

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	<i>Numbers: Integers and Floats</i>	43
2.2.3	<i>Strings</i>	46
2.2.4	<i>Complex Numbers</i>	49
2.3	<i>Collections in Python</i>	51
2.3.1	<i>Lists</i>	52
2.3.2	<i>List Comprehension</i>	60
2.3.3	<i>Tuples</i>	61
2.3.4	<i>Dictionaries</i>	66
2.3.5	<i>Sets</i>	72
2.4	<i>The Beginning of Wisdom: Logic & Control Flow</i>	80
2.4.1	<i>Booleans and Logical Operators</i>	80
2.4.2	<i>Conditional Statements</i>	82
2.4.3	<i>While Loop</i>	85
2.4.4	<i>For Loop</i>	87
2.5	<i>Functions</i>	89
2.6	<i>Scripts and Modules</i>	94

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

99

3.1	<i>Numerical Python – NumPy</i>	100
3.1.1	<i>Matrices and Vectors</i>	101
3.1.2	<i>N-Dimensional Arrays</i>	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	<i>Arithmetic Mean</i>	152
4.3.4	<i>Geometric Mean</i>	155
4.3.5	<i>Harmonic Mean</i>	159
4.4	<i>Dispersion</i>	163
4.4.1	<i>Setting the Boundaries: Range</i>	163
4.4.2	<i>Splitting One's Sides: Quantiles, Quartiles, Percentiles and More</i>	166
4.4.3	<i>Mean Deviation</i>	169
4.4.4	<i>Variance and Standard Deviation</i>	171
4.5	<i>Data Description – Descriptive Statistics Revisited</i>	176
5	<i>Definitely Maybe: Probability and Distributions</i>	179
5.1	<i>Probability</i>	180
5.2	<i>Random Variables and Probability Distributions</i>	182
5.2.1	<i>Random Variables</i>	183
5.2.2	<i>Discrete and Continuous Distributions</i>	185
5.2.3	<i>Expected Value and Variance</i>	186
5.3	<i>Discrete Probability Distributions</i>	191
5.3.1	<i>Uniform Distribution</i>	191
5.3.2	<i>Bernoulli Distribution</i>	197
5.3.3	<i>Binomial Distribution</i>	201
5.3.4	<i>Hypergeometric Distribution</i>	208
5.3.5	<i>Poisson Distribution</i>	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	<i>Linear Correlation and Regression</i>	296
6.4.1	<i>Pearson Correlation</i>	296
6.4.2	<i>Linear Regression</i>	301
6.4.3	<i>Spearman Correlation</i>	308
6.5	<i>Hypothesis Testing with One Sample</i>	312
6.5.1	<i>One-Sample t-test for the Population Mean</i>	312
6.5.2	<i>One-Sample z-test for Proportions</i>	316
6.5.3	<i>Wilcoxon Signed Rank with One-Sample</i>	320
6.6	<i>Hypothesis Testing with Two Samples</i>	324
6.6.1	<i>Two-Sample t-test – Comparing Means, Same Variances</i>	325
6.6.2	<i>Levene's Test – Testing Homoscedasticity</i>	330
6.6.3	<i>Welch's t-test – Comparing Means, Different Variances</i>	332
6.6.4	<i>Mann-Whitney Test – Testing Non-normal Samples</i>	334
6.6.5	<i>Paired Sample t-test</i>	338
6.6.6	<i>Wilcoxon Matched Pairs</i>	342
6.7	<i>Analysis of Variance</i>	345
6.7.1	<i>One-factor or One-way ANOVA</i>	347
6.7.2	<i>Tukey's Range Test</i>	360
6.7.3	<i>Repeated Measures ANOVA</i>	361
6.7.4	<i>Kruskal-Wallis – Non-parametric One-way ANOVA</i>	365
6.7.5	<i>Two-factor or Two-way ANOVA</i>	369

6.8	Tests as Linear Models	376
6.8.1	Pearson and Spearman Correlations	377
6.8.2	One-sample <i>t</i> - and Wilcoxon Signed Rank Tests	378
6.8.3	Two-Sample <i>t</i> - and Mann-Whitney Tests	379
6.8.4	Paired Sample <i>t</i> - 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 Visualisation 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	<i>The Binomial Coefficient</i>	483
D.1	<i>Some Useful Properties of the Binomial Coefficient</i>	484
E	<i>The Hypergeometric Distribution</i>	485
E.1	<i>The Hypergeometric vs Binomial Distribution</i>	485
F	<i>The Poisson Distribution</i>	487
F.1	<i>Derivation of the Poisson Distribution</i>	487
F.2	<i>The Poisson Distribution as a Limit of the Binomial Distribution</i>	488
G	<i>The Normal Distribution</i>	491
G.1	<i>Integrating the PDF of the Normal Distribution</i>	491
G.2	<i>Maximum and Inflection Points of the Normal Distribution</i>	493
H	<i>Skewness and Kurtosis</i>	495
I	<i>Kruskal-Wallis Test – No Ties</i>	497
	<i>Bibliography</i>	501
	<i>Index</i>	511