1. Record Nr. UNINA9910880800503321

Autore	Bruce Peter C
Titolo	Statistics for Data Science and Analytics
Pubbl/distr/stampa	Newark : , : John Wiley & Sons, Incorporated, , 2024 ©2025
ISBN	1-394-25383-4 1-394-25382-6 1-394-25381-8
Edizione	[1st ed.]
Descrizione fisica	1 online resource (381 pages)
Altri autori (Persone)	GedeckPeter DobbinsJanet
Lingua di pubblicazione	Inglese
Formato	Materiale a stampa
Livello bibliografico	Monografia
Nota di contenuto	Cover -- Title Page -- Copyright -- Contents -- About the Authors -- Acknowledgments -- About the Companion Website -- Introduction -- Chapter 1 Statistics and Data Science -- 1.1 Big Data: Predicting Pregnancy -- 1.2 Phantom Protection from Vitamin E -- 1.3 Statistician, Heal Thyself -- 1.4 Identifying Terrorists in Airports -- 1.5 Looking Ahead -- 1.6 Big Data and Statisticians -- 1.6.1 Data Scientists -- Chapter 2 Designing and Carrying Out a Statistical Study -- 2.1 Statistical Science -- 2.2 Big Data -- 2.3 Data Science -- 2.4 Example: Hospital Errors -- 2.5 Experiment -- 2.6 Designing an Experiment -- 2.6.1 A/B Tests -- A Controlled Experiment for the Hospital Plans -- 2.6.2 Randomizing -- 2.6.3 Planning -- 2.6.4 Bias -- 2.6.4.1 Placebo -- 2.6.4.2 Blinding -- 2.6.4.3 Beforeafter Pairing -- 2.7 The Data -- 2.7.1 Dataframe Format -- 2.8 Variables and Their Flavors -- 2.8.1 Numeric Variables -- 2.8.2 Categorical Variables -- 2.8.3 Binary Variables -- 2.8.4 Text Data -- 2.8.5 Random Variables -- 2.8.6 Simplified Columnar Format -- 2.9 Python: Data Structures and Operations -- 2.9.1 Primary Data Types -- 2.9.2 Comments -- 2.9.3 Variables -- 2.9.4 Operations on Data -- 2.9.4.1 Converting Data Types -- 2.9.5 Advanced Data Structures -- 2.9.5.1 Classes and Objects -- 2.9.5.2 Data Types and Their Declaration -- 2.10 Are We

Sure We Made a Difference? -- 2.11 Is Chance Responsible? The Foundation of Hypothesis Testing -- 2.11.1 Looking at Just One Hospital -- 2.12 Probability -- 2.12.1 Interpreting Our Result -- 2.13 Significance or Alpha Level -- 2.13.1 Increasing the Sample Size -- 2.13.2 Simulating Probabilities with Random Numbers -- 2.14 Other Kinds of Studies -- 2.15 When to Use Hypothesis Tests -- 2.16 Experiments Falling Short of the Gold Standard -- 2.17 Summary -- 2.18 Python: Iterations and Conditional Execution -- 2.18.1 if Statements.
2.18.2 for Statements -- 2.18.3 while Statements -- 2.18.4 break and continue Statements -- 2.18.5 Example: Calculate Mean, Standard Deviation, Subsetting -- 2.18.6 List Comprehensions -- 2.19 Python: Numpy, scipy, and pandas-The Workhorses of Data Science -- 2.19.1 Numpy -- 2.19.2 Scipy -- 2.19.3 Pandas -- 2.19.3.1 Reading and Writing Data -- 2.19.3.2 Accessing Data -- 2.19.3.3 Manipulating Data -- 2.19.3.4 Iterating Over a DataFrame -- 2.19.3.5 And a Lot More -- Exercises -- Chapter 3 Exploring and Displaying the Data -- 3.1 Exploratory Data Analysis -- 3.2 What to Measure-Central Location -- 3.2.1 Mean -- 3.2.2 Median -- 3.2.3 Mode -- 3.2.4 Expected Value -- 3.2.5 Proportions for Binary Data -- 3.2.5.1 Percents -- 3.3 What to Measure-Variability -- 3.3.1 Range -- 3.3.2 Percentiles -- 3.3.3 Interquartile Range -- 3.3.4 Deviations and Residuals -- 3.3.5 Mean Absolute Deviation -- 3.3.6 Variance and Standard Deviation -- 3.3.6.1 Denominator of N or N-1? -- 3.3.7 Population Variance -- 3.3.8 Degrees of Freedom -- 3.4 What to Measure-Distance (Nearness) -- 3.5 Test Statistic -- 3.5.1 Test Statistic for this Study -- 3.6 Examining and Displaying the Data -- 3.6.1 Frequency Tables -- 3.6.2 Histograms -- 3.6.3 Bar Chart -- 3.6.4 Box Plots -- 3.6.5 Tails and Skew -- 3.6.6 Errors and Outliers Are Not the Same Thing! -- 3.7 Python: Exploratory Data Analysis/Data Visualization -- 3.7.1 Matplotlib -- 3.7.2 Data Visualization Using Pandas and Seaborn -- Exercises -- Chapter 4 Accounting for Chance-Statistical Inference -- 4.1 Avoid Being Fooled by Chance -- 4.2 The Null Hypothesis -- 4.3 Repeating the Experiment -- 4.3.1 Shuffling and Picking Numbers from a Hat or Box -- 4.3.2 How Many Reshuffles? -- 4.3.3 The tTest -- 4.3.4 Conclusion -- 4.4 Statistical Significance -- 4.4.1 Bottom Line -- 4.4.1.1 Statistical Significance as a Screening Device.
4.4.2 Torturing the Data -- 4.4.3 Practical Significance -- 4.5 Power -- 4.6 The Normal Distribution -- 4.6.1 The Exact Test -- 4.7 Summary -- 4.8 Python: Random Numbers -- 4.8.1 Generating Random Numbers Using the random Package -- 4.8.2 Random Numbers in numpy and scipy -- 4.8.3 Using Random Numbers in Other Packages -- 4.8.4 Example: Implement a Resampling Experiment -- 4.8.5 Write Functions for Code Reuse -- 4.8.6 Organizing Code into Modules -- Exercises -- Chapter 5 Probability -- 5.1 What Is Probability -- 5.2 Simple Probability -- 5.2.1 Venn Diagrams -- 5.3 Probability Distributions -- 5.3.1 Binomial Distribution -- 5.3.1.1 Example -- 5.4 From Binomial to Normal Distribution -- 5.4.1 Standardization (Normalization) -- 5.4.2 Standard Normal Distribution -- 5.4.2.1 zTables -- 5.4.3 The 95 Percent Rule -- 5.5 Appendix: Binomial Formula and Normal Approximation -- 5.5.1 Normal Approximation -- 5.6 Python: Probability -- 5.6.1 Converting Counts to Probabilities -- 5.6.2 Probability Distributions in Python -- 5.6.3 Probability Distributions in random -- 5.6.4 Probability Distributions in the scipy Package -- 5.6.4.1 Continuous Distributions -- 5.6.4.2 Discrete Distributions -- Exercises -- Chapter 6 Categorical Variables -- 6.1 Twoway Tables -- 6.2 Conditional Probability -- 6.2.1 From Numbers to Percentages to Conditional Probabilities -- 6.3 Bayesian Estimates -- 6.3.1 Let's

Review the Different Probabilities -- 6.3.2 Bayesian Calculations -- 6.4 Independence -- 6.4.1 Chisquare Test -- 6.4.1.1 Sensor Calibration -- 6.4.1.2 Standardizing Departure from Expected -- 6.5 Multiplication Rule -- 6.6 Simpson's Paradox -- 6.7 Python: Counting and Contingency Tables -- 6.7.1 Counting in Python -- 6.7.2 Counting in Pandas -- 6.7.3 Twoway Tables Using Pandas -- 6.7.4 Chisquare Test -- Exercises -- Chapter 7 Surveys and Sampling.
7.1 Literary Digest-Sampling Trumps "All Data" -- 7.2 Simple Random Samples -- 7.3 Margin of Error: Sampling Distribution for a Proportion -- 7.3.1 The Confidence Interval -- 7.3.2 A More Manageable Box: Sampling with Replacement -- 7.3.3 Summing Up -- 7.4 Sampling Distribution for a Mean -- 7.4.1 Simulating the Behavior of Samples from a Hypothetical Population -- 7.5 The Bootstrap -- 7.5.1 Resampling Procedure (Bootstrap) -- 7.6 Rationale for the Bootstrap -- 7.6.1 Let's Recap -- 7.6.2 Formulabased Counterparts to Resampling -- 7.6.2.1 FORMULA: The Zinterval -- 7.6.2.2 Proportions -- 7.6.3 For a Mean: Tinterval -- 7.6.4 Example-Manual Calculations -- 7.6.5 Example-Software -- 7.6.6 A Bit of History-1906 at Guinness Brewery -- 7.6.7 The Bootstrap Today -- 7.6.8 Central Limit Theorem -- 7.7 Standard Error -- 7.7.1 Standard Error via Formula -- 7.8 Other Sampling Methods -- 7.8.1 Stratified Sampling -- 7.8.2 Cluster Sampling -- 7.8.3 Systematic Sampling -- 7.8.4 Multistage Sampling -- 7.8.5 Convenience Sampling -- 7.8.6 Selfselection -- 7.8.7 Nonresponse Bias -- 7.9 Absolute vs. Relative Sample Size -- 7.10 Python: Random Sampling Strategies -- 7.10.1 Implement Simple Random Sample (SRS) -- 7.10.2 Determining Confidence Intervals -- 7.10.3 Bootstrap Sampling to Determine Confidence Intervals for a Mean -- 7.10.4 Advanced Sampling Techniques -- 7.10.4.1 Stratified Sampling for Categorical Variables -- 7.10.4.2 Stratified Sampling of Continuous Variables -- Exercises -- Chapter 8 More than Two Samples or Categories -- 8.1 Count Data-R×C Tables -- 8.2 The Role of Experiments (Many Are Costly) -- 8.2.1 Example: Marriage Therapy -- 8.3 ChiSquare Test -- 8.3.1 Alternate Option -- 8.3.2 Testing for the Role of Chance -- 8.3.3 Standardization to the ChiSquare Statistic -- 8.3.4 ChiSquare Example on the Computer -- 8.4 Single Sample-GoodnessofFit.
8.4.1 Resampling Procedure -- 8.5 Numeric Data: ANOVA -- 8.6 Components of Variance -- 8.6.1 From ANOVA to Regression -- 8.7 Factorial Design -- 8.7.1 Stratification and Blocking -- 8.7.2 Blocking -- 8.8 The Problem of Multiple Inference -- 8.9 Continuous Testing -- 8.9.1 Medicine -- 8.9.2 Business -- 8.10 Bandit Algorithms -- 8.10.1 Web Testing -- 8.11 Appendix: ANOVA, the Factor Diagram, and the F Statistic -- 8.11.1 Decomposition: The Factor Diagram -- 8.11.2 Constructing the ANOVA Table -- 8.11.3 Inference Using the ANOVA Table -- 8.11.4 The FDistribution -- 8.11.5 Different Sized Groups -- 8.11.5.1 Resampling Method -- 8.11.5.2 Formula Method -- 8.11.6 Caveats and Assumptions -- 8.12 More than One Factor or Variable-From ANOVA to Statistical Models -- 8.13 Python: Contingency Tables and Chisquare Test -- 8.13.1 Example: Marriage Therapy -- 8.13.2 Example: ImanishiKari Data -- 8.14 Python: ANOVA -- 8.14.1 Visual Comparison of Groups -- 8.14.2 ANOVA Using Resampling Test -- 8.14.3 ANOVA Using the FStatistic -- Exercises -- Chapter 9 Correlation -- 9.1 Example: Delta Wire -- 9.2 Example: Cotton Dust and Lung Disease -- 9.3 The Vector Product Sum Test -- 9.3.1 Example: Baseball Payroll -- 9.3.1.1 Resampling Procedure -- 9.4 Correlation Coefficient -- 9.4.1 Inference for the Correlation Coefficient-Resampling -- 9.4.1.1 Hypothesis Test-Resampling -- 9.4.1.2 Example: Baseball Again -- 9.4.1.3 Inference for the Correlation

Coefficient: Formulas -- 9.5 Correlation is not Causation -- 9.5.1 A Lurking External Cause -- 9.5.2 Coincidence -- 9.6 Other Forms of Association -- 9.7 Python: Correlation -- 9.7.1 Vector Operations -- 9.7.2 Resampling Test for Vector Product Sums -- 9.7.3 Calculating Correlation Coefficient -- 9.7.4 Calculate Correlation with numpy, pandas -- 9.7.5 Hypothesis Tests for Correlation -- 9.7.6 Using the t Statistic.
9.7.7 Visualizing Correlation.