

# MODERN STATISTICS:

## INTUITION, MATH, PYTHON, R

Dr. Mike X Cohen

---

# 0.1

## Front matter

This page contains some important details about the book that basically no one reads but somehow is always in the first page.

© Copyright 2023 Michael X Cohen.

*All rights reserved. No part of this book may be reproduced or transmitted in any form or by any means, electronic or mechanical, including photocopying, recording, or by any information storage and retrieval system without the written permission of the author, except where permitted by law.*

ISBN: 9798867723736, edition 1.

This book was written and formatted in  $\text{\LaTeX}$  by Mike X Cohen.

# 0.2

## Book cover

The cover of this book, designed by Yuva Oz ([art-4-science.com](http://art-4-science.com)), portrays the synergy of simulated (blue dots) and real (orange dots) data to use statistics as a lens that brings nature's hidden patterns into focus.

# 0.3

## Dedication

If you're reading this, then the book is dedicated to you. I wrote this book for *you*. Now turn the page and start learning statistics!

# 0.4

## Forward

The past is immutable and the present is fleeting. Forward is the only direction.

# Contents

0.1	Front matter . . . . .	2
0.2	Book cover . . . . .	2
0.3	Dedication . . . . .	2
0.4	Forward . . . . .	2
<b>1</b>	<b>Introduction to this book</b>	<b>13</b>
1.1	What is statistics and why learn it? . . . . .	14
1.2	Statistics, data science, machine learning, etc. . . . .	15
1.3	Target audience . . . . .	17
1.4	Prerequisites . . . . .	18
1.5	Exercises . . . . .	19
1.6	Learning from simulated data . . . . .	20
1.7	Using the code with this book . . . . .	21
1.7.1	Which language to use? . . . . .	21
1.7.2	Following along with Python . . . . .	22
1.7.3	Following along with R . . . . .	24
1.7.4	Modifying and reposting my code . . . . .	25
1.8	Online resources . . . . .	25
1.9	AI assistance . . . . .	27
1.9.1	ChatGPT-4 . . . . .	27
1.9.2	Book figures and DALL-E-2 . . . . .	28
<b>2</b>	<b>What are (is?) data?</b>	<b>29</b>
2.1	Is "data" singular or plural? . . . . .	30
2.2	Where do data come from, what do they mean? . . . . .	30
2.3	What do data look like? . . . . .	31
2.4	Limitations of data . . . . .	33
2.5	Accuracy, precision, resolution, range . . . . .	37
2.6	Data types . . . . .	38
2.7	From anecdotes to populations . . . . .	43
2.7.1	Sample vs. population . . . . .	45
2.7.2	"Big enough" samples . . . . .	46
2.7.3	Problems with $N=1$ studies . . . . .	46
2.8	Data management . . . . .	47
2.9	The ethics of making up data . . . . .	48

<b>3</b>	<b>Visualizing data</b>	<b>51</b>
3.1	Why visualize data? . . . . .	52
3.2	How to visualize data . . . . .	52
3.3	Bar plots . . . . .	54
3.3.1	Bar plots for grouped data . . . . .	56
3.3.2	Error bars . . . . .	57
3.4	Pie charts . . . . .	58
3.5	Box plots . . . . .	59
3.6	Histograms . . . . .	60
3.6.1	Histogram vs. bar plot . . . . .	64
3.6.2	Counts vs. proportions . . . . .	65
3.7	Lines vs. bars in a histogram . . . . .	66
3.8	Violin plots . . . . .	68
3.9	Linear vs. logarithmic axis scaling . . . . .	69
3.10	Discretizing continuous data . . . . .	70
3.11	Radial plots . . . . .	71
3.12	Color . . . . .	74
3.12.1	Which colors to use? . . . . .	75
3.13	Exercises . . . . .	77
<b>4</b>	<b>Descriptive statistics</b>	<b>83</b>
4.1	Descriptive vs. inferential statistics . . . . .	84
4.2	Data distributions . . . . .	85
4.2.1	Empirical vs. analytical distributions . . . . .	87
4.2.2	The uses of data distributions . . . . .	89
4.2.3	Examples of distributions . . . . .	90
4.2.4	Quantifying qualitative characteristics . . . . .	92
4.3	Central tendency . . . . .	93
4.3.1	Mean . . . . .	93
4.3.2	Median . . . . .	96
4.3.3	Mode . . . . .	98
4.4	Measures of dispersion . . . . .	99
4.4.1	Variance . . . . .	100
4.4.2	Standard deviation . . . . .	104
4.4.3	Heteroscedasticity and Homoscedasticity . . . . .	105
4.4.4	Full width at half maximum (FWHM) . . . . .	106
4.4.5	Fano factor and CV . . . . .	107
4.5	Interquartile range (IQR) . . . . .	108
4.6	QQ plots . . . . .	109
4.7	Statistical "moments" . . . . .	112
4.7.1	Unstandardized and standardized moments . . . . .	113
4.7.2	First moment: mean . . . . .	113
4.7.3	Second moment: variance . . . . .	114
4.7.4	Third moment: skew . . . . .	115

4.7.5	Fourth moment: kurtosis . . . . .	115
4.7.6	What to memorize . . . . .	117
4.8	Histograms part 2: Number of bins . . . . .	117
4.8.1	Other descriptive stats . . . . .	118
4.9	Exercises . . . . .	120
<b>5</b>	<b>Simulating data</b>	<b>135</b>
5.1	Why simulate data? . . . . .	136
5.2	Random data from distributions . . . . .	137
5.2.1	Normally distributed random data . . . . .	138
5.2.2	Uniformly distributed data . . . . .	140
5.2.3	Random data from other distributions . . . . .	143
5.2.4	Random integers . . . . .	144
5.3	Random elements of a set . . . . .	145
5.4	Random permutations . . . . .	146
5.5	Reproducing randomness . . . . .	149
5.6	Running experiments with random numbers . . . . .	151
5.6.1	Experiment: Impact of standard deviation on mean	152
5.7	The amazing world of data-simulations . . . . .	155
5.8	Finding publicly available real datasets . . . . .	155
5.9	Exercises . . . . .	157
<b>6</b>	<b>Transformations</b>	<b>171</b>
6.1	What, why, and how of data transformations . . . . .	172
6.1.1	What are data transformations? . . . . .	172
6.1.2	Why transform data? . . . . .	172
6.1.3	How to transform data? . . . . .	173
6.1.4	What kinds of transformations are there? . . . . .	174
6.2	Z-score standardization . . . . .	175
6.2.1	Z-score math . . . . .	176
6.2.2	Interpretation . . . . .	177
6.2.3	Hard and soft assumptions . . . . .	178
6.2.4	The modified z-score method . . . . .	180
6.3	Min-max normalization . . . . .	182
6.3.1	Interpretation . . . . .	184
6.4	Z-scoring vs. min-max scaling . . . . .	184
6.5	Percent change . . . . .	185
6.6	Nonlinear data transformations . . . . .	186
6.6.1	Rank-transform . . . . .	187
6.6.2	Logarithm and square root transformation . . . . .	189
6.6.3	Fisher-Z . . . . .	190
6.6.4	Transform any distribution to Gaussian . . . . .	192
6.7	Interpreting transformed data . . . . .	192
6.7.1	When to transform your data . . . . .	194
6.8	Exercises . . . . .	195

<b>7</b>	<b>Assess and improve data quality</b>	<b>203</b>
7.1	Data quality matters . . . . .	204
7.1.1	Data quality influences data-driven decisions . . .	204
7.2	Data cleaning phases . . . . .	205
7.3	Assessing data quality . . . . .	207
7.4	Improving data quality through transformations . . . . .	210
7.5	What are outliers? . . . . .	210
7.5.1	How to think about outliers . . . . .	211
7.6	Identifying outliers . . . . .	213
7.6.1	Absolute threshold detection . . . . .	214
7.6.2	The $z$ -score method . . . . .	214
7.6.3	Iterative $z$ -score method . . . . .	217
7.6.4	Removing data by trimming . . . . .	219
7.6.5	Manual, automatic, and semi-automatic cleaning .	220
7.6.6	What happens to rejected outliers? . . . . .	220
7.7	Analysis-based solutions to outliers . . . . .	221
7.8	Missing data . . . . .	222
7.9	Exercises . . . . .	224
<b>8</b>	<b>Probability theory</b>	<b>231</b>
8.1	From descriptive to inferential statistics . . . . .	232
8.2	What is probability? . . . . .	233
8.2.1	The problem with probability . . . . .	233
8.2.2	When do we need probabilities? . . . . .	235
8.3	Probability vs. proportion . . . . .	236
8.4	Computing probabilities . . . . .	237
8.4.1	Computing analytical probabilities . . . . .	238
8.4.2	Computing empirical probabilities . . . . .	241
8.5	Probability functions, mass, and density . . . . .	243
8.6	Cumulative distribution function (cdf) . . . . .	247
8.7	Expected value . . . . .	249
8.7.1	Computing expected value . . . . .	250
8.7.2	Expected value and statistical moments . . . . .	251
8.8	Softmax . . . . .	253
8.9	Exercises . . . . .	256
<b>9</b>	<b>Sampling and distributions</b>	<b>267</b>
9.1	Sampling variability and its annoyances . . . . .	268
9.1.1	An example with random data . . . . .	269
9.1.2	Where does sampling variability come from? . . .	269
9.2	Creating sample estimate distributions . . . . .	271
9.3	Standard error of the mean . . . . .	273
9.3.1	Standard error of the mean vs. standard deviation	274
9.4	Random and representative sampling . . . . .	275
9.4.1	Independent and identically distributed data . . .	277

9.5	The Law of Large Numbers . . . . .	277
9.5.1	LLN and sample size (LLN demo #1) . . . . .	278
9.5.2	LLN and repeated samples (LLN demo #2) . . . . .	279
9.6	The Central Limit Theorem . . . . .	282
9.6.1	CLT part 1: sampling distributions . . . . .	283
9.6.2	CLT part 2: mixing variables . . . . .	284
9.6.3	The distribution of sample means . . . . .	285
9.6.4	Implications of the CLT . . . . .	286
9.7	Exercises . . . . .	288
<b>10</b>	<b>Hypothesis testing</b>	<b>295</b>
10.1	Hypotheses . . . . .	296
10.1.1	How to specify a hypothesis . . . . .	296
10.1.2	(Why) do we need hypotheses? . . . . .	297
10.1.3	Strong and weak hypotheses . . . . .	298
10.2	IVs, DVs, models, and other stats lingo . . . . .	300
10.3	Can you prove a hypothesis? . . . . .	303
10.4	Sample distributions under $H_0$ and $H_A$ . . . . .	305
10.5	Where do $H_0$ distributions come from? . . . . .	309
10.6	$P$ -values: definition and misinterpretations . . . . .	310
10.6.1	$P$ -values and statistical significance . . . . .	311
10.6.2	$P$ -values and distribution tails . . . . .	312
10.6.3	Where do $p$ -values come from? . . . . .	314
10.6.4	$P$ - $z$ combinations to memorize . . . . .	315
10.6.5	Misinterpretations . . . . .	316
10.6.6	Problems with $p$ -values . . . . .	318
10.7	$P$ -values and significance categorization . . . . .	319
10.8	Type-I and Type-II errors . . . . .	320
10.8.1	The balance of Type-I and Type-II errors . . . . .	321
10.9	Various interpretations of "significant" . . . . .	323
10.10	Multiple comparisons . . . . .	325
10.10.1	Solutions to the multiple comparisons problem . . . . .	326
10.11	Degrees of freedom . . . . .	328
10.12	Exercises . . . . .	330
<b>11</b>	<b>The <math>t</math>-test family</b>	<b>339</b>
11.1	Purpose and interpretation of the $t$ -test . . . . .	340
11.1.1	The purpose of a $t$ -test . . . . .	340
11.1.2	General $t$ -test formula . . . . .	341
11.1.3	Degrees of freedom of $t$ -tests . . . . .	343
11.1.4	$P$ -values from $t$ -values . . . . .	343
11.1.5	$T$ -values from $p$ -values . . . . .	346
11.1.6	Determining significance of a $t$ -test . . . . .	348
11.1.7	Determining significance by critical $t$ -values . . . . .	349
11.1.8	Assumptions of the $t$ -test . . . . .	350

11.1.9	Testing for normality . . . . .	351
11.2	How to make a $t$ -test significant . . . . .	352
11.3	One-sample $t$ -test . . . . .	354
11.4	Two-sample $t$ -tests . . . . .	357
11.4.1	Paired samples $t$ -test . . . . .	357
11.4.2	Independent samples $t$ -test . . . . .	361
11.5	Effect size . . . . .	364
11.5.1	Effect size vs. $t$ -value . . . . .	365
11.6	Nonparametric $t$ -test alternatives . . . . .	366
11.6.1	Wilcoxon signed-rank . . . . .	366
11.6.2	Mann-Whitney U test . . . . .	369
11.6.3	Permutation testing . . . . .	370
11.7	More than two samples? . . . . .	370
11.8	Exercises . . . . .	371
<b>12</b>	<b>Correlations</b>	<b>387</b>
12.1	Motivation and description of correlation . . . . .	388
12.1.1	The correlation coefficient . . . . .	389
12.2	Covariance and correlation: formulas . . . . .	391
12.2.1	Covariance . . . . .	392
12.2.2	"Autocovariance" . . . . .	395
12.2.3	Correlation . . . . .	395
12.3	Correlation matrix . . . . .	397
12.3.1	Linear algebra . . . . .	398
12.4	Correlations in code . . . . .	399
12.5	Assumptions of correlation . . . . .	401
12.6	Simulating correlated data . . . . .	403
12.7	Nonparametric correlations . . . . .	405
12.7.1	The problem with Pearson . . . . .	405
12.7.2	Spearman . . . . .	406
12.7.3	Kendall's correlation for ordinal data . . . . .	407
12.8	Statistical significance . . . . .	408
12.8.1	Fisher- $z$ transformation . . . . .	409
12.9	The subgroups correlation paradox . . . . .	410
12.10	Cosine similarity . . . . .	411
12.11	Exercises . . . . .	415
<b>13</b>	<b>Confidence intervals</b>	<b>427</b>
13.1	Using and interpreting confidence intervals . . . . .	428
13.2	Confidence interval vs. standard deviation . . . . .	430
13.3	Analytical confidence intervals . . . . .	431
13.3.1	Assumptions of analytical confidence intervals . . . . .	433
13.4	Empirical confidence intervals . . . . .	434
13.4.1	Bootstrapping . . . . .	435
13.4.2	Bootstrapping confidence intervals . . . . .	436



13.4.3	Comments and assumptions . . . . .	437
13.5	Confidence intervals & hypothesis testing . . . . .	438
13.5.1	Confidence intervals vs. $p$ -values . . . . .	439
13.5.2	Confidence in confidence intervals . . . . .	440
13.6	Exercises . . . . .	442
<b>14</b>	<b>ANOVA</b>	<b>453</b>
14.1	ANOVA: introduction and overview . . . . .	454
14.1.1	When to use an ANOVA . . . . .	454
14.2	ANOVA terminology . . . . .	455
14.2.1	Factorial design table . . . . .	458
14.2.2	Assumptions of ANOVA . . . . .	459
14.3	The math of the ANOVA . . . . .	460
14.3.1	The ANOVA model . . . . .	461
14.3.2	ANOVA hypotheses . . . . .	461
14.3.3	Sum of squares . . . . .	462
14.3.4	ANOVA as a partition of variability . . . . .	463
14.3.5	Mean square and the $F$ -statistic . . . . .	465
14.4	The ANOVA table . . . . .	467
14.5	Post-hoc comparisons . . . . .	470
14.5.1	Pass the Tukey . . . . .	471
14.5.2	Other post-hoc tests . . . . .	472
14.5.3	When and what to test? . . . . .	473
14.6	Effect size . . . . .	474
14.6.1	Less biased estimators . . . . .	475
14.6.2	Effect size vs. $p$ -value . . . . .	477
14.7	One-way ANOVA example . . . . .	477
14.7.1	ANOVA in Python . . . . .	479
14.7.2	ANOVA in R . . . . .	480
14.8	One-way repeated-measures ANOVA . . . . .	481
14.8.1	Advantages of rmANOVA . . . . .	483
14.8.2	The math of rmANOVA . . . . .	484
14.8.3	Example one-way rmANOVA . . . . .	486
14.9	ANOVA residuals . . . . .	491
14.9.1	Calculate the residuals . . . . .	491
14.9.2	Inspect the residuals . . . . .	492
14.9.3	What to do when the residuals are non-Gaussian? . . . . .	493
14.10	The two-way ANOVA . . . . .	494
14.10.1	Interpreting main effects and interactions . . . . .	494
14.10.2	The math of the two-way ANOVA . . . . .	496
14.10.3	How many ways? . . . . .	499
14.11	"Types" of sums of squares . . . . .	500
14.12	Sphericity and its corrections . . . . .	502
14.12.1	Mauchley's test . . . . .	503

14.13	Simulating data for ANOVAs . . . . .	504
14.13.1	Simulation 1: One-way ANOVA . . . . .	504
14.13.2	Simulation 2: One-way rmANOVA . . . . .	507
14.13.3	Simulation 3: Two-way ANOVA . . . . .	509
14.13.4	Simulation 4: Two-way mixed-effects ANOVA . . .	511
14.13.5	A world of ANOVA explorations . . . . .	512
14.14	Nonparametric ANOVA alternatives . . . . .	512
14.15	Exercises . . . . .	514
<b>15</b>	<b>Regression</b>	<b>525</b>
15.1	Introduction to regression . . . . .	526
15.2	Regression terminology and notation . . . . .	528
15.3	The picture of regression . . . . .	531
15.4	A simple example . . . . .	532
15.5	Least-squares solution to the GLM . . . . .	535
15.5.1	Predicted data and residuals . . . . .	536
15.5.2	Proof of the least-squares equation . . . . .	537
15.6	Evaluating regression models . . . . .	538
15.6.1	Overall model fit . . . . .	539
15.6.2	Comparing "nested" models . . . . .	542
15.6.3	Evaluating individual regressors . . . . .	544
15.7	Standardizing regression coefficients . . . . .	545
15.7.1	Two methods to standardize coefficients . . . . .	547
15.7.2	When to standardize? . . . . .	548
15.8	Regression in Python . . . . .	548
15.8.1	Interpreting the output of <code>sm.OLS</code> . . . . .	550
15.9	Regression in R . . . . .	552
15.9.1	Interpreting the output of <code>lm</code> . . . . .	554
15.10	Simulating data for regression . . . . .	556
15.10.1	Example 1 (one continuous regressor) . . . . .	557
15.10.2	Example 2 (one continuous, one categorical IV) . .	560
15.10.3	Example 3 (two continuous regressors) . . . . .	564
15.11	Assumptions of regression . . . . .	565
15.12	Other regression models . . . . .	567
15.12.1	Weighted regression . . . . .	567
15.12.2	Piecewise regression . . . . .	568
15.12.3	Polynomial regression . . . . .	569
15.12.4	Logistic regression . . . . .	572
15.13	Exercises . . . . .	576
<b>16</b>	<b>Permutation tests</b>	<b>593</b>
16.1	When and why to use permutation testing? . . . . .	594
16.2	Creating an empirical $H_0$ distribution . . . . .	595
16.2.1	One randomized shuffle . . . . .	595
16.2.2	A distribution of shuffled statistics . . . . .	596

16.3	Computing $p$ -values . . . . .	598
16.3.1	$P$ -value based on normalized distance . . . . .	598
16.3.2	$P$ -value based on counts . . . . .	599
16.4	Permutation testing for means . . . . .	600
16.4.1	Permutation testing for a one-sample mean . . . . .	600
16.4.2	Permutation testing for a paired-sample mean . . . . .	602
16.5	Permutation testing for correlation . . . . .	602
16.6	How many permutes? . . . . .	603
16.7	What to permute? . . . . .	605
16.7.1	Permutation world . . . . .	605
16.8	Permutation testing vs. bootstrapping . . . . .	606
16.9	Why not always use permutation testing? . . . . .	607
16.10	Exercises . . . . .	609
<b>17</b>	<b>Power and sample sizes</b>	<b>617</b>
17.1	What is statistical power? . . . . .	618
17.1.1	Statistical power in a graph . . . . .	619
17.1.2	How much power is enough? . . . . .	620
17.2	Estimating statistical power . . . . .	621
17.2.1	Statistical power of a one-sample $t$ -test . . . . .	622
17.3	How to increase statistical power . . . . .	625
17.4	Estimating a required sample size . . . . .	626
17.4.1	Where do the expected values come from? . . . . .	627
17.5	Computing statistical power in practice . . . . .	628
17.5.1	Using <code>statsmodels</code> in Python . . . . .	629
17.5.2	Using <code>pwr</code> in R . . . . .	631
17.5.3	Using G*Power software . . . . .	632
17.6	A priori power vs. post-hoc power . . . . .	634
17.7	Assumptions of power calculations . . . . .	636
17.8	Exercises . . . . .	637
<b>18</b>	<b>Biases</b>	<b>649</b>
18.1	Science vs. the real world . . . . .	650
18.2	Sources of biases . . . . .	651
18.2.1	Unintentional individual biases . . . . .	652
18.2.2	Intentional individual biases . . . . .	654
18.2.3	Culture and tradition biases . . . . .	657
18.3	Conclusions . . . . .	660
18.4	Exercises . . . . .	661
<b>19</b>	<b>Data communication</b>	<b>665</b>
19.1	What is data communication? . . . . .	666
19.2	Tell a story by crafting a data narrative . . . . .	666
19.2.1	What is a data narrative? . . . . .	667
19.3	A few tips . . . . .	668

19.3.1	Generate and resolve conflict . . . . .	668
19.3.2	Humanize previous research . . . . .	669
19.3.3	Highlight the importance of your findings . . . . .	670
19.3.4	Various tips for writing a Results section . . . . .	670
19.3.5	How much to report? . . . . .	671
19.4	Outlets for publishing data . . . . .	672
<b>20</b>	<b>Table of exercises</b>	<b>675</b>
20.1	Table of exercises . . . . .	676