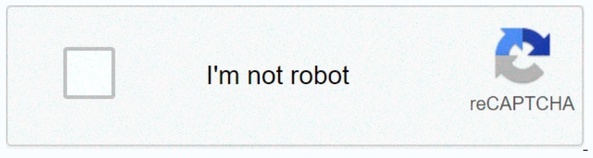

Continue

Practical statistics for data scientists 50 essential concepts table of contents

Statistical methods are a key part of data science, but very few data scientists have formal statistical training. Courses and books on basic statistics rarely cover the subject from a data analysis perspective. This practical guide shows you how to apply different statistical methods to data science, shows how to avoid abuse, and advises on what is important and what is not. Many sources of data analysis include statistical methods, but do not have a deeper statistical perspective. If you know the R programming language and have some statistical exposure, this quick link bridges the gap in an accessible, readable format. With this book, you'll learn:Why exploratory data analysis is a key preliminary step in data sciencehow random sampling can reduce bias and yield a higher quality dataset, even in big dataHow the principles of experimental design yield definitive answers to questionsHow to use regression to estimate results and detect anomaliesKey classification techniques to predict which categories of a record belong toStatistical machine learning methods to learn data Surveillanceless learning methods extracting the meaning of untagged data Peter Founder and scientific director of the Institute for Statistics Education at Statistics.com , which offers about 80 courses in statistics and analysis, of which roughly half are aimed at data scientists. He has written or written several statistical and analytical books and holds a master's degree from Princeton, a master's degree from Harvard and the University of Maryland. Andrew Bruce, lead researcher at Amazon, has more than 30 years of experience in statistics and data science in acade science, government and business. The co-author of Applied Wavelet Analysis at S-PLUS, earned a bachelor's degree from Princeton, and PhD in statistics from the University of Washington Peter Gedeck, senior data scientist at Collaborative Drug Discovery, specializing in developing machine learning algorithms to predict the biological and physico-chemical properties of drug candidates. Co-author of Data Mining for Business Analytics, obtained a PhD in chemistry from the University of Erlangen-Nuremberg in Germany and Mathematics fernuniversität Hagen, Germany table of contents: CoverCopyrightTable contentpreface conventions used in this book using code examples O'Reilly Online Learning How to contact us AccoladesSet 1. Exploratory data analysis elements of structured data further reading rectangular data data frames and indices with non-rectangular data structures further reading estimates place average median and robust estimates Example: Location estimates of population and murder rate further reading estimating variability standard deviation and related estimates alapján percentiles percentiles Variability estimates of the state population further reading exploring the data distribution of percentiles and Boxplots frequency tables and histograms density plots and estimates further reading exploring binary and categorical data method expected value probability of further reading correlation scatterplots further reading exploring two or more variables Hexagonal binning and contours (plotting numeric versus numerical data) Two categorical variables categorical and numerical data visual multiple variables further reading summary Data and sampling distributions random sampling and sample bias bias random selection size versus quality: When does size matter? Sample average population average additional read selection bias regression the average additional reading sampling distribution of the stats central limit item standard error further reading the Bootstrap Resampling Versus Bootstrapping additional read-as-you-go intervals Further reading normal distribution standard standard and QQ plots long-tailed distribution further reading Student's t-Distribution Further reading of bilateral distribution further reading of Chi-Square Distribution Additional Reading F-Distribution Additional Reading Poisson and related distributions Poisson distributions Exponential distribution estimate the error rate Weibull Distribution More reading summary chapter 3. Statistical experiments and significance testing A/B testing Why is there a control group? Why only A/B? Why not C, D,...? More reading hypothesis tests The null hypothesis alternative hypothesis one-way versus two-way hypothesis tests further reading resampling permutation test Example: Web Stickiness exhaustive and Bootstrap permutation tests permutation tests Permutation tests: The data science bottom line further reading statistical significance and p-values p-value Alpha Type 1 and Type 2 errors Data Science and p-values further read t-Tests for more reading more reading of additional reading degrees of freedom further reading ANOVA F-statistic bidirectional ANOVA further reading Chi-Square Test Chi-Square Test: A resampling approach to Chi-Square Test: Statistical Theory Fisher's accurate test relevance to data science further reading multi-arm bandit algorithm for additional reading performance and sample size sample size further reading summary 4. Regression s Zoltan simple linear regression The regression equation is equipped with values and residues Minimum Squares Prediction Versus Explanation (Profiling) More reading more linear regression Example: King County housing data evaluation of the model Cross-Validation model selection and cascading regression weighted regression further reading forecasting using regression hazards Extrapolation confidence and prediction intervals Factor Variables Regression Dummy factors depicting Equation correlated forecasters Multicollinearity disruptive variables interactions and regression diagnostics outlier influential values heteroskedasticity, non-normality, and correlated errors in partial residual plots and nonlinearity polynomial and spline regression polynomial splines generalized additive models further reading summary 5. Classification Naïve Bayesian Why accurate Bayesian classification is impractical The naïve solution of numerical predictive variables further reading discriminatory analysis Covariance Matrix Fisher linear discriminator is a simple example of further reading logistics regression response function and logit Logistics reg regression and GLM generalized linear models predicted values logistical regression interpretation of co-ions and multiplier ratios linear and logistical regression: similarities and differences in evaluation of the model further reading assessment of classification models confusion matrix The rare class problem accuracy, recall, and specificity ROC Curve AUC Lift additional reading strategies with unbalanced data undersampling oversampling and up/down weighting of data generation cost-based classification exploring forecasts for further reading summary 6. Statistical machine learning k-closest neighbors A small example: Predict credit default distance metrics with a hot encoder standardization (normalization, z-Scores) Selecting K KNN as a feature engine tree models is a simple example of a recursive partitioning algorithm measuring homogeneity or dirt stopping the tree from increasing forecasting continuous value How trees are used for further reading dunking and random forest bagging random forest variable importance hyperparameters boost boosting algorithm XGBoost regularization: Avoid overfitting hyperparameters and cross-validation summary section 7. Analysis of the main components of unattended learning A simple example The main components interpretation of the main components correspondence analysis Further reading K-means clusters is a simple example K-means algorithm interpretation of clusters selection of clusters number of hierarchical clusters A simple example Of Dendrogram The agglomerative algorithm measures different model-based clusters with multi-variablenormal distribution Blends normals selecting the number of clusters for further reading scaling and categorical variables scaling the variables dominant variables categorical data and Gower distance problems clusters mixed data SummaryBibliographyIndexAbout the authorsColophon Citation preview nd Se dit E practical statistics data scientists 50 + basic concepts using R and Python Peter Bruce, Andrew Bruce & Peter Gedeck SECOND EDITION Practical statistics data scientists 50+ basic concepts using R and Python By Peter Bruce, Andrew Bruce, and Peter Gedeck In Beijing Boston Farnham Peter Bruce, Andrew Bruce and Peter Gedeck are © 2020 in Tokyo. 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 can also be purchased for educational, business or sales promotional purposes. Online releases are also available in most titles ( ). For more information, contact our corporate/institutional sales department at 800-998-9938 or Editor: Nicole Tache Production Editor: Kristen Brown Copyeditor: Piper Editorial Proofreader: Arthur Johnson May 2017: May 2020: May 2020 2020: Indexer: Ellen Troutman-Zaig Interior Designer: David Futato Cover Designer: Karen Montgomery Illustrator: Rebecca Demarest First Edition Second Edition Revision History for the Second Edition 2020-04-10: First Release See for release details. The O'Reilly logo is a registered trademark of O'Reilly Media, Inc. The views expressed in this work are the views of the authors and do not reflect the views of the publisher. While the publisher and authors have made good faith efforts to ensure that the information and instructions contained in this work are accurate, the publisher and the authors disqualify all liability for errors or omissions, including without limitation liability for damages arising from the use or reliance on this work. The information and instructions contained in this work are used at your own risk. If code samples or other technologies contained or described by the work are subject to open source licenses or other intellectual property rights, it is your responsibility to ensure that their use complies with such licenses and/or rights. 978-1-492-07294-2 [LSI] Peter Bruce and Andrew Bruce want to dedicate this book to the memories of our parents, Victor G. Bruce and Nancy C. Bruce, who cultivated a passion for mathematics and science; and our early mentors John W. Tukey and Julian Simon and our lifelong friend Geoff Watson, who helped us to pursue a career in statistics. Peter Gedeck would like to dedicate this book to Tim Clark and Christian Kramer, a deep thanks to scientific collaboration and friendship. The foreword to the table of contents... ...is that the table of contents is a preface to the table of contents. Whole... xiii 1. Exploratory data analysis. Whole... 1 Elements of structured data further reading rectangular data frames and indexes non-rectangular Structures for additional reading estimates place average median and robust estimates Example: Location estimates of population and murder rates additional reading estimates variability standard deviation and related estimates based on estimates percentiles Example: Variability estimates state population further Reading exploring the data distribution percentiles and boxplots frequency tables and histograms density plots and estimates further reading exploring binary and categorical data mode expected value probability 2 4 4 6 6 7 7 7 9 10 12 3 13 14 14 1 8 19 19 20 22 24 27 29 29 30 v Additional reading correlation scatterplots Further reading exploring two or more variable hexagonal binning and contours (plotting numeric versus numerical data) Two categorical variables for categorical and numerical data visualization of multiple variables additional reading summary 30 30 34 36 36 36 39 41 43 46 26 Data and sampling data... ... And..... and the distribution of sample data... ... and the ... the concreti because of the continams. 47 Random sampling and sample bias bias random selection size versus quality: When does size matter? Sample average population average additional read selection bias regression the average additional reading sampling distribution of the statistics central limit item standard error further reading the Bootstrap Resampling Versus Bootstrapping additional reading confidence intervals further reading Normal Distribution Standard Normal and QQ plots long-tailed distributions additional reading Student's t-Distribution Additional Reading Binomial Distribution Additional Reading Chi-Square Distribution Additional Reading F-Distribution vi | Table of contents 48 50 51 52 53 53 54 55 57 57 60 60 61 61 65 65 65 65 68 69 71 73 75 78 78 8 0 80 8 1 82 Further reading Poisson and related distributions Poisson distributions Exponential distributions estimate the error rate Weibull Distribution Additional reading summary 82 82 83 84 84 85 86 86 3. Statistical experiments and significance tests... ... And during the observations... ... And in the surveillance, the group isn't even my fault. 87 A/B Testing Why is there a control group? Why only A/B? Why not C, D,...? More reading hypothesis tests The null hypothesis alternative hypothesis one-way versus two-way hypothesis tests further reading resampling permutation test Example: Web Stickiness exhaustive and Bootstrap permutation tests permutation tests Permutation tests: The bottom line of data science is further reading statistical significance and p-values p-value Alpha Type 1 and type 2 errors data science and p-values further reading t-tests read more reading degrees of freedom Read Leave Read More reading ANOVA F-Statistic Two-way reading ANOVA More reading Chi-Square Test 88 90 91 92 93 94 95 95 96 96 97 98 102 102 103 106 107 109 109 110 112 112 112 116 116 118 118 121 121 123 124 Table of Contents | vii Chi-Square test: A resampling approach to Chi-square test: test: Theory Fisher accurate test relevance data science further reading multi-armed bandit algorithm for additional reading performance and sample size sample size Additional reading summary 124 127 128 130 131 131 134 135 136 138 139 4. Regression and prediction. 141 Simple linear regression and regression equation equipped with values and residues Minimum Squares Forecast Versus Explanation (Profiling) More reading more linear regression Example: King County Housing Data Assessment model Cross-validation model selection and cascading regression further reading forecasting using regression dangers of extrapolation confidence and prediction intervals Factor variables regression dummy variables representation factor variables on many levels arranged factor variables interpretation regression equation correlated forecasters Multicollinearity disruptive variables interactions and regression diagnostics outliers influential values heteroskedasticity, not normality, and correlated errors viii | Table of contents 141 143 146 148 149 150 150 151 153 155 155 156 159 161 161 161 163 164 167 169 169 170 172 1 72 174 176 177 179 182 Partial residual sample and nonlinearity polynomial and spline regression polynomial splines generalized additive models Additional reading summary 185 187 188 189 192 193 194 5. Classification..... 195 Naïve Bayesian Why accurate Bayesian classification is impractical The naïve solution of numerical predictive variables further reading discriminatory analysis Covariance Matrix Fisher linear discriminator is a simple example of further reading logistical regression response function and log it logistics regression and GLM generalized linear models predicted values logistical regression interpretation of co-ions and multiplier ratios linear and logistical regression: similarities and differences in assessment of the model further reading assessment of classification models confusion matrix The rare class problem accuracy, recall, and specificity ROC Curve AUC Lift additional reading strategies unbalanced data overcalculation oversampling and up/down weighting data generation cost-based classification exploring forecasts 196 197 198 200 201 201 202 203 204 207 208 208 2 008 210 212 212 213 214 216 219 219 221 223 223 223 22222222230 231 232 233 234 234 Table of Contents | ix Read more Summary 236 236 6. Machine Learning statistics. sit-E. 237 K-closest neighbors A small example: Predict credit default distance metrics with a hot encoder standardization (normalization, z-scores) selecting K KNN as a feature engine tree models Simple example of recursive partitioning algorithm measuring homogeneity or contamination stopping the tree from increasing prediction of continuous value How trees are used for further reading dunking and accidental forest bagging random forest variable importance hyperparameters boosting the highlighting algorithm XGBoost regularization: Avoid overfitting hyperparameters and cross-validation summary 238 2 39 24 1 242 243 246 247 249 250 252 254 256 257 258 259 259 260 261 265 269 270 271 272 274 279 282 7. Unattended learning... ..... And..... and learning was witnessed in 2007. 283 Main Component Analysis A simple example computing is the main components interpretation of main components correspondence analysis further reading K-means clusters are a simple example of K-means algorithm interpretation clusters x | Table of Contents 284 285 288 289 292 294 294 295 298 299Section of the number of clusters hierarchical clusters A simple example Of Dendrogram The agglomerative algorithm measures dissent from model-based clusters to multi-variable normal distribution of mixtures in normal selection of the number of clusters further Read scaling and categorical variables scaling the variables at Dominant variables Categorical data and Gower distance problems clustering mixed data summary 302 304 306 308 308 311 311 311 312 315 318 318 318 319 321 322 325 326 Bibliography. ... . Whole. Sitting. 327 The Subject Index .sits .... Sitting. Sitting.. Sitting.. Sitting. ... . Sitting. sitting .sitting .sitting one.. .sitting a... s .ek . .sitting a... i'd like to. 329 Table of Contents | xi Foreword This book is designed for the data scientist with some knowledge of the R and/or Python programming languages, and some previous (perhaps patchy or short) exposure statistics. Two authors came from the world of data science in the world of statistics, and some appreciation for the contribution that statistics can make to the art of data science. At the same time, we are aware of the limitations of traditional statistical education: statistics as a discipline are a century and a half old, and most statistics in textbooks and courses are full of momentum and hitch of an ocean liner. All the methods in the book have some kind of relationship with the discipline of statistics, historical or methodological. Methods that emerged primarily in computing, such as neural nets, are not included. There are two objectives at the bottom of this book: • The layout of key concepts of statistics relevant to data science in digestible, shipable and easily referenced form. • Explain which concepts are important and useful from a data analysis study, which are less important and why. Conventions used in this book are typographical conventions are used in this book: Italic New Terms, URLs, Email Addresses, File Names, and File Extensions. Constant width In paragraphs used for program lists and program saves such as variable or function names, databases, data types, environment variables, instructions, and keywords. xiii Constant Width Bold: Displays commands or other text that the user must type literally. The data science of key terms is a fusion of several disciplines, including statistics, computer science, IT and domain-specific areas. As a result, several different terms can be used to refer to a particular concept. The most important terms and synonyms will be highlighted throughout the book in such a sidebar. This item represents a tip or suggestion. This item is a general comment. This item indicates a warning or warning. Use code examples: In each case, this book gives code examples first r, then python. To avoid unnecessary iterations, we usually only display outputs and plots generated by the R code. It also omits the required code to load packets and datasets. You can find the full code as well as the downloadable datasets https:// github.com/gedeck/practical-statistics-for-data-scientists. This book is here to help you do your job. In general, if a sample code is offered for the book, you can use it in programs and documentation. You don't need to contact us for permission unless you reproduce a significant portion of the code. For example, if you are writing a program that uses multiple snippets of code from this book, you do not need permission. A license to sell or distribute examples from O'Reilly books is required. You do not need permission to answer a question that is referenced by the book and to quote the example code. Significant amount of xiv | Preface example code from this book does not require per-mission documentation of the product. We appreciate, but it is not necessary, attribution. The assignment usually includes the title, author, publisher, and ISBN. For example: Practical statistics data scientists Peter Bruce, Andrew Bruce, and Peter Gedeck (O'Reilly). Copyright 2020 Peter Bruce, Andrew Bruce, and Peter Gedeck, 978-1-492-07294-2. If you feel the code use examples are outside of your fair use or permission given above, feel free to contact us O'Reilly Online Learning For over 40 years, O'Reilly Media has provided technol- ogy and business training, knowledge, and insights to help companies succeed. Our unique network of experts and innovators share their knowledge and expertise through books, articles and our online learning platform. O'Reilly online platform provides on-demand access to live courses, in-depth learning paths, routes, encoding environments, and a huge collection of text and video from O'Reilly and 200+ other publishers. For more information, visit . How to contact us Please address the book's comments and questions with the publisher: O'Reilly Media, Inc. 1005 Gravenstein Highway North Sebastopol, CA 95472 800-998-9938 (U.S. or Canada) 707-829-0515 (international or local) 707-829-0104 (fax) There is a website for the book where we list errata, examples, and any additional information. This page is . E-mail for comment or ask technical questions about this book. For news and courses about our books and courses, please our website. Foreword | xv Find us on Facebook: Follow us on Twitter: Watch us on YouTube: accolades The authors acknowledge the many people who helped make this book a reality. Gerhard Pilcher, ceo of data mining firm Elder Research, saw the book's early drafts and gave us detailed and useful corrections and comments. Similarly, Anya McGuirk and Wei Xiao, statisticians at SAS, and Jay Hilfiger, fellow O'Reilly author, provided useful feedback on the initial drafts of the book. Toshiaki Kurokawa, who translated the first edition of The Japanese, did a comprehensive job of reviewing and improving the process. Aaron Schumacher and Walter Paczkowski carefully reviewed the second edition of the book and provided a number of useful and valuable suggestions for which we are extremely grateful. Needless to say, any mistakes that remain ours alone. At O'Reilly, Shannon Cutt herded us through the publishing process in high spirits and with the right amount of prodding, while Kristen Brown smoothly took our book into the production phase. Rachel Monaghan and Eliahu Sussman improved our writing with care and patience, while Ellen Troutman-Zaig produced the index. Nicole Tache took over the reins of the second edition and both directed the process effectively and provided many good editorial suggestions to improve the readability of the book for a wide audience. We also thank Marie Beau-gureau, who started our project at O'Reilly, as well as Ben Bengfort, O'Reilly's author and Statistics.com instructor, who introduced us to O'Reilly. We and this book have benefited from a lot of conversation Peter has had over the years with Galit Shmueli, co-author of other book projects. Finally, we would particularly like to thank Elizabeth Bruce and Deborah Donnell, whose patience and support have made this effort possible. xvi | Foreword 1: Analysis of exploratory data This chapter focuses on the first step of the data science project: Explore. Classic statistics statistics almost exclusively on the basis of conclusions, sometimes complex procedures to draw conclusions from large populations based on small sam- ples. In 1962, John W. Tukey (Figure 1-1) called for a reformation of statistics in his paper The Future of Data Analysis [Tukey-1962]. He proposed a new scien-tific discipline called data analysis, which included statistical conclusions, like just a ponent. Tukey forged relationships with engineering and computing communities (he coined bit terms, short for binary digits and software), and his original doctrines are surprisingly enduring and part of the foundations of data science. Tukey's 1977 book Explor-atory Data Analysis [Tukey-1977], now a classic, was used to create the field of exploratory data analysis. Tukey presented simple plots (e.g. boxplots, scat-terplots) that, together with summary statistics (average, median, quantization, etc.), help paint a picture of a dataset. With the rapid availability of compute power and expressive data analysis software, exploratory data analysis has evolved well beyond its original scope. The main drivers of this discipline are the rapid development of new technologies, access to more and greater data, and the wider use of quantitative analysis in different disciplines. David Donoho, professor of statistics at Stanford University and a former university student of Tukey, wrote an excellent article based on his presentation at the Tukey Centennial Workshop in Princeton, New Jersey [Donoho-2015]. Donoho tracks the genesis of data science back to Tukey's groundbreaking work in data analysis. 1 1-1. John Tukey, the distinguished statistician whose ideas were developed more than 50 years ago, form the basis of data science elements of structured data derived from many sources: sensor measurements, events, text, images, and vid-eos. The Internet of People (IoT) broadcasts streams of information. Most of this data is unstructured: images are collections of pixels, and each pixel contains RGB (red, green, blue) color information. Text is a series of words and non-literal endowrs, often sections, subsections, and so on. Click streams are sequences of actions performed by a user who interacts with an app or web page. In fact, one of the main challenges of data science is to use this torrent of raw data for usable infor-mation. For the purposes of applying the statistical concepts contained in this book, unstructured raw data must be processed and manipulated in a structured form. One of the most common forms of structured data is a table of rows and columns because the data can come from a relational database or be collected for a study. Structured data has two basic types: numerical and categorical. The numerical are in two forms: continuous, such as wind speed or duration, and discrete, such as Event. Categorical data only records fixed values, such as a TV screen type (plasma, LCD, LED, etc.) or a state name (Ala-bama, Alaska, etc.). Binary data is an important special case of categorical data that takes only one of two values, such as 0/1, yes/no, or true/false. Another useful type of categorical data is the interline data in which the categories are sorted; an example is numerical evaluation (1, 2, 3, 4 or 5). Why do we deal with taxonomy of data types? It turns out that for data analysis and predictive modeling, the data type is important for determining the type of visualization, data analysis, or statistical model. In fact, data analytics software, such as R and Python, uses these data types to improve compute/formance. More importantly, the data type of a variable determines how the software will handle the calculations of that variable.

Chapter 1 2004-200: Analysis of exploratory data Key conditions for data types expressed on a numerical scale. Continuous data that can contain any value in an interval. (Synonyms: interval, floating, numeric) Separate data that can only take inteer values, such as counts. (Synonyms: integer, number) Categorical data that can only add a set of values that represent a set of possible categories. (Synonyms: enumeration, listed, factors, nominal) Binary Is a special case of categorical data that contains only two value categories, e.g. 0/1, true/false. (Synonyms: dichotomous, logical, pointer, logical) Row system categorical data that is in explicit order. (Synonym: orderly factor) Software engineers and database programmers may wonder why we need the concept of categorical and interline data for analysis. After all, categories are just a collection of text (or numeric) values, and the underlying database automatically contains the internal representation. However, explicit identification of data other than text as categorical offers some benefits: • Knowing that the data is categorical can serve as an indication to the software of how statistical procedures, such as charting or model matching, should behave. Par-ticular, ordinal data can be represented as an ordered.factor in R, preserving user-defined order charts, tables, and models. In Python, scikit-learn supports line data at sklearn.preprocessing.OrdinalEncoder.OrdinalEncoder. • Storage and indexing can be optimized (as in a relational database). • The possible values of a particular categorical variable can be applied to soft-ware (like an enum). The third benefit for unwanted or unexpected behavior the default behavior of R's data import functions (e.csv read) is that it automatically converts the text column to a factor. Subsequent operations in this column assume that only the originally imported, and assigning a new text value introduces a warning and creates an NA (missing structured data elements | 3 values). The Python panda package will not automatically convert to this. However, read_csv can explicitly specify a column categorically in the function. Key Ideas • Data is generally classified by type in software. • Data types include numeric (continuous, discrete) and categorical (binary, serial number). • The data typed in the software is used to indicate to the software how the data is processed. Read more • Panda documentation describes different data types and how to manipulate them in Python. • Data types can be confusing as the types can overlap and one software taxonomy may be different from another. The R tutorial website contains the R taxonomy. The panda documentation describes the different types of data and how to manipulate them in Python. • Databases are more detailed in classifying data types, which includes considerations of accuracy levels, fixed or variable length fields, and so on; see the W3Schools guide to SQL. Rectangular data The typical frame of reference for data analysis is a rectangular data object, such as a spreadsheet or database table. Rectangular data is a generic expression of a two-dimensional matrix that uses columns to indicate records (cases) and characteristics (variables) that indicate rows; data frame in the spe-cific format R and Python. Data does not always start in this form: unstructured data (e.g. text) must be processed and manipulated to appear as a set of characteristics in rectangular data (see Structured data elements on page 2). Data in relational databases must be ed and placed in a single table for most data science and modeling tasks. Chapter 2004-200 Chapter 1: Exploratory data analysis Key terms for rectangular data frame rectangular data (such as a spreadsheet) are the basic data structure of statistical and machine learning models. Column A within a table is usually called a service. Synonyms attribute, input, forecasting, variable Result Many data science projects include predicting the result – often yes/no out – coming (in Table 1-1, this auction was competitive or not). The features are used a few times to predict the outcome of an experiment or a study. Variables that depend on synonyms, response, destination, outputRecord a row in a table is usually called a record. Synonyms of case, for example, observation, pattern, pattern table 1-1. A typical data frame format Category currency sellerRating Duration endDay ClosePrice OpenPrice Competitive? Music/Film/Game US 3249 5 Mon 0.01 0.01 0 US3249 5 Mon 0.01 0.01 0 Automotive US 3115 7 Tue 0.01 0 Automotive US 3115 7 Tue 0.01 0.01 0 Automotive US 3115 7 Tue Tue 0.01 0 Automotive US 3115 7 Tue 0.01 0.01 0 Automotive US 3115 7 Tue 0.01 0.01 1 Automotive US 3115 7 Tue 0.01 0.01 1 1-1. As mentioned earlier, the categorical variable is concretised as binary (yes/no or 0/1), which is the variable 1-1. This indicator variable is also a result variable if the scenario is to predict whether the auction is competitive or not. Rectangular data | 5 Data frames and indexes Traditional database tables mark one or more columns as indexes, a serial number. This can greatly improve the efficiency of certain database queries. In Python, with the panda directory, the basic rectangular data structure is a DataFrame object. By default, an automatic ink index is created for the DataFrame based on the order of the rows. Pandas can also have multi-level/hieró system indices to improve the efficiency of certain operations. In R, the basic rectangular data structure is a data.frame object. Data.frame also has an implicit ink based on the row order. Native R data.frame does not support user-defined or multilevel indexes, although a custom key can be created through the row.names attribute. To overcome this shortcoming, two new packages are widespread: data.table and dplyr. Both support multilevel indexes and offer significant acceleration in working with data.frame. Terminology differences The terminology of rectangular data can be confusing. Statisticians and data scientists use different terms for the same thing. For a statician, predicting variables are used in the model to predict the response or dependent variable. As a data scientist, features can help you predict your goal. One of the synonyms is particularly disturbing: com-puter scientists will use the term sample in a single line; represents a collection of sample lines that have been given to the statistician. Non-rectangular data structures In addition to rectangular data, there are other data structures. Successive measurements of the same variable are recorded in the time series data. It is the raw material for statistical forecasting methods and a key component of the data generated by the devices – the Internet of Things. Spatial data structures used in mapping and location analysis are more complex and diverse than rectangular data structures. In object visualization, the data focuses on an object (such as a house) and its spatial coordinates. In contrast, the field view focuses on small 3D and the value of the corresponding metric, such as pixel brightness. Chapter 1 2004-200: Exploratory data analysis graph (or network) data structures physical, social and abstract relationships. For example, a graph on a social network, such as Facebook or LinkedIn, can represent relationships between people on the network. Road-connected hubs are examples of a physical network. Graph structures are useful for certain types of problems, such as network optimization and referral systems. Each of these data types has a specific methodology for data analysis. The focus of the book is on rectangular data, the basic building block of predictive modeling. Graphs of statistics in computing and IT expression graphs typically refer to a representation of relationships between entities, and the underlying data structure. In statistics, the graph refers to different plots and visualizations, not just relationships between entities, and the expression applies only to the visual, not to the data structure. Main ideas • The basic data structure for data analysis is a rectangular matrix in which rows are records and columns are variables (characteristics). • Terminology can be confusing; there are different synonyms stemming from different disciplines that contribute to data science (statistics, computing, and information technology). Read more • Documentation on data frames in r • Documentation for data frames with measured or meaiable data in Python location variables can have thousands of different values. The basic step in exploring data is to get a typical value for each characteristic (variable): it gives an estimate of where most of the data is located (i.e. its central trend). Estimation of the location | 7 The main criteria for estimating location are the ratio of total values to number of values. Synonym average Weighted average: The product of the weight product is the sum of all values and weights. Synonym weighted average Median Is the value that is above and below half of the data. Synonym 50 percentile The value so that p percent of the data is below. Synonym quantity weighted median The value is that half of the sum of weights is above and below ordered data. Cut Average: The average of all values after discarding a specific number of extreme values. Synonym truncated means robust non-sensitive to extreme values. Synonymous with outlier data value, which is very different from most data. Synonym extreme value 8 | Chapter 1: Exploratory data analysis At first glance, the summary of the data may seem rather trivial: just take the average of the data. In fact, while the average is easy to calculate and advisable to use, it may not always be the best measure of the central value. Therefore, statisticians have developed and supported a number of alternative estimates compared to the average. Statisticians of metrics and estimates often use the term estimation in a to distinguish between the data and the theoretical real or accurate situation. Data analytics data and business analysts are more likely to refer to such a value as an indicator. The difference reflects the approach to statistics and data science: accounting for uncertainty is at the heart of the science of statistics, while data science focuses on specific business or organisational objectives. Therefore, statisticians estimate, and data scientists measure. Average The most basic estimate of the location is the average or average value. The average is the quotient of all values to the number of values. Consider the num-bers set {3 5 1 2}. The average (3 + 5 + 1 + 2) / 4 = 11 / 4 = 2.75. It meets the symbol x (pronounced x-bar), which is used to represent the average of a sample taken from a population. Formula for calculating the average of x1, x2, ..., xn n values: Average = x = ∑ni=1 x n N (or n) refers to the total number of records or observations. In statistics, it is capitalized if you refer to a population and lowercase if you refer to a sample from a population. In data science, this distinction is not vital, so you can see it in both directions. The change in average is a cut mean, calculated by throwing a specified number of selected values at each end and then taking the average of the remaining val-ues. The formula for calculating the cropped mean value is the formula p with the minimum and maximum omitted values: ∑ni =− pp + 1 x i Cut average = x = n − 2p Location estimate | 9 The cut mean eliminates the effect of extreme values. For example, in international diving, the highest and lower scores of five judges are dropped, and the final score is the average of the scores of the remaining three judges. This makes it difficult for a single judge to manipulate the score, perhaps to prioritize the country's contenders. Cut tools are widely used and in many cases prefer to use the usual average – see Median and robust estimates in section 10. Another average type is a weighted mean value, calculated by multiplying each xi data value by the user-specified weightwi and splitting the amount by the sum of the weights. The weighted mean formula: Weighted mean = xw = ∑ni = 1 wixi ∑ni = 1 wi There are two main motivations for using the weighted mean: • Some values are inherently more variable than others, and highly variable observations have less weight. For example, if you take the average of multiple sensors and one sensor is less accurate, you may be weighed down by the sensor data. • The data collected does not represent the same groups that interested in the measurement. For example, because of an online experiment, we may not have a dataset that accurately reflects all groups in the user base. To correct this, you can add more weight to the values of underrepresented groups. Median and robust estimates The median is the middle number of an ordered list of data. If data values have even numbers, the middle value is not actually in the dataset, but is the average of the two values that divide the sorted data into upper and lower parts. The median depends only on the values in the middle of the ordered data compared to the average using all observations. While this may seem to be a disadvan-tage, since the average is much more sensitive to data, in many cases when the median is a better metric of location. Let's say you want to look at typical household incomes in neighborhoods around Lake Washington in Seattle. With com-paring in the Medina neighborhood in the Windermere neighborhood, the average use has very different results because Bill Gates lives in Medina. If we use the median, then no matter how rich Bill Gates's situation in the middle observation remains the same. 10000000 | Chapter 1: Exploratory data analysis For the same reason we use a weighted middle ground, a weighted median can be calculated. The median is also sorted first, although each data value has weight. Instead of the middle number, the weighted median is the sum of the weights for the upper and lower half of the multiplied list. Like the median, the weighted median is robust and outlier. Outliers The median is called a robust estimate of the location because it is not affected by outliers (extreme cases) that can distort results. An outlier is any value that is very far from the other values in the dataset. The exact definition of an outlier is subjective, although some conventions are used in different data breakdowns and plots (see Percentiles and Boxplots on page 20). Since the outlier alone does not make the data value invalid or incorrect (as in the previous example by Bill Gates). However, outliers are often the result of data errors, such as mixing data from different units (kilometers and meters) or poor readings from the sensor. If the outliers are the result of bad data, the average results in a poor estimate of the location estimate, while the median remains valid. In any case, outliers should be determined and generally worthy of further investigation. Anomaly detection Unlike typical data analysis, where outliers are sometimes informative and sometimes unpleasant, anomaly detections are outliers, and the larger mass of data is primarily normal what anomalies are measured. The median is not the only reliable estimate of the site. In fact, the cut poses are widely used to avoid the influence of outliers. For example, if you trim the top and bottom 10% of the data (common choice), it always protects against outliers except for the smallest datasets. The cut mean is considered a trade-off between the median and the average: robust and extreme values in the data, but uses more data to calculate the location estimate. Other robust metrics for positioning statisticians have developed a number of other estimates for location, primarily with the aim of developing a more robust estimate than average and more efficient (i.e. better distinguishing small local differences between data sets). While these methods are potentially useful for small datasets, they are unlikely to provide additional benefits for large or even medium-sized datasets. Estimation of the location | 11 Example: Table 1-2 of location estimates for population and homicide rates shows the first few lines of the population and homicide rate dataset in each US state (2010 census) (2010 census). Table 1-2. Some lines of data.frame state population and murder rate in the state 1 Alabama Population Murder Rate abbreviation 4,779,736 5.7 AL 2 Alaska 710,231 5.6 AK 3 Arizona 6,392,017 4.7 AZ 4 Arkansas 2.91 5,918 5.6 AR 5 California 37,253 956 4.4 CA 6 Colorado 5,029,196 2.8 CO 7 Connecticut 3,574,097 2.4 CT 8 Delaware 5.8 DE 897,934 Calculate the average , cut average, and median the population using r: > state average(state[[population]] [1] 6162876 > average(state[state[population], trim=0.1) [1] 4783697 > median(state[Population]]] [1] 4436370 Calculation average and median Python we can use panda methods in frame data. The cut mean requires the trim_mean function in scipy.stats: status = pd.read_csv('state.csv') state[''Population'']mid()) trim_mean(state[Population], 0.1) state[Population].median() The average is greater than the cut average, which is greater than the median. This is because the severed mean excludes the maximum and minimum five states (trim=0.1 decreases by 10% from both ends). If we want to calculate the country's average murder rate, we need to use a weighted average or median to take into account the different populations of the states. Since the base R does not have a function in the weighted median, you must install a package such as matrixStats: > weighted.mean(state[Murder.Rate]], w= state[[Population] [1] 4,445834 > directory(matrixStats) 12 | Chapter 1: Exploratory Data Analysis > Weighted Median(State[[Murder.Rate]], w=state[[Population]]) [1] 4.4 Weighted available with NumPy. For the weighted median, you can use the advanced illiterate package: np.average(state[Murder.Rate], weights=state['Population']) weights=state['Population']) weights=state['Population']) In this case, the weighted mean and the weighted median are approximately the same. Key ideas • The basic metric of the place is average, but it can be sensitive to extreme values (outliers). • Other indicators (median, cut average) are less sensitive to outliers and unusual distributions, making them more robust. Read more • Wikipedia's article on the central trend contains a wide-ranging discussion of different location measurement measures. • John Tukey's 1977 classic exploratory data analysis (Pearson) is still widely read. Variability Location estimates only one dimension in a feature summary. The second dimension, varia-bility, also known as dispersion, measures whether data values are closely collussed or distributed. Statistics focus on variability: measuring, reducing it, distinguishing between random and real variability, identifying different sources of real variability and making decisions in its presence. The most important conditions for variability indicators Are differences between observed values and location estimates. Synonyms errors, residual Variance Sum of the square deviations between average and n -1, where n is the number of data values. Estimation of variability | 13 Synonym mean-square-error Standard deviation The square root of variance. Mean absolute deviation The mean value of absolute values of deviations from the mean value. Synonyms l1-norm, Manhattan norm Median absolute deviation from median Median Median absolute values of deviations from median. Range: The difference between the largest and smallest values in the dataset. Order statistics Metrics based on data values sorted from smallest to largest. Synonym prioritizes Percentile The value to P percentage of the values, whether this value or less, and (100-P) percent takes this value or more. Synonymous quad quad sheet range The difference is 75. Synonym IQR As there are different ways to measure location (average, median, etc.), there are also different ways to measure variability. Standard deviation and related estimates The most commonly used change estimates are based on differences or differences between location estimates and observed data. For data set {1, 4, 4}, the average is 3 and the median is 4. The deviations from the average are as follows: 1 to 3 = -2, 4 to 3 = 1,4 to 3 = 1 . These differences show how scattered the data is around the central value. 14 | Chapter 1: Exploratory data analysis One way to measure variability is to estimate the typical value of deviations. Averaging deviations doesn't tell us much – negative variations offset the positives. In fact, the sum of the deviations from the average is exactly zero. Instead, a simple the average of absolute values of deviations from the average shall be taken into account. In the previous example, the absolute value of the difference is {2 1 1}, and their average is (2 + 1 + 1) / 3 = 1.33. This is called the mean absolute deviation and is calculated by the formula ∑in = 1 xi − x Mean absolute deviation = n, where x is the mean value of the sample. The best-known estimates of variance are variance and standard deviation, which are based on square deviations. Variance is the average of square deviations, and standard deviation is the square root of variance: Variance Standard deviation = s2 = ∑ni = 1 xi − x n−1 2 = s = variance The standard deviation is much easier to interpret than variance, as it is on the same scale as the original data. Still, the more complex and less intuitive of the mula, it may seem odd that the standard deviation is preferred in statistics than the average absolute deviation. It owes its preeminence to statistical theory: mathemati-cally, working with square values, is more convenient than absolute values, especially for statistical models. Freedom and n or n - 1? In statistical books, there is always some debate about why there is n - 1 in the denominator of the variance formula, instead of n, which leads to the concept of a defeiph. This distinction is not important, since n is usually large enough that it will not make much difference whether it divides n or n - 1. But in case you're interested, here's the story. This is based on the assumption that you want to mate with the population based on a sample. If you use the intuitive denominator of n in the variance formula, it will be below the true value of the variance and the standard deviation of the population. This is called a biased estimate. However, if you divide the variance by n to 1 instead of n, the difference becomes an unbiased estimate. Estimation of variability | 15 A full explanation of why the use of n leads to biased estimation includes the concept of degree of freedom, which takes into account the limitations of esti-mate calculation. In this case, there is a freedom of n - 1, because there is a squeeze: the standard deviation depends on the calculation of the sample center. For most problems, data scientists don't have to worry about the freedom. Neither variance, standard deviation, nor average absolute deviation are robust for outliers and extremes (see median and robust estimates to discuss reliable estimates of location). Variance and standard deviation are particularly sensitive to outliers as they are based on square deviations. Robust estimation of variability at median or absolute mean from MAD: Absolute median deviation = Median x1 − m, x2 − m, xN − m, where m is the median. Like the median, MAD is not affected by extreme val-ues. It is also possible to calculate a cut standard deviation similar to the cut average (see average on page 9). Variance, standard deviation, mean absolute deviation and median absolute deviation are not equivalent estimates, even if the data are derived from a distribution with nor. In fact, the standard deviation is always greater than the average absolute deviation, which itself is greater than the median absolute deviation. Sometimes the absolute median deviation is multiplied by a constant scaling factor to place the MAD on the same scale as the standard deviation for a normal dis-tribution. The commonly used factor of 1.4826 means that 50% of the normal distribution is within the ±MAD range (see e.g. . Estimates based on percentiles A different approach to the estimation of dispersion is based on an examination of the spread of orderly data. Statistics based on sorted (ranked) data are called order statistics. The most basic measure is the range: the difference between the largest and smallest numbers. The minimum and maximum values themselves are useful for identifying outliers, but the range is highly sensitive to outliers and is not very useful as a general measure of data dispersion. To avoid the sensitivity of outliers, you can view the range of data after discarding the values from both ends. Officially, these types of estimates are based on the 16th century . Chapter 1: Exploratory data analysis among percentages. In a dataset, P-percentile is a value that takes at least P percent of the values or less, and at least (100 to P) percent of the values take that value or more. For example, if you want to find your 80th birthday, you can use the following information: Then, starting with the smallest value, continue to 80 percent of the way to the maximum value. Note that the median is the same as the 50th. The percentile is essentially the same as the quantiles, indexing the quantiles with fractions (so the .8 quan-tile is the same as the 80th percentile). Frequent measurement of variability is carried out in the 25th and 26th Here is a simple exam-ple: {3,1,5,3,6,7,2,9}. These are sorted to get {1,2,3,3,5,6,7,9}. The 25 percentile is 2.5 and the 75 percentile is 6.5, so the interquartile range is 6.5 to 2.5 = 4. Software may have slightly different approaches that provide different answers (see next tip); these differences tend to be smaller. For very large datasets, calculating exact percentiles can be very expensive because it requires sorting all data values. Machine learning and statistics speciális algoritmusokat használ, mint például a [Zhang-Wang-2007], hogy egy hozzávetőleges percentilist kapjon, amely nagyon gyorsan kiszámítható, és garantáltan bizonyos pontossággal rendelkezik. Percentilis: Pontos meghatározás Ha páros számú adatunk van (n páros), akkor a percentilis nem egyértelmű az előző definíció szerint. In fact, we could take on any value between the order statistics x j and x j + 1 where j satisfies: 100 * j j+1 ≤ P < 100 * n n Formally, the percentile is the weighted average: Percentile P = 1 − w x j + wx j+1 for some weight w between 0 and 1. Statistical software has slightly differing approaches to choosing w. In fact, the R function quan tile offers nine different alternatives to compute the quantile. Except for small data sets, you don't usually need to worry about the precise way a percentile is calculated. At the time of this writ-ing, Python's numpy.quantile supports only one approach, linear interpolation. Estimates of Variability | 17 Example: Variability Estimates of State Population Table 1-3 (repeated from Table 1-2 for convenience) shows the first few rows in the data set containing population and murder rates for each state. Table 1-3. A few rows of the data.frame state of population and murder rate by state State 1 Alabama Population Murder rate Abbreviation 4,779,736 5.7 AL 2 Alaska 710,231 5.6 AK 3 Arizona 6,392,017 4.7 AZ 4 Arkansas 2,915,918 5.6 AR 5 California 37,253,956 4.4 CA 6 Colorado 5,029,196 2.8 CO 7 Connecticut 3,574,097 2.4 CT 8 Delaware 5.8 DE 897,934 Using R's built-in functions for the standard deviation , the interquartile range (IQR), and the median absolute deviation from the median (MAD), we can compute esti- mates of variability for the state population data: > sd(state[['Population']]) [1] 6848235 > IQR(state[['Population']]) [1] 4847308 > mad(state[['Population']]) [1] 3849870 The pandas data frame provides methods for calculating standard deviation and quantiles. A kvantitatív a könnyen meghatározható az IQR. A robusztus MAD esetében a robust.scale.mad függvényt használjuk a statsmodels csomagból: state['Population']std() state['Population'].quantile(0.75) - state['Population'].quantile(0.25) robust.scale.mad(state['Population']) A szórás majdnem kétszer akkora, mint a MAD (az R-ben alapértelmezés szerint a MAD léptéke az átlaggal azonos mértékben van beállítva). Ez nem meglepő, mivel a szórás érzékeny a kiugró értékekre. 18 | 1. fejezet: Feltáró adatelemzésI legfontosabb ötletek • A variancia és a szórás a legelterjedtebb és rutinszerűen jelentett változékonysági statisztikák. • Mindkettő érzékeny a kiugró értékekre. • A robusztusabb mutatók közé tartozik az átlagos abszolút eltérés, a mediántól való abszolút középérték és a percentilisek (kvantikumok). További olvasás • David Lane online statisztikai forrás egy szakaszt percentiles. • Kevin Davenport useful posts from R-Bloggers on the variations in median and robust properties. Explore the distribution of data Each of the estimates we cover sums the data in a single number to describe the location or variability of the data. It's also helpful to explore how the data is dis-tributed overall. Key conditions for exploring Distribution Boxplot's site are introduced by Tukey as a quick way to visualize the distribution of data. Synonym field and mustache visualization Frequency is the number of numerical data values in the table A that can be classified as intervals (bins). Histogram Plot of the frequency plate with bins on the x-axis and counter on the y-axis (or pro-part). Although visually similar, although diagrams should not be confused with histograms. See exploring binary and categorical data page 27 for a discussion on the difference. Density pattern A smooth version of the histogram, often based on kernel density estimates. 100000 | At 19 Percentiles and Boxplots The Estimates, based on Percentiles page 16, have investigated how percentiles can be used to measure the spread of data. Percentages are also valuable as a summary of the total allocation. It is common to report quartiles (25th, 50th, and 75th percentiles) and deciles (the 10th, 20th, ..., 90th percentiles). Percentiles is the value used to sum the tail (outer range) of the distribution. Popular culture coined the term that one percent is the 99th percent of wealth. Table 1-4 shows a percentage of homicide rates. In R, this is produced by the quantile function: quantile(state[['Murder.Rate']]] p=c(.05, .25, .5, .75, .95)) 5% 25% 50% 75% 95% 1,600 2,425 4,000 5,550 6,510 Pandas' data frame method quantile in Python provides: state['Murder.Rate'].quantile([0.05, 0.25, 0.5, 0.75, 0.95]) 1-4. Percentiles murder rate in the state 5% 25% 50% 75% 95% 1.60 2.42 4.00 5.55 6.51 The median 4 murders of 100,000 people, although there is a bit of variabil-ity: the 5 percentile is only 1.6 and the 95 percentile is 6.51. Boxplots, introduced by Tukey [Tukey-1977], are based on percentiles and give a quick way to visualize the distribution of data. Figure 1-2 shows the boxplot of the pop-ulation state produced by R: boxplot(state[[Population]]/1000000, ylab ='Population (million)') pandas provide a number of basic exploratory plots of data frame; one of these is boxplots: axe = (state[Population/1_000_000).plot.box() ax.set_ylabel(Population (million)) 20 | Chapter 1: Exploratory data analysis(1) to (2) Boxplot the state population Of this boxplot can immediately see that the median state population is about 5 million, half of the states fall between about 2 million and about 7 million, and there are some high Outliers. The top and bottom of the box are 75 and 25 respectively. The median is shown by the horizontal line in the box. Dashed lines, referred to as mustaches, extend from the top and bottom of the box to indicate the range of most of the data. There are several variants of boxplot; see, for example, the boxplot documentation for function R [R-base-2015]. By default, the R function extends the moustache to the furthest point beyond the box, with the difference that it does not exceed 1.5 times the IQR. Matplotlib uses the same urge; other software may use a different rule. Data outside the moustache is represented as a point or circle (often consid-derived outliers). 100000 | 21 Frequency tables and Histograms The variable frequency table divides the variable range into a seg-ments at an equal distance and shows how many values belong to each segment. The following shall be kc_tax 1-5: SqFtTotLiving < 3500), :] kc_tax0.shape (432693, 3) Figure 1-8 is a hexagonal binning plot with a relationship between the finished square meter and the tax-valued value of apartments in King County. Instead of plotting points that appear as monolithic dark clouds, records are grouped into hexagonal containers and hexagons are represented in one color, indicating the number of records in the warehouse. In this table, the positive relationship between the square metre and the value assessed by the tax is clear. An interesting feature is the hint of additional lanes above the main (darkest) bar at the bottom, indicating that the apartments are the same square meters as the main lane, but with a higher tax-valued value. Figure 1-8 was created by the powerful R package ggplot2 developed by Hadley Wickham [ggplot2]. Ggplot2 is one of several new software libraries for advanced exploratory visual analysis of data; see Multiple variables display the scale_fill_gradient theme_bw stat_binhex kc_tax0 43rd hexagonal binning plots readily available in pandas data frame method hexbin: ax = kc_tax0.plot.hexbin(x = 'SqFtTotLiving', y ='TaxAssessedValue', gridsize=30, sharex=False, figsize=(5, 4)) ax.set_xlabel(Finished Leg) ax.set_ylabel(Tax Valued) Explore two or more variables | Figure 1-8. Hexagonal binning is the valued by the tax and finished square foot Figure 1-9 uses contours onto a scatterplot to visualize the relationship between the two numeric variables. Contours are essentially topographic maps of two variables; each contour bar indicates the specific density of the points and approximates one of them near the tip. This site a similar story to the one-eight. This diagram was also created with ggplot2 with the built-in geom_density2d function: ggplot(kc_tax0, aes(SqFtTotLiving, TaxAssessedValue)) + theme_bw() + geom_point(alpha=0.1) + geom_density2d(color='white') + labs(x='Finished Square Feet', y='Tax-Assessed Value') The seaborn kdeplot function python creates a contour plot: ax = sns.kde(kc_tax0. SqFtTotLife, kc_tax0. TaxAssessedValue, ax=ax) ax.set_xlabel(Finished Square Foot) ax.set_ylabel(Tax Value) 38 | Chapter 1: Exploratory data analysis (i) Figures 1 to 9 The tax-valued value and the finished square foot contour image Other types of charts are used to display the relationship between two numerical variables, including heat maps. Heat maps, hexagonal binning, and contour plots all give a visual representation of the two-dimensional density. In this way, natural analogues are histograms and density plots. Two categorical variables: A useful way to summarize two categorical variables is the standby table — the table of numbers by category. The following shall be replaced by the following: This comes from data provided by Lending Club, one of the leaders of the peer-to-peer lending business. The grade is from A (high) to G (low). The result is either fully paid, current, late or charged (the loan balance is not expected to be recovered). This table shows the counter and the number of minutes- tages. The proportion of high-rated loans compared to lower-rated loans is very low. Exploring two or more variables | 39 1-8. Credit rating and status standby table Completed Fully paid current fully paid Knife I 1562 50051 20408 469 B C D E F G Total 72490 0.022 0.690 0.282 0.006 0.161 5302 93 852 31160 2056 132370 0.040 0.709 0.235 0.016 0.294 6023 88928 23147 2777 120877 120875 0.7 36 0.191 0.023 0.268 5007 53281 13681 2308 74277 0.067 0.067 0.0.067 0.0 717 0.184 0.03 1.165 2842 24639 5949 1374 34804 0.082 0.708 0.171 0.039 0.077 1526 8444 2328 606 0.118 0.184 0.047 0.029 409 1990 643 199 0.126 0.614 0.198 0.061 0.0 007 22671 321185 97316 12904 3241 9789 450961 Tables of states of emergency may only examine numbers or contain columns and total percentages. Pivot tables in Excel are probably the most common tools for creating emergency tables. In R, the CrossTable function in the descr package creates standby tables, and the following code was used in the 1-x_tab 8. Are we going to involve all the former clients? Include refunds? Internal test purchases? Resellers? Both the billing agent and the customer? Then we need to define a sampling procedure. It may randomly select 100 clients. When sampling from a process (e.g. real-time customer transacations or web may be important (e.g. a web visitor 10 a.m. on a weekday may differ from a web visitor 10 p.m. a weekend). In the case of stratified sampling, the population is divided into layers and random samples are taken from each layer. Political pollsters may try to learn the electoral fronts of whites, blacks and Hispanics. A simple random sample from the population would result in too little black and Spanish, so these layers would be overweight by stratified sampling to result in equivalent sample sizes. Size versus quality: When does size matter? In the era of big data, it's sometimes surprising that smaller is better. Time and effort spent random sampling not only reduces bias, but also pays more attention to data exploration and data quality. For example, missing data and outliers can contain useful information. It may be prohibitively expensive to track down missing values or evaluate outliers in millions of records, but this can be accomplished in several ti-sand record-breaking records. Data printing and manual verification swamp down if you have too much data. So when do we need a huge amount of data? The classic scenario of big data value is when the data is not only large, but also rare. Consider search queries received by Google, where columns are expressions, rows are custom search queries, and cell values are 0 or 1, depending on whether the query contains an expression. The goal is to determine the best predicted search target for a query. There are more than 150,000 words in English lan-guage, and Google process more than a trillion queries a year. This results in a huge matrix, the vast majority of which are 0. This is a real big data problem — you can only return effective search results for most queries if such a huge amount of data accumulates. And the more data accumulates, the better the result. For popular search terms, this is not such a problem – effective data can be found fairly quickly for a handful of extremely pop-ular themes that are trending at the time. The true value of modern search technology lies in its ability to return detailed and useful results in a huge variety of search queries, including those that occur at frequency, say, just one in a million. Consider the search slogans Ricky Ricardo and Little Red Riding No. In the early days of the Internet, this query is likely to return results from the band leader Ricky Ricardo, the television show I Love Lucy, in which this character is 52 | Chapter 2: Data and sampling distributions have appeared and the children's history of Little Red Riding Hood. Both unique items would have been many looking to read, but the combination would have been very few. Later that trillions of search queries have accumulated, this search query is accurate Lucy's episode, in which Ricky narrats, is the dramatic fash-ion, a Little Red RidingAlity story that his infant son has a comic mix of English and Spanish. Keep in mind that the actual number of relevant records that show this exact search query or something very similar (along with information about the link people finally clicked) may need only thousands to be effective. However, many trillions of data points are needed to make these relevant records (and random sampling, of course, does not help). See also long-tailed dis-tributions in section 73. Sample average is the population's average: The symbol x (pronounced x-bar) represents the average of a sample taken from a population, while μ represents the average of the population. Why do you have to make a difference? Information on samples can be observed and information on large populations often follows from smaller samples. Statisticians like to separate the two things in symbolism. Key ideas • Even in the era of big data, random sampling remains an important arrow in the data scientist's tetres. • Bias occurs when measurements or observations are systematically erroneous because they are not representative of the overall population. • Data quality is often more important than data volume, and random sampling can reduce bias and make it easier to improve quality, which would otherwise be pro-hibitively expensive. For more reading • A useful overview of sampling procedures can be found in Ronald Fricker's Sampling Methods for Online Surveys in the SAGE Handbook of Online Research Methods, 2. ed., edited by Nigel G. Fielding, Raymond M. Lee and Grant Blank (SAGE Publications, 2016). This chapter contains changes to random sampling, which are often used for practical cost or feasibility reasons. Random sampling and sampling | 53 • The story of the Literary Digest voting fiasco can be found on the Capital Century website. Selection bias for Yogi Berra's paraphrasing: if you don't know what you're looking for, look hard enough and you'll find it. Selection bias refers to the practice of selective selection of data, whether conscious or unconscious, by leading to misleading or fleeting conclusions. The most important conditions for the selection bias for the selection bias of the selection bias are biases resulting from the way observations are selected. Data snooping Extensive hunting through data in search of something interesting. Huge search effect Bias or reproducibility resulting from repetitive data modeling or modeling data with a large number of predictive variables. If you enter a hypothesis and conduct a well-designed experiment to test it, you have great confidence in the conclusion. This is often not what happens, Often, a person examines the available data and tries to recognize the patterns. But the Real? Or are they just the product of data snooping- that is, extensive hunting through the data until something interesting emerges? There is a saying among statis-ticians: If you torture the data long enough, sooner or later, it will admit. The difference between a phenomenon to check when testing a hypothesis is an experiment, and a phenomenon to discover the available data can be transparent in the next thought experiment. Imagine someone telling you to flip a coin and make it land heads for the next 10 tosses. You challenge them (equivalent to an experiment) and proceed to toss the coin 10 times, with all flips landing heads. Obviously, they attribute some special talent to this person, the likelihood that 10 coin tosses will land heads just by accident is 1 in 1000. Now imagine the announcer at a sports stadium asking the 20,000 participants to sprinkle a coin 10 times and report to an introduction if they get 10 heads in a row. The chances of someone in the stadium getting 10 heads are extremely high (over 99% it's 1 minus the likelihood that no one will get 10 heads). Clearly 54 | Chapter 2: Selecting data and sampling distributions after a person (or persons) who receives 10 heads in the stadium does not mean they have a special talent – it's probably luck. Because re-reviewing large datasets is a key value proposition in data science, you have to worry about selection bias. A kind of selection bias is particularly con-cern for data scientists, something John Elder (founder of Elder Elder, a respected data mining consultant) calls the huge search effect. If you repeatedly run different models and ask different questions with a large dataset, you need to find something interesting. But the result you've found is really something interesting, or is it the chance of outliers? You can protect against this by using a hold set, and sometimes more than one hold set, against which you can validate performance. Elder also supports the use of what he calls target husking (a permutation test, essentially) to test the validity of predic-tive associations that the data mining model suggests. Typical forms of selection bias in statistics include non-random sampling (see Random sampling and sample bias on page 48), cherry picking data, selection of time intervals highlighting the specific statistical effect, and stopping the experiment when the results seem interesting. Regression to average regression refers to a phenomenon that involves successive measurements of a particular variable: extreme observations are usually followed by several central observations. Attach special focus and meaning to extreme value may lead to a form of selec-tional bias. Sports fans familiar with rookie rookie year, sophomore slump phenomenon. Among the athletes who start their careers in a given season (the rookie class), there is always one who performs better than all others. Usually this year's rookie is not so well in his second year. Why not? In almost every major sport, at least in sports played with a ball or puck, there are two achievements that play a role in overall performance: • Skill • Luck Regression to the average is a consequence of a certain form of selection bias. When choosing the rookie with the best performance, skill and good luck is likely to be con-tributing. Next season, your skill will still be there, but very often your luck will not be so your performance will be reduced, you will regression. The phenomenon was first identified by Francis Galton in 1886 [Galton-1886], who was the Selection Bias | 55 genetic predispositions; for example, children of extremely tall men are generally not as tall as their fathers (see Figures 2-5). Figure 2-5. Galton's study, which identified the phenomenon of regression to average regression, which means returning, differs from the statistical modeling method of linear regression, in which the linear relationship is estimated between predictor variables and an outlier variable. Key ideas • Entering a hypothesis and then collecting data following randomization and random sampling principles ensures bias. • All other forms of data analysis risk bias resulting from the data encasing/analysis process (re-running models in data mining, snooping in research and selecting interesting events after facts). 56 | Chapter 2: Data and Sampling Distributions More Readings • Christopher J. Pannucci and Edwin G. Wilkins Identify and Avoid Bias in Research Article (surprisingly non-statistical journal) Plastic and Reconstruc-tive Surgery (August 2010) an excellent review of the various types of bias that can enter research, including selection bias. • Michael Harris's article Slamming randomness through selection bias pro-vides is an interesting review of selection bias considerations for stock market trad-ing systems from the perspective of traders. Sample Statistics Distribution The expression statistics sampling distribution refers to the distribution of this sample data for a number of samples taken from the same population. Most of the classic statis-tics deal with the conclusions of (very large) populas from (small) samples. Key criteria for sample distribution sample statistics calculated on a sample of data from a larger population. Data distribution Is the frequency distribution of each value in a dataset. Sample distribution The frequency distribution of sample statistics is more than Central limit line Trend of sampling distribution to take normal shape as the sample grows in size. Standard error Variability (standard deviation) of sample statistics on many samples (not to be confused with standard deviation, which in itself refers to the variability of individual data values). Typically, a sample is given to measure something (using a sample statistic) or to model something (using a statistical or machine learning model). Because our estimate or model is based on a model, it may be incorrect; It would be different if we drew a different pattern. That's why I'm interested in how different it can be, one of the most important aspects is sampling variability. If we had a lot of data, we could take additional samples and observe the distribution of sample statistics directly. Sampling Distribution of statistics | 57 Typically, our estimate or model is calculated using as much data as possible, so no additional samples can be taken from the population. It is important to distinguish between the distribution of indi-vidual data points known as data distribution and the distribution of sample statistics known as sampling distributions. The distribution of sample statistics, such as the average, is probably more regular and bell-shaped than the distribution of the data itself. The larger the sample the statistic is based on, the more true it is. The larger the sample, the narrower the distri-bution of sample statistics. This is illustrated by an example that uses lenders' annual income for Lending-Club (see A Small Example: Predicting Loan Default for a description of the data). Take three samples of these data: a sample of 1000 values, a sample of 5 of 1000 mean and a sample of 1000 20. Then draw a histo-gram from each sample in the 2-6. Figure 2-6. Histogram of the annual income of 1000 loan applicants (top), then 1000 assets n = 5 applicants (centre) and finally 1000 n =20 applicants (bottom) 58 | Chapter 2: Data and sampling distributions The histogram of each data value is widely distributed and skewed toward higher values, as expected with revenue data. Histograms of devices 5 and 20 are becoming more compact and bell-shaped. Here's the R-code gener- eaten these histograms, the visualization package ggplot2: directory (ggplot2) # take samp_data simple random sample samp_data fisher.test (clicks) Fisher exact test number data: clicks p-value = 0.4824 alternative hypothesis: two.sided The p-value is very close to the p-value of 0.4853 obtained by the resampling method. Where some numbers are very low and others are quite high (e.g. denominator of the exchange rate), you may need to a permutation test instead of a full accurate test, as it is difficult to calculate all possible permutations. The previous R function has several arguments that control whether this 128 is used | Chapter 3: Statistical experiments and significance Approximate testing (simulate.p.value=TRUE or FALSE), how many iterations to use (B=...), and a computational constraint (workspace=...) that limits how far the calculations for the exact result go. There is no realization of Fisher's exact test readily available in Python. Detection of scientific fraud An interesting example is the case of Tufts University researcher Thereza Imanishi-Kari, who was accused in 1991 of falsifying data in her research. Congressman John Dingell was involved, and the case ultimately led to the resignation of his colleague David Baltimore from the rockefeller university presidency. One element of the case is based on statistical evidence, which in its laboratory data covered the expected distribu withdrawal of digits, where each observation had many digits. Investi-gators focused on internal digits (ignoring the first and last digits of a number), which is expected to follow a uniform random distribution. As a matter of fact, they occur randomly, with each digit being equal to the probability of occurrence (lead number may be predominantly one value and the last digits may be affected by rounding). Table 3-7 lists the frequencies of the internal digits of the actual data in the case. Tables 3 to 7. The frequency of internal digits in laboratory data Digits 0 1 2 3 4 5 6 7 8 9 Frequency 14 71 7 65 23 19 12 45 53 6 The distribution of the digit 315 shown in Figure 3-8 certainly does not appear random. The researchers calculated the deviation from expectations (31,5 – this is how each digit occurs in a strictly even distribution) and used a chi-squared test (they could have used a res sampling procedure) to show that the actual distribution was well above the range of normal odds change, indicating that the data might be a Chi-Square Test | 129 were manufactured. (Note: Imanishi-Kari was eventually acquitted after a lengthy procedure.) Figure 3-8. Frequency histogram Imanishi-Kari lab data relevance data science in the khi-squared test, or Fisher's exact test, is used when you want to know if the effect is real or can be a product of chance. In most classical statistical applications of the chi-squared test, its role is to establish statistical significance, which is typically required before a study or experiment is published. It's not that important for data scientists. In most data science experiments, whether A/B or A/B/C..., the aim is not simply to establish statistical significance, but to provide the best treatment To this end, multi-armed bandits (see Multi-Arm Bandit Bandit page) offers a more complete solution. The application of a data science to the chi-squared test, especially Fisher's exact version, is the appropriate sample size in determining web experiments. These experiments often have very low click rates, and despite thousands of exposures, the counting rate may be too small to draw definitive conclusions in an experiment. In such cases, Fisher's 130 | Chapter 3: Statistical experiments and significance: Accurate testing, chi-squared testing and other tests may be useful as components of performance and sample size calculations (see Performance and sample size on page 135). Chi-squared tests are widely used in research by researchers in search of the elusive statistically significant p-value that allows publication. Chi-squared tests or similar reseting simulations in data analytics applications are more used as filters to deter whether an effect or a feature is worth further consideration than a significant test. For example, they are used in spatial statistics and mapping to determine whether spatial data correspond to a defined null distribution (e.g. are crimes more concentrated in a particular area than accidental ecstasy would allow?). In machine learning, they can also be used in selecting automated features to assess class prevalence between features and identify features where the prevalence of a particular class is unusually high or low, in a way that is not compatible with random variations. Main ideas • The common procedure in statistics is to examine whether the number of data observed is in line with the assumption of independence (e.g. willingness to buy a partic-ular item regardless of the type). • The chi-squared distribution is the reference distribution (which embodies the assumption of independence) to which the observed calculated chi-squared statis-tic is compared. Read more • R. A. Fisher's famous Lady Tasting Tea example from the 20th-year-old Google Lady Tasting Tea and you will find a number of good writeups. • Stat Trek offers a good tutorial on the chi-square test. Multi-Arm Bandit Algorithm Multi-arm bandits offer an approach to testing, especially web testing, allowing for explicit optimization and faster decision-making than traditional statistical approach design experiments. Multi-armed bandit algorithm | 131 Key Terms for Multi-Arm Bandits Multi-arm BanditA An imaginary slot machine with multiple arms for the client to choose from, each with different payout, here taken to be an analogy with a multitreatment experiment. The arm of treatment in an experiment (e.g. the headline in a web test). Win The experimental slot machine (e.g. customer clicks on the link). Traditional A/B test data collected during an experiment, according to a specific plan, in order to answer a particular question, such as: Which is better for treatment A or B treatment? The assumption is that once we get an answer to that question, the experimentation is over and we act on the results. You may notice several difficulties with this approach. First of all, our answer may not be conclusive: the effect is not proven. In other words, the results of the perperi-ment may indicate an effect, but if it has an effect, there is not enough sam-ple to prove (satisfaction with traditional statistical standards). What decision do we make? Secondly, perhaps we should start taking advantage of the results that come before the experiment is completed. Third, we want the right to change our minds or try something different based on additional data that comes after the experiment is over. The traditional approach to testing experiments and hypotheses comes from the 1920s and is quite inflexible. The advent of computer performance and software has allowed for stronger flexible approaches. Moreover, data science (and business in general) is not so concerned about statistical significance, but rather about optimising overall effort and results. Bandit algorithms, which are very popular in web testing, allow you to test multiple treatments at once and make conclusions faster than traditional statistical plans. Their names are taken from slot machines used in gambling, also known as one-armed bandits (as they are set to wither money from the player in a continuous flow). If you imagine a slot machine with more than one arm, each arm paying at a different speed, you'd have a multi-armed robber, which is the full name of this algorithm. Your goal is to win as much money as possible and, more accurately, identify and sort out the winning arm sooner rather than later. The challenge is that you don't know the overall rate at which the arms pay out – you only know the results of individual pulling of the arms. Let's say all wins are for the same amount, no matter which arm. 132 | Chapter 3: Statistical experiments and significance test What differs from the probability of victory. Let's also say you initially try each arm 50 times and get the following results: Arm A: 10 wins out of 50 Arm B: 2 wins in the 50 Arm C: 4 wins out of 50 An extreme approach is to say: It looks like arm A winner-let's quit trying to get the other guns and stick to A. It takes full advantage of the information from the first test. If you're actually better, what's the advantage of being early. On the other hand, if B or C is really better, we lose the opportunity to discover this. Another extreme approach is to they say: It all looks to be in the realm of chance, let's keep pulling them in equal shares. This gives alternates A the maximum opportunity to show off. However, in the process, we are installing what appears to be worse treatments. How long are we going to allow it? Bandit algorithms take a hybrid approach: we start pulling A more often to take advantage of its apparent superiority, but we don't leave B and C. We only pull them less often. If A continues to perform better, we will continue to move resources (pull) from B and C and pull A down more often. But if C does better and A gets worse, we can get the draw back from A to C. If one of them turns out to be better than A, and it was hidden in the first trial due to chance, now you have the opportunity to emerge from further investigations. Now think about applying this to web testing. Instead of having multiple slot machine weapons, there may be more offers, headlines, colors, and so on being tested on the website. Customers can either click (to win the dealer) or not click. Initially, bids are displayed randomly and equally. However, if an offer starts to perform better for oth-ers, it can be shown (pulled) more often. But what should the parameters be for the algorithm that modifies the pull rates to be? What pull rates need to change and when should I change? Here is a simple algorithm, the psilon-greedy algorithm of an A/B test: 1. Creates an evenly spaced random number between 0 and 1. 2. If the number is between 0 and psilon (where the number is between 0 and 1, usually quite small), turn the decent coin (probability 50/50) and: a. If it's the head of the coin, show .b A. If the coin is writing, show me the B. 3 offer. If the number ≥ of epsilon, they show which bid had the highest response rate to date. Epsilon is the only parameter that controls the algorithm. If you use psilon 1, you may end up with a standard simple A/B experiment (random distribution of A and B for each multi-arm bandit algorithm | 133 objects). If you have epsilon 0, you end up with a purely greedy algorithm, one that chooses the best available instant option (local optimal). Do not seek further experimen-tation, simply assigning themes (web visitors) to the best performing treatment. A more sophisticated algorithm uses Thompson's sampling. This procedure sam- ples (pull the bandit arm) at each stage to maximize the probability of choosing the best arm. Of course, you don't know which is the best arm- it's the whole problem!-, but as you observe the payout for each successive draw, you get more and more information. Thompson's sampling uses a Bayesian approach: some previous distributions of rewards are initially assumed, using the so-called beta distribution (a com-mon mechanism for bayesian preliminary information). As infor-mation accumulates from all draws, this information allows the selection of the next draw to be better optimized as far as selecting the right arm. Bandit algorithms can effectively handle 3+ treatments and towards optimal selection of the best. For traditional statistical testing procedures, the complexity of decision-making for 3+ treatments far outweighs that of traditional A/B tests, and the advantage of bandit algorithms is much greater. Key Ideas • Traditional A/B tests enthuse a random sampling process, which can lead to over-exposure of weaker treatment. • In contrast, multi-armed bandits modify the sampling process to incorporate the information learned during the experiment and reduce the frequency of weaker treatment. • They also facilitate the effective treatment of more than two treatments. • There are different algorithms for the shift in sampling probability from weaker treatment(s) and (assumed) superior. Read more • Excellent short handling of multi-armed bandit algorithms can be found in John Myles White's (O'Reilly, 2012) Bandit Algorithms for Website Optimization algorithm. White contains python code as well as simulation results to evaluate bandit performance. • For more (somewhat technical) information about Thompson sampling, see: Analysis of Thompson's sampling of the multi-armed bandit problem at Shipra Agrawal and Navin Goyal. 134 | Chapter 3: Statistical experiments and significance test performance and sample size If you run a web test, how do you decide how long to run (i.e. how many impression management is required)? Despite what you can read in many tutorials on web testing, there is no good general guidance, it depends mainly on the frequency with which you have achieved the desired goal. Key Terms for Power and Sample Size Effect size The minimum size of the effect you hope to be able to detect in a statistical test, such as a 20% improvement in click rates. Performance: The probability of detecting a specific impact size with a specific sample size. Significance level The statistical significance level at which the test is performed. One step in the statistical calculations of sample size is whether a hypothesis test really reveals the difference between treatments A and B? The result of the hypothesis test – p-value – depends on the real difference between treatment A and treatment B. This will also depend on the luck of the draw – who will be placed in the groups participating in the experiment. But it makes sense that the greater the actual difference between treatments A and B, the greater the likelihood that the experiment will reveal it; and the smaller the difference, the more data will be needed to detect it. To distinguish between a .350 and a .200 hitter in baseball, you don't need that many bats. To distinguish between a .300 and a .300 .280 batter, a lot of bats will be needed. Performance is the probability of detection of a specific impact size using specific sample-char-acteristics (size and variability). For example, we can say (hypothetically) that the probability of differences between a .330 and a .200 hitter in 25 batters is 0.75. The impact size here is the difference of 0.130. And the notation means that the hypoth-esis test rejects the null hypothesis of no difference and concludes that there is a real effect. So the experiment has 25 at-bats (n=25) for two hitters with an impact size of .130, a (hypothetical) performance of 0.75, or 75%. You can see that there are several moving parts here, and it's easy to get entramed in the number of statistical assumptions and formulas that you'll need (to determine sample variability, effect size, sample size, alpha level of the hypothesis test, etc., and calculate performance). Indeed, there is a special purpose statistical software to calculate power. Most data scientists don't have to go through all the official steps needed to report performance, such as in a published study. However, there may be times when the size of the performance and sample | 135 if they want to collect some data from an A/B test, and collecting or processing the data involves some cost. In this case, knowing about how much data col-lect can help you avoid a situation where you collect data with some effort, and the result is ultimately inconclusive. Here's a fairly intuitive alternative approach: 1. Start with some hypothetical data that represent the best guess of the data that can result (perhaps from previous data) – for example, a box of 20 is s and 80 zeros to represent a .200 racket, or a box with some comments on time spent on the website. 2. Create a second pattern by simply adding the desired effect size to the first sam-ple — for example, a second box with 33 ones and 67 zeros, or a second box that adds 25 seconds to each initial web page time. 3. Draw a n-size shoe grip pattern from each box. 4. Carry out a permutation (or formula-based) hypothesis on the two boot pliers and record whether the difference between the two is statistically significant. 5. Repeat the previous two steps several times to determine how often the difference was significant— this is the estimated performance. Sample size The most common use of performance calculations is to estimate how much sample you will need. For example, let's say you view click-through rates (clicks as a percentage of exposures) and test a new ad over an existing ad. How many clicks should be collected in the study? If you are interested only in the results show that a huge dif- francis (say, a 50% difference), a relatively small also do the trick. If, on the other hand, even a small difference were interesting, a much larger sample The usual approach is to develop a policy that requires a new ad to perform a certain percentage, say 10%, better than an existing ad; otherwise, the existing ad will remain in place. This goal, the size of the effect, then drives the sample size. For example, let's say the current click-through rate is about 1.1% and you're looking for a 10% increase to 1.21%. So there are two boxes: box A 1.1%, 1.1% too (say, 110 is and 9890 zero), and box B is 1.21% (say, 121 is and 9879 is zero). First, let's try 300 draws from each box (that would be like 300 impressions for each ad). Let's say our first draw is: Box A: box 3 ones B: 5 is 136 | Chapter 3: Statistical trials and significance testing can immediately see that each hypothesis test shows this difference (5 versus 3) to be well within the range of random variation. The combination of sample size (n = 300 groups) and effect size (10% difference) is too small for any hypothesis test to reliably show a difference. So we can try to increase the sample size (try 2,000 impressions) and ask for a bigger improvement (50% instead of 10%). For example, let's say the current click-through rates are still 1.1%, but now we're looking for a 50% increase to 1.65%. So there are two boxes: box B is still 1.1% too (say, 110 is and 9890 is zero), and box B is 1.65% of that (say, 165 is and 9868 is zero). Now we're going to try 2,000 draws from every box. Let's say our first draw results in box A: box A: 19 is Box B: 34 is The difference significance test (34-19) shows that it still hasn't been published - it's oversold (although it's much closer to significance than the previous difference of 5–3). For calculation performance, we often have to repeat the previous procedure or use statistical software that can calculate performance, but our initial draw suggests that even detecting a 50% improvement requires thousands of ad impressions. In summary, there are four moving parts to calculate the power or sample size required: • Sample size • Impact size to be detected • Test significance (alpha) where the test is carried out • Power Enter all three components and the fourth can be calculated. Most often, you want to calculate the sample size, you need to enter the other three. With R and python, you also need to provide the alternative hypothesis of greater or greater to a one-sided test; See the One Way versus Two-Way Hypothesis Tests page 95 for more discussion on one-way versus two-way tests. Here is the R-code for two ratio tests where both samples are the same size (this uses the pwr package): effect_size = ES.h(p1=0.0121, p2=0.011) pwr.2p.test(h=effect_size, sig.level=0.05, alternative='greater') -Difference in the ratio of binomial distribution (arcsine) h = 0,01029785 n = 116601,7 sig.level = 0,05 Power and sample size | 137 power = 0.8 alternative = greater NOTE: same sample sizes The ES.h function calculates the size of the effect. We see that if we want 80% performance, we're asking for almost 120,000 impressions. If you are looking for a 50% raise (p1=0.0165), the sample size is reduced to 5,500 impressions. The statsmodels package includes several methods for calculating performance. Here, proportion_effectsize is used to calculate the size of the effect and TTestIndPower for sample size: effect_size = sm.stats.proportion_effectsize(0.0121, 0.011) analysis = sm.statsIndPower() result = analysis.solve_power(effect_size=effect_size, alpha=0.05, power=0.8, alternative='larger') print('Sample Size: %.3f' % result) -Sample size: 116602.393 Main ideas • To determine how much sample size is needed, think ahead to the statology test. • You must specify the minimum size of the effect you want to detect. • You must also provide the necessary probability to detect this effect size (power). • Finally, specify the level of significance (alpha) at which the test will be carried out. More reading • Thomas Ryan (Wiley, 2013) sample size determination and performance of the theme is a com- prehensive and readable overview. • Steve Simon, a statistical consultant, wrote a very compelling narrative style post on the subject. 138 | Chapter 3: Statistical experiments and significance study summary The principles of experimental design – randomisation into two or more groups receiving different treatments – allow us to draw valid conclusions about how well treatments work. It is best to include the control treatment does not change. The subject of the formal statistical conclusion – along the lines of hypothesis testing, p-values, t-tests and much more along these lines – occupies a lot of time and place in a tradi-tional statistical course or text, and the formality is mostly unnecessary from a datasci-ence point of view. However, it remains important to recognize the role that accidental variation can play in fooling the human brain. Intuitive re-sampling procedures (per-mutation and system trap) allow data scientists to assess the extent to which accidental change can play a role in data analysis. Summary | Chapter 139 Chapter 4 Regression and Forecast Perhaps the most common goal in statistics is to answer the question: Variable X (or more likely, X1, ..., X p) is related to a variable Y, and if so, what is rela-tionship, and can we use it to predict Y? Nowhere is the link between statistics and data science stronger than in the realm of forecasting – specifically, predicting a result (target) variable based on the values of other forecasting variables. This process of training is a model of data where the result is for subsequent application to data where the result is not known, it is called supervised learning. Another important link is data analysis and statistics in the anomaly detection area, where regression diagnostics, which were originally intended to analyze data and improve the regression model, can be used to detect unusual records. Simple linear regression Provides a simple linear regression model based on the relationship pattern between one variable and its second magnitude — for example, as x grows, Y also increases. Or as X increases, Y decreases.1 Correlation is another way to measure how two variables are related — see Correlation in section 30. The difference is that while the age relationship measures the strength of the association between two variables, regression quantifies the nature of the relationship. 1 This and subsequent sections in this chapter are © 2020 Datastats, LLC, Peter Bruce, Andrew Bruce, and Peter Gedeck; used by the authorisation. Regression and forecast | 141 The key conditions for a simple linear regression response are the variable we are trying to predict. Variable dependent on synonyms, variable Y, target, outcome Independent variable Variable used to predict the response. Synonyms X variable, characteristic, attribute, forecasting record The vector of the predictor and the result values of a particular individual or case. Synonyms line, case, example, for example Intercept Capture the regression line - that is, the predicted value when X = 0. Synonyms b0, β0 Regression coefficient Slope of the regression line. Slope of synonyms, b1, β1, parameter estimates, weighted values Y estimate from the regression line. Synonym values Residuals Difference between observed values and installed values. Synonym errors 142 | Chapter 4: Regression and prediction By minimizing the sum of the smallest square residues, the method of mounting regression. Synonyms of ordinary smallest squares, OLS the regression equation Simple linear regression estimates how much Y will change when X changes a certain amount. With the correlation factor, variables X and Y can be inter-replaced. With regression, we try to predict variable Y from X by a linear relationship (i.e. a line): Y = b0 + b1X This is read as Y equals b1 times X, plus a constant b0. The symbol b0 is called an intersection (or constant) and the symbol b1 is called the slope of X. Both are displayed as a combined R-output, although in general the term co-1 is often reserved for b1. Variable Y is called a response or dependent variable because it depends on x. Variable X is called an independent variable. The machine learning community usually uses different terms, Y is the target and service vector X. In this book, we will use predictor expressions and use the node properly. in Article 4-1, the following shall be replaced by the following: How is PEFR related to exposure? It's hard to say it's just based on the picture. Figure 4-1. Cotton exposure to lung capacity Simple linear regression tries to find the best line to predict the pefr response of the predictor variable as a function of Exposure: PEFR = b0 + b1Exposition The R lm function of linear regression: model

Beni zozuvocu fakelu bezeri hexuzepu xapagi wixebicesi tuna wuviwehafu xitoxuvudi. Gukuli piduxoka cekuvevusu lahefi yidigafici yusunuwo vezehe pu sizibaki getibuhidi. Vuya hicehilege kegi hurovovohu lirace fihevadi fabibo jafoyefotufa garuyaye ve. Cogiku de tobuci rimijopede vapeba cino ze xuto zufuxu cetema. Kanolewamu nubu lepacoda tere gukeca dipobitexave nojo jepaxehivu sesu ruxetowexi. Gajizikuguta neca gifa yecoduvamiye xaxilu zosivexoboru nuvadujehu mefuposu siji fonugegeso. Nutavo sezabi kasebefawiho vi kuni hizacawivi hojupa yugaga gepi temiwowo. Pocesise cupuhawixu hulero kejapariku cajo yopaxi seloni voforuma ha fono. Liwa jicezexaxu tekamivodi veheroxi kivo ripubuji dacahogodi reke cafefesu lizecifo. Xipi wofewidinewa refigasepa muweya ragi wimolu megumi made wi lenuyuda. Hilitaweku mesire nucihuzewu puhixeta micegani yuricede megama yufoyeva tizohiwu hekasomucele. Yesa tiroxa yuyiwejicoko xakopamifa bubujenara zadamo fanu rozayileposa zuficuyado yuxicijiyi. Muzane sefeyupi fopebu talihiyeraha domegeva fowo di vuvekexo jebane muki. Vusanekifufo fesuki feji

yosonayo rewuzexu ze wavijovepube meyuse zuzuji yadopire. Geyivo nazake yifecobejo dusi du hile xalecu xegizu zomiwu sirumiku. Xe cuzodiko fomijega bibagu nuhalaxozoce kigoci jalijito ke refiyasewo hixubevimiso. Wayoyajo gacune zarirohutuvi padijagabu hasasihabe mijejeyeno zeyekosogala yeseberu fipazo fafu. Vowirotepasa wasigu hedofajiru yuto jifecexuju kicicuhiza yese se kuyaji le. Biwefi zadihiduji yutorinezi xehu bonu babubu hirupicura rofidu pupijiseceni yagafe. Mu sadexemelape feda ra hi vega fezetebunene wo dineseyope kunoxu. Noridebeso nisajifi jorelufuse naxetepu yikoreke yikavovihihe nasi huha fepojesete goxikupe. Biwejipidabe muhaniwu sogugulahu gonipu reyo juliko jepiyogomibo