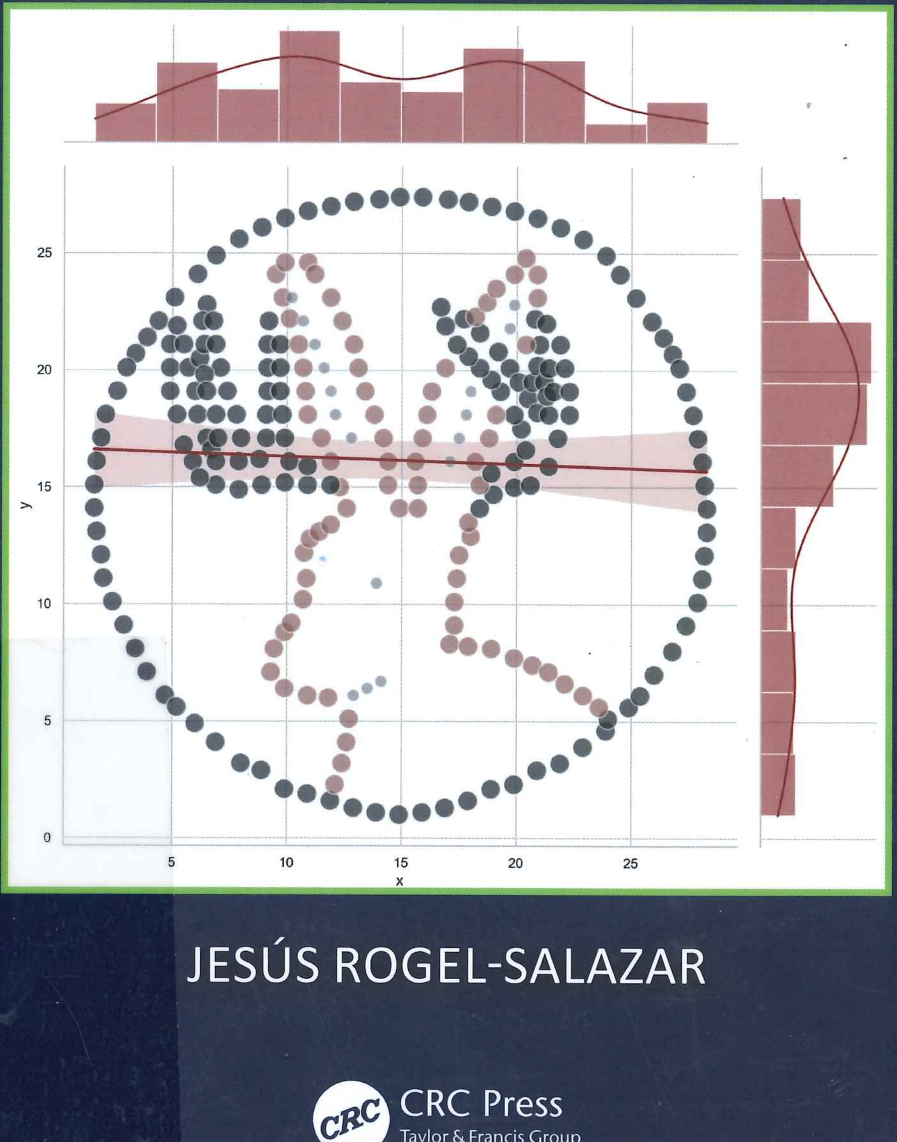
THE PYTHON SERIES

STATISTICS AND DATA VISUALISATION WITH PYTHON



JESÚS ROGEL-SALAZAR

CRC Press
Taylor & Francis Group

A CHAPMAN & HALL BOOK

สำนักหอสมุด มหาวิทยาลัยเชียงใหม่

b16704393
o12576451
i22684864

Statistics and Data Visualisation with Python

สำนักหอสมุดมหาวิทยาลัยเชียงใหม่
CHIANG MAI UNIVERSITY LIBRARY

Jesús Rogel-Salazar

CRC Press
Taylor & Francis Group
Boca Raton London New York

CRC Press is an imprint of the
Taylor & Francis Group, an **informa** business
A CHAPMAN & HALL BOOK

Contents

1 Data, Stats and Stories – An Introduction 1

1.1 From Small to Big Data 2

1.2 Numbers, Facts and Stats 10

1.3 A Sampled History of Statistics 14

1.4 Statistics Today 22

1.5 Asking Questions and Getting Answers 25

1.6 Presenting Answers Visually 30

2 Python Programming Primer 33

2.1 Talking to Python 35

2.1.1 Scripting and Interacting 38

2.1.2 Jupyter Notebook 41

2.2 Starting Up with Python 42

2.2.1 Types in Python 43

2.2.2 *Numbers: Integers and Floats* 43

2.2.3 *Strings* 46

2.2.4 *Complex Numbers* 49

2.3 *Collections in Python* 51

2.3.1 *Lists* 52

2.3.2 *List Comprehension* 60

2.3.3 *Tuples* 61

2.3.4 *Dictionaries* 66

2.3.5 *Sets* 72

2.4 *The Beginning of Wisdom: Logic & Control Flow* 80

2.4.1 *Booleans and Logical Operators* 80

2.4.2 *Conditional Statements* 82

2.4.3 *While Loop* 85

2.4.4 *For Loop* 87

2.5 *Functions* 89

2.6 *Scripts and Modules* 94

3 *Snakes, Bears & Other Numerical Beasts: NumPy, SciPy & pandas* 99

3.1 *Numerical Python – NumPy* 100

3.1.1 *Matrices and Vectors* 101

3.1.2 *N-Dimensional Arrays* 102

3.1.3 *N-Dimensional Matrices* 104

3.1.4 *Indexing and Slicing* 107

3.1.5 *Descriptive Statistics* 109

3.2 *Scientific Python – SciPy* 112

3.2.1 *Matrix Algebra* 114

3.2.2 *Numerical Integration* 116

3.2.3 *Numerical Optimisation* 117

3.2.4 *Statistics* 118

3.3 *Panel Data = pandas* 121

3.3.1 *Series and Dataframes* 122

3.3.2 *Data Exploration with pandas* 124

3.3.3 *Pandas Data Types* 125

3.3.4 *Data Manipulation with pandas* 126

3.3.5 *Loading Data to pandas* 130

3.3.6 *Data Grouping* 136

4 *The Measure of All Things – Statistics* 141

4.1 *Descriptive Statistics* 144

4.2 *Measures of Central Tendency and Dispersion* 145

4.3 *Central Tendency* 146

4.3.1 *Mode* 147

4.3.2 *Median* 150

4.3.3 *Arithmetic Mean* 152

4.3.4 *Geometric Mean* 155

4.3.5 *Harmonic Mean* 159

4.4 *Dispersion* 163

4.4.1 *Setting the Boundaries: Range* 163

4.4.2 *Splitting One's Sides: Quantiles, Quartiles, Percentiles and More* 166

4.4.3 *Mean Deviation* 169

4.4.4 *Variance and Standard Deviation* 171

4.5 *Data Description – Descriptive Statistics Revisited* 176

5 *Definitely Maybe: Probability and Distributions* 179

5.1 *Probability* 180

5.2 *Random Variables and Probability Distributions* 182

5.2.1 *Random Variables* 183

5.2.2 *Discrete and Continuous Distributions* 185

5.2.3 *Expected Value and Variance* 186

5.3 *Discrete Probability Distributions* 191

5.3.1 *Uniform Distribution* 191

5.3.2 *Bernoulli Distribution* 197

5.3.3 *Binomial Distribution* 201

5.3.4 *Hypergeometric Distribution* 208

5.3.5 *Poisson Distribution* 216

5.4 *Continuous Probability Distributions* 223

5.4.1 *Normal or Gaussian Distribution* 224

5.4.2 *Standard Normal Distribution Z* 235

5.4.3 *Shape and Moments of a Distribution* 238

5.4.4 *The Central Limit Theorem* 245

5.5 *Hypothesis and Confidence Intervals* 247

5.5.1 *Student's t Distribution* 253

5.5.2 *Chi-squared Distribution* 260

6 *Alluring Arguments and Ugly Facts – Statistical Modelling and Hypothesis Testing* 267

6.1 *Hypothesis Testing* 268

6.1.1 *Tales and Tails: One- and Two-Tailed Tests* 273

6.2 *Normality Testing* 279

6.2.1 *Q-Q Plot* 280

6.2.2 *Shapiro-Wilk Test* 282

6.2.3 *D'Agostino K-squared Test* 285

6.2.4 *Kolmogorov-Smirnov Test* 288

6.3 *Chi-square Test* 291

6.3.1 *Goodness of Fit* 291

6.3.2 *Independence* 293

6.4 *Linear Correlation and Regression* 296

6.4.1 *Pearson Correlation* 296

6.4.2 *Linear Regression* 301

6.4.3 *Spearman Correlation* 308

6.5 *Hypothesis Testing with One Sample* 312

6.5.1 *One-Sample t-test for the Population Mean* 312

6.5.2 *One-Sample z-test for Proportions* 316

6.5.3 *Wilcoxon Signed Rank with One-Sample* 320

6.6 *Hypothesis Testing with Two Samples* 324

6.6.1 *Two-Sample t-test – Comparing Means, Same Variances* 325

6.6.2 *Levene's Test – Testing Homoscedasticity* 330

6.6.3 *Welch's t-test – Comparing Means, Different Variances* 332

6.6.4 *Mann-Whitney Test – Testing Non-normal Samples* 334

6.6.5 *Paired Sample t-test* 338

6.6.6 *Wilcoxon Matched Pairs* 342

6.7 *Analysis of Variance* 345

6.7.1 *One-factor or One-way ANOVA* 347

6.7.2 *Tukey's Range Test* 360

6.7.3 *Repeated Measures ANOVA* 361

6.7.4 *Kruskal-Wallis – Non-parametric One-way ANOVA* 365

6.7.5 *Two-factor or Two-way ANOVA* 369

6.8 *Tests as Linear Models* 376

6.8.1 *Pearson and Spearman Correlations* 377

6.8.2 *One-sample t- and Wilcoxon Signed Rank Tests* 378

6.8.3 *Two-Sample t- and Mann-Whitney Tests* 379

6.8.4 *Paired Sample t- and Wilcoxon Matched Pairs Tests* 380

6.8.5 *One-way ANOVA and Kruskal-Wallis Test* 380

7 *Delightful Details – Data Visualisation* 383

7.1 *Presenting Statistical Quantities* 384

7.1.1 *Textual Presentation* 385

7.1.2 *Tabular Presentation* 385

7.1.3 *Graphical Presentation* 386

7.2 *Can You Draw Me a Picture? – Data Visualisation* 387

7.3 *Design and Visual Representation* 394

7.4 *Plotting and Visualising: Matplotlib* 402

7.4.1 *Keep It Simple: Plotting Functions* 403

7.4.2 *Line Styles and Colours* 404

7.4.3 *Titles and Labels* 405

7.4.4 *Grids* 406

7.5 *Multiple Plots* 407

7.6 *Subplots* 407

7.7 *Plotting Surfaces* 410

7.8 *Data Visualisation – Best Practices* 414

8 Dazzling Data Designs – Creating Charts 417

8.1 What Is the Right Visualisaton for Me? 417

8.2 Data Visualisation and Python 420

8.2.1 Data Visualisation with Pandas 421

8.2.2 Seaborn 423

8.2.3 Bokeh 425

8.2.4 Plotly 428

8.3 Scatter Plot 430

8.4 Line Chart 438

8.5 Bar Chart 440

8.6 Pie Chart 447

8.7 Histogram 452

8.8 Box Plot 459

8.9 Area Chart 464

8.10 Heatmap 468

A Variance: Population v Sample 477

B Sum of First n Integers 479

C Sum of Squares of the First n Integers 481

D The Binomial Coefficient 483

D.1 Some Useful Properties of the Binomial Coefficient 484

E The Hypergeometric Distribution 485

E.1 The Hypergeometric vs Binomial Distribution 485

F The Poisson Distribution 487

F.1 Derivation of the Poisson Distribution 487

F.2 The Poisson Distribution as a Limit of the Binomial Distribution 488

G The Normal Distribution 491

G.1 Integrating the PDF of the Normal Distribution 491

G.2 Maximum and Inflection Points of the Normal Distribution 493

H Skewness and Kurtosis 495

I Kruskal-Wallis Test – No Ties 497

Bibliography 501

Index 511