed as trademarks. Where those designations appear in this book, and O'Reilly Media, Inc. was aware of a trademark claim, the designations have been printed in caps or initial caps.

While the publisher and the author have used good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the author disclaim all responsibility for errors or omissions, including without limitation responsibility for damages resulting from the use of or reliance on this work. Use of the information and instructions contained in this work is at your own risk. If any code samples or other technology this work contains or describes is subject to open source licenses or the intellectual property rights of others, it is your responsibility to ensure that your use thereof complies with such licenses and/or rights.

Think Stats is available under the *Creative Commons Attribution-NonCommercial-ShareAlike 3.0 Unported License*. The author maintains an online version at <http://thinkstats2.com>.

ISBN: 978-1-491-90733-7

[LSI]

Table of Contents

Preface	ix
1. Exploratory Data Analysis	1
A Statistical Approach	2
The National Survey of Family Growth	2
Importing the Data	3
DataFrames	4
Variables	6
Transformation	7
Validation	8
Interpretation	9
Exercises	11
Glossary	12
2. Distributions	15
Representing Histograms	16
Plotting Histograms	16
NSFG Variables	17
Outliers	19
First Babies	20
Summarizing Distributions	22
Variance	23
Effect Size	23
Reporting Results	24
Exercises	25
Glossary	25
3. Probability Mass Functions	27
Pmfs	27

Plotting PMFs	28
Other Visualizations	30
The Class Size Paradox	30
DataFrame Indexing	34
Exercises	35
Glossary	37
4. Cumulative Distribution Functions.....	39
The Limits of PMFs	39
Percentiles	40
CDFs	41
Representing CDFs	42
Comparing CDFs	44
Percentile-Based Statistics	44
Random Numbers	45
Comparing Percentile Ranks	47
Exercises	47
Glossary	48
5. Modeling Distributions.....	49
The Exponential Distribution	49
The Normal Distribution	52
Normal Probability Plot	54
The lognormal Distribution	55
The Pareto Distribution	57
Generating Random Numbers	60
Why Model?	61
Exercises	61
Glossary	63
6. Probability Density Functions.....	65
PDFs	65
Kernel Density Estimation	67
The Distribution Framework	69
Hist Implementation	69
Pmf Implementation	70
Cdf Implementation	71
Moments	72
Skewness	73
Exercises	75
Glossary	77

7. Relationships Between Variables.....	79
Scatter Plots	79
Characterizing Relationships	82
Correlation	83
Covariance	84
Pearson's Correlation	85
Nonlinear Relationships	86
Spearman's Rank Correlation	87
Correlation and Causation	88
Exercises	88
Glossary	89
8. Estimation.....	91
The Estimation Game	91
Guess the Variance	93
Sampling Distributions	94
Sampling Bias	97
Exponential Distributions	98
Exercises	99
Glossary	100
9. Hypothesis Testing.....	101
Classical Hypothesis Testing	101
Hypothesis Test	102
Testing a Difference in Means	104
Other Test Statistics	105
Testing a Correlation	107
Testing Proportions	108
Chi-Squared Tests	109
First Babies Again	110
Errors	111
Power	112
Replication	113
Exercises	114
Glossary	114
10. Linear Least Squares.....	117
Least Squares Fit	117
Implementation	118
Residuals	119
Estimation	120
Goodness of Fit	122

Testing a Linear Model	124
Weighted Resampling	126
Exercises	127
Glossary	128
11. Regression.....	129
StatsModels	130
Multiple Regression	131
Nonlinear Relationships	133
Data Mining	134
Prediction	135
Logistic Regression	137
Estimating Parameters	139
Implementation	140
Accuracy	141
Exercises	142
Glossary	143
12. Time Series Analysis.....	145
Importing and Cleaning	145
Plotting	147
Linear Regression	148
Moving Averages	151
Missing Values	153
Serial Correlation	153
Autocorrelation	155
Prediction	157
Further Reading	161
Exercises	161
Glossary	162
13. Survival Analysis.....	165
Survival Curves	165
Hazard Function	167
Estimating Survival Curves	168
Kaplan-Meier Estimation	169
The Marriage Curve	170
Estimating the Survival Function	171
Confidence Intervals	172
Cohort Effects	173
Extrapolation	176
Expected Remaining Lifetime	178

Exercises	180
Glossary	181
14. Analytic Methods.....	183
Normal Distributions	183
Sampling Distributions	184
Representing Normal Distributions	185
Central Limit Theorem	186
Testing the CLT	187
Applying the CLT	190
Correlation Test	191
Chi-Squared Test	193
Discussion	194
Exercises	195
Index.....	197