The Elements of Data Analytic Style

Jeff Leek

The Elements of Data Analytic Style

A guide for people who want to analyze data.

Jeff Leek

This book is for sale at http://leanpub.com/datastyle

This version was published on 2015-03-02

This is a Leanpub book. Leanpub empowers authors and publishers with the Lean Publishing process. Lean Publishing is the act of publishing an in-progress ebook using lightweight tools and many iterations to get reader feedback, pivot until you have the right book and build traction once you do.

An @simplystats publication.

Thank you to Karl Broman and Alyssa Frazee for constructive and really helpful feedback on the first draft of this manuscript. Thanks to Roger Peng, Brian Caffo, and Rafael Irizarry for helpful discussions about data analysis.

Contents

1. Introduction

The dramatic change in the price and accessibility of data demands a new focus on data analytic literacy. This book is intended for use by people who perform regular data analyses. It aims to give a brief summary of the key ideas, practices, and pitfalls of modern data analysis. One goal is to summarize in a succinct way the most common difficulties encountered by practicing data analysts. It may serve as a guide for peer reviewers who may refer to specific section numbers when evaluating manuscripts. As will become apparent, it is modeled loosely in format and aim on the Elements of Style by William Strunk.

The book includes a basic checklist that may be useful as a guide for beginning data analysts or as a rubric for evaluating data analyses. It has been used in the author's data analysis class to evaluate student projects. Both the checklist and this book cover a small fraction of the field of data analysis, but the experience of the author is that once these elements are mastered, data analysts benefit most from hands on experience in their own discipline of application, and that many principles may be non-transferable beyond the basics.

If you want a more complete introduction to the analysis of data one option is the free Johns Hopkins Data Science Specialization[1].

As with rhetoric, it is true that the best data analysts sometimes disregard the rules in their analyses. Experts usually do

[1] https://www.coursera.org/specialization/jhudatascience/1

this to reveal some characteristic of the data that would be obscured by a rigid application of data analytic principles. Unless an analyst is certain of the improvement, they will often be better served by following the rules. After mastering the basic principles, analysts may look to experts in their subject domain for more creative and advanced data analytic ideas.

2. The data analytic question

2.1 Define the data analytic question first

Data can be used to answer many questions, but not all of them. One of the most innovative data scientists of all time said it best.

> The data may not contain the answer. The combination of some data and an aching desire for an answer does not ensure that a reasonable answer can be extracted from a given body of data.
>
> John Tukey

Before performing a data analysis the key is to define the type of question being asked. Some questions are easier to answer with data and some are harder. This is a broad categorization of the types of data analysis questions, ranked by how easy it is to answer the question with data. You can also use the data analysis question type flow chart to help define the question type (Figure 2.1)

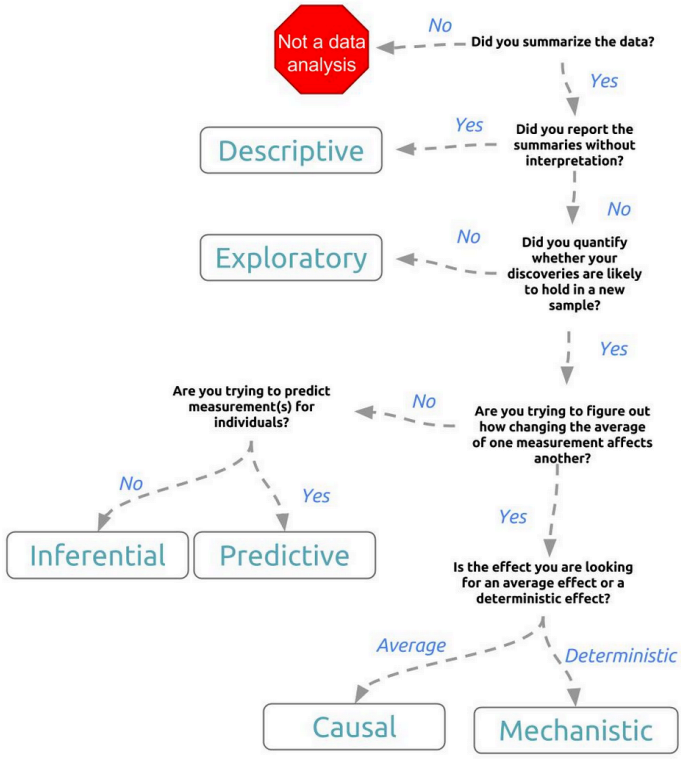

Figure 2.1 The data analysis question type flow chart

2.2 Descriptive

A descriptive data analysis seeks to summarize the measurements in a single data set without further interpretation. An example is the United States Census. The Census collects data on the residence type, location, age, sex, and race of all people in the United States at a fixed time. The Census is descriptive because the goal is to summarize the measurements in this fixed data set into population counts and describe how many

people live in different parts of the United States. The interpretation and use of these counts is left to Congress and the public, but is not part of the data analysis.

2.3 Exploratory

An exploratory data analysis builds on a descriptive analysis by searching for discoveries, trends, correlations, or relationships between the measurements of multiple variables to generate ideas or hypotheses. An example is the discovery of a four-planet solar system by amateur astronomers using public astronomical data from the Kepler telescope. The data was made available through the planethunters.org website, that asked amateur astronomers to look for a characteristic pattern of light indicating potential planets. An exploratory analysis like this one seeks to make discoveries, but rarely can confirm those discoveries. In the case of the amateur astronomers, follow-up studies and additional data were needed to confirm the existence of the four-planet system.

2.4 Inferential

An inferential data analysis goes beyond an exploratory analysis by quantifying whether an observed pattern will likely hold beyond the data set in hand. Inferential data analyses are the most common statistical analysis in the formal scientific literature. An example is a study of whether air pollution correlates with life expectancy at the state level in the United States. The goal is to identify the strength of the relationship in both the specific data set and to determine whether that relationship will hold in future data. In non-randomized

experiments, it is usually only possible to observe whether a relationship between two measurements exists. It is often impossible to determine how or why the relationship exists - it could be due to unmeasured data, relationships, or incomplete modeling.

2.5 Predictive

While an inferential data analysis quantifies the relationships among measurements at population-scale, a predictive data analysis uses a subset of measurements (the features) to predict another measurement (the outcome) on a single person or unit. An example is when organizations like FiveThirtyEight.com use polling data to predict how people will vote on election day. In some cases, the set of measurements used to predict the outcome will be intuitive. There is an obvious reason why polling data may be useful for predicting voting behavior. But predictive data analyses only show that you can predict one measurement from another, they don't necessarily explain why that choice of prediction works.

2.6 Causal

A causal data analysis seeks to find out what happens to one measurement if you make another measurement change. An example is a randomized clinical trial to identify whether fecal transplants reduces infections due to *Clostridium difficile*. In this study, patients were randomized to receive a fecal transplant plus standard care or simply standard care. In the resulting data, the researchers identified a relationship between transplants and infection outcomes. The researchers

were able to determine that fecal transplants caused a reduction in infection outcomes. Unlike a predictive or inferential data analysis, a causal data analysis identifies both the magnitude and direction of relationships between variables.

2.7 Mechanistic

Causal data analyses seek to identify average effects between often noisy variables. For example, decades of data show a clear causal relationship between smoking and cancer. If you smoke, it is a sure thing that your risk of cancer will increase. But it is not a sure thing that you will get cancer. The causal effect is real, but it is an effect on your average risk. A mechanistic data analysis seeks to demonstrate that changing one measurement always and exclusively leads to a specific, deterministic behavior in another. The goal is to not only understand that there is an effect, but how that effect operates. An example of a mechanistic analysis is analyzing data on how wing design changes air flow over a wing, leading to decreased drag. Outside of engineering, mechanistic data analysis is extremely challenging and rarely undertaken.

2.8 Common mistakes

2.8.1 Correlation does not imply causation

Interpreting an inferential analysis as causal.

Most data analyses involve inference or prediction. Unless a randomized study is performed, it is difficult to infer why

there is a relationship between two variables. Some great examples of correlations that can be calculated but are clearly not causally related appear at http://tylervigen.com/[1] (Figure 2.2).

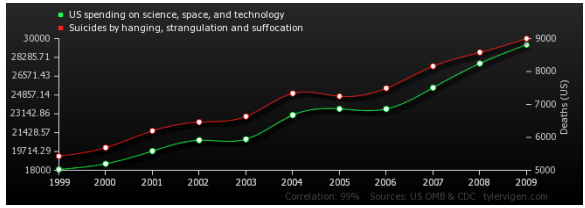

Figure 2.2 A spurious correlation

Particular caution should be used when applying words such as "cause" and "effect" when performing inferential analysis. Causal language applied to even clearly labeled inferential analyses may lead to misinterpretation - a phenomenon called causation creep[2].

2.8.2 Overfitting

Interpreting an exploratory analysis as predictive

A common mistake is to use a single, unsplit data set for both model building and testing. If you apply a prediction model to the same data set used to build the model you can only estimate "resubstitution error" or "training set error". These estimates are very optimistic estimates of the error you would get if using the model in practice. If you try enough models on the same set of data, you eventually can predict perfectly.

[1] http://tylervigen.com/

[2] http://junkcharts.typepad.com/numbersruleyourworld/causation-creep/

2.8.3 n of 1 analysis

> Descriptive versus inferential analysis.

When you have a very small sample size, it is often impossible to explore the data, let alone make inference to a larger population. An extreme example is when measurements are taken on a single person or sample. With this kind of data it is possible to describe the person or sample, but generally impossible to infer anything about a population they come from.

2.8.4 Data dredging

> Interpreting an exploratory analysis as inferential

Similar to the idea of overfitting, if you fit a large number of models to a data set, it is generally possible to identify at least one model that will fit the observed data very well. This is especially true if you fit very flexible models that might also capture both signal and noise. Picking any of the single exploratory models and using it to infer something about the whole population will usually lead to mistakes. As Ronald Coase[3] said:

> "If you torture the data enough, nature will always confess."

This chapter builds on and expands the paper "What is the question?"[4] co-authored by the author of this book.

[3] http://en.m.wikiquote.org/wiki/Ronald_Coase

[4] http://www.sciencemag.org/content/early/2015/02/25/science.aaa6146.full

3. Tidying the data

The point of creating a tidy data set is to get the data into a format that can be easily shared, computed on, and analyzed.

3.1 The components of a data set

The work of converting the data from raw form to directly analyzable form is the first step of any data analysis. It is important to see the raw data, understand the steps in the processing pipeline, and be able to incorporate hidden sources of variability in one's data analysis. On the other hand, for many data types, the processing steps are well documented and standardized.

These are the components of a processed data set:

1. The raw data.
2. A tidy data set.
3. A code book describing each variable and its values in the tidy data set.
4. An explicit and exact recipe you used to go from 1 to 2 and 3.

3.2 Raw data

It is critical that you include the rawest form of the data that you have access to. Some examples of the raw form of data are as follows.

1. The strange binary file your measurement machine spits out
2. The unformatted Excel file with 10 worksheets the company you contracted with sent you
3. The complicated JSON data you got from scraping the Twitter API
4. The hand-entered numbers you collected looking through a microscope

You know the raw data is in the right format if you ran no software on the data, did not manipulate any of the numbers in the data, did not remove any data from the data set, and did not summarize the data in any way.

If you did any manipulation of the data at all it is not the raw form of the data. Reporting manipulated data as raw data is a very common way to slow down the analysis process, since the analyst will often have to do a forensic study of your data to figure out why the raw data looks weird.

3.3 Raw data is relative

The raw data will be different to each person that handles the data. For example, a machine that measures blood pressure does an internal calculation that you may not have access to when you are given a set of blood pressure measurements. In general you should endeavor to obtain the rawest form of the data possible, but some pre-processing is usually inevitable.

3.4 Tidy data

The general principles of tidy data are laid out by Hadley Wickham in this paper[1] and this video[2]. The paper and the video are both focused on the R package, which you may or may not know how to use. Regardless the four general principles you should pay attention to are:

- Each variable you measure should be in one column
- Each different observation of that variable should be in a different row
- There should be one table for each "kind" of variable
- If you have multiple tables, they should include a column in the table that allows them to be linked

While these are the hard and fast rules, there are a number of other things that will make your data set much easier to handle.

3.5 Include a row at the top of each data table/spreadsheet that contains full row names.

So if you measured age at diagnosis for patients, you would head that column with the name AgeAtDiagnosis instead of something like ADx or another abbreviation that may be hard for another person to understand.

[1] http://vita.had.co.nz/papers/tidy-data.pdf
[2] https://vimeo.com/33727555

3.6 If you are sharing your data with the collaborator in Excel, the tidy data should be in one Excel file per table.

They should not have multiple worksheets, no macros should be applied to the data, and no columns/cells should be highlighted. Alternatively share the data in a CSV or TAB-delimited text file.

3.7 The code book

For almost any data set, the measurements you calculate will need to be described in more detail than you will sneak into the spreadsheet. The code book contains this information. At minimum it should contain:

- Information about the variables (including units!) in the data set not contained in the tidy data
- Information about the summary choices you made
- Information about the experimental study design you used

In our genomics example, the analyst would want to know what the unit of measurement for each clinical/demographic variable is (age in years, treatment by name/dose, level of diagnosis and how heterogeneous). They would also want to know how you picked the exons you used for summarizing the genomic data (UCSC/Ensembl, etc.). They would also

want to know any other information about how you did the data collection/study design. For example, are these the first 20 patients that walked into the clinic? Are they 20 highly selected patients by some characteristic like age? Are they randomized to treatments?

A common format for this document is a Word file. There should be a section called "Study design" that has a thorough description of how you collected the data. There is a section called "Code book" that describes each variable and its units.

3.8 The instruction list or script must be explicit

You may have heard this before, but reproducibility is kind of a big deal in computational science. That means, when you submit your paper, the reviewers and the rest of the world should be able to exactly replicate the analyses from raw data all the way to final results. If you are trying to be efficient, you will likely perform some summarization/data analysis steps before the data can be considered tidy.

3.9 The ideal instruction list is a script

The ideal thing for you to do when performing summarization is to create a computer script (in R, Python, or something else) that takes the raw data as input and produces the tidy data you are sharing as output. You can try running your script a couple of times and see if the code produces the same output.

3.10 If there is no script, be very detailed about parameters, versions, and order of software

In many cases, the person who collected the data has incentive to make it tidy for a statistician to speed the process of collaboration. They may not know how to code in a scripting language. In that case, what you should provide the statistician is something called pseudocode. It should look something like:

- Step 1 - take the raw file, run version 3.1.2 of summarize software with parameters a=1, b=2, c=3
- Step 2 - run the software separately for each sample
- Step 3 - take column three of outputfile.out for each sample and that is the corresponding row in the output data set

You should also include information about which system (Mac/Windows/Linux) you used the software on and whether you tried it more than once to confirm it gave the same results. Ideally, you will run this by a fellow student/labmate to confirm that they can obtain the same output file you did.

3.11 Common mistakes

3.11.1 Combining multiple variables into a single column

A common mistake is to make one column in a data set represent two variables. For example, combining sex and age range into a single variable. .

3.11.2 Merging unrelated data into a single file

If you have measurements on very different topics - for example a person's finance and their health - it is often a better idea to save each one as a different table or data set with a common identifier.

3.11.3 An instruction list that isn't explicit

When an instruction list that is not a computer script is used, a common mistake is to not report the parameters or versions of software used to perform an analysis.

This chapter builds on and expands the book author's data sharing guide[3].

[3] https://github.com/jtleek/datasharing

4. Checking the data

Data munging or processing is required for basically every data set that you will have access to. Even when the data are neatly formatted like you get from open data sources like Data.gov[1], you'll frequently need to do things that make it slightly easier to analyze or use the data for modeling.

The first thing to do with any new data set is to understand the quirks of the data set and potential errors. This is usually done with a set of standard summary measures. The checks should be performed on the rawest version of the data set you have available. A useful approach is to think of every possible thing that could go wrong and make a plot of the data to check if it did.

4.1 How to code variables

When you put variables into a spreadsheet there are several main categories you will run into depending on their data type:

- Continuous
- Ordinal
- Categorical
- Missing
- Censored

[1] http://www.data.gov/

Continuous variables are anything measured on a quantitative scale that could be any fractional number. An example would be something like weight measured in kg. Ordinal data are data that have a fixed, small (< 100) number of possible values, called levels, that are ordered. This could be for example survey responses where the choices are: poor, fair, good.

Categorical data are data where there are multiple categories, but they aren't ordered. One example would be sex: male or female.

Missing data are data that are missing and you don't know the mechanism. You should use a single common code for all missing values (for example, "NA"), rather than leaving any entries blank.

Censored data are data where you know the missingness mechanism on some level. Common examples are a measurement being below a detection limit ora patient being lost to follow-up. They should also be coded as NA when you don't have the data. But you should also add a new column to your tidy data called, "VariableNameCensored" which should have values of TRUE if censored and FALSE if not.

4.2 In the code book you should explain why censored values are missing.

It is absolutely critical to report if there is a reason you know about that some of the data are missing. The statistical models used to treat missing data and censored data are completely different.

4.3 Avoid coding categorical or ordinal variables as numbers.

When you enter the value for sex in the tidy data, it should be “male” or “female”. The ordinal values in the data set should be “poor”, “fair”, and “good” not 1, 2 ,3. This will avoid potential mixups about which direction effects go and will help identify coding errors.

4.4 Always encode every piece of information about your observations using text.

For example, if you are storing data in Excel and use a form of colored text or cell background formatting to indicate information about an observation (“red variable entries were observed in experiment 1.”) then this information will not be exported (and will be lost!) when the data is exported as raw text. Every piece of data should be encoded as actual text that can be exported. For example, rather than highlighting certain data points as questionable, you should include an additional column that indicates which measurements are questionable and which are not.

4.5 Identify the missing value indicator

There are a number of different ways that missing values can be encoded in data sets. Some common choices are “NA”, “.”,

"999", and "-1". There is sometimes also a missing data indicator variable. Missing values coded as numeric values like "-1" are particularly dangerous to an analysis as they will skew any results. The best ways to find the missing value indicator are to go through the code book or to make histograms and tables of common values. In the R programming language be sure to use `useNA` argument to highlight missing values `table(x,useNA="ifany")`.

4.6 Check for clear coding errors

It is common for variables to be miscoded. For example, a variable that should take values 0,1,2 may have some values of 9. The first step is to determine whether these are missing values, miscodings, or whether the scale was incorrectly communicated. As an example, it is common for the male patients in a clinical study to be labelled as both "men" and "males". This should be consolidated to a single value of the variable.

4.7 Check for label switching

When data on the same individuals are stored in multiple tables, a very common error is to have mislabled data. The best way to detect these mislabelings are to look for logical inconsistencies. For example if the same person is labeled as "male" and "female" in two different tables, that is a potential label switching or coding error.

4.8 If you have data in multiple files, ensure that data that should be identical across files is identical

In some cases you will have the same measurements recorded twice. For example, you may have the sex of a patient recorded in two seperate data tables. You should check that for each patient in the two files the sex is recorded the same.

4.9 Check the units (or lack of units)

It is important to check that all variables you have take values on the unit scale you expect. For example, if you observe the people are listed at 180 inches tall, it is a good bet that the measurement is actually in centimeters. This mistake is so pervasive it even caused the loss of a mars satellite[2]. Histograms and boxplots are good ways to check that the measurements you observe fall on the right scale.

4.10 Common mistakes

4.10.1 Failing to check the data at all

A common temptation in data analysis is to load the data and immediately leap to statistical modeling. Checking the data before analysis is a critical step in the process.

[2] http://www.wired.com/2010/11/1110mars-climate-observer-report/

4.10.2 Encoding factors as quantitative numbers

If a scale is qualitative, but the variable is encoded as 1, 2, 3, etc. then statistical modeling functions may interpret this variable as a quantitative variable and incorrectly order the values.

4.10.3 Not making sufficient plots

A common mistake is to only make tabular summaries of the data when doing data checking. Creating a broad range of data visualizations, one for each potential problem in a data set, is the best way to identify problems.

4.10.4 Failing to look for outliers or missing values

A common mistake is to assume that all measurements follow the appropriate distribution. Plots of the distribution of the data for each measured variable can help to identify outliers.

This chapter builds on and expands the book author's data sharing guide[3].

[3] https://github.com/jtleek/datasharing

5. Exploratory analysis

Exploratory analysis is largely concerned with summarizing and visualizing data before performing formal modeling. The reasons for creating graphs for data exploration are:

- To understand properties of the data
- To inspect qualitative features rather than a huge table of raw data
- To discover new patterns or associations

5.1 Interactive analysis is the best way to explore data

If you want to understand a data set you play around with it and explore it. You need to make plots make tables, identify quirks, outliers, missing data patterns and problems with the data. To do this you need to interact with the data quickly. One way is to analyze the whole data set at once using tools like Hive, Hadoop, or Pig. But an often easier, better, and more cost effective approach is to use random sampling . As Robert Gentleman put it "make big data as small as possible as quickly as possible"[1].

[1] https://twitter.com/EllieMcDonagh/status/469184554549248000

5.2 Plot as much of the actual data as you can

These boxplots look very similar:

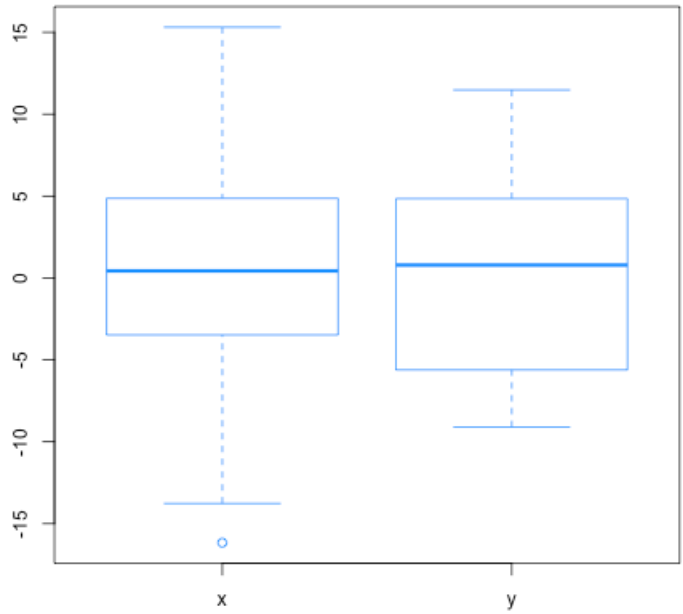

Figure 5.1 Boxplots that look similar

but if you overlay the actual data points you can see that they have very different distributions.

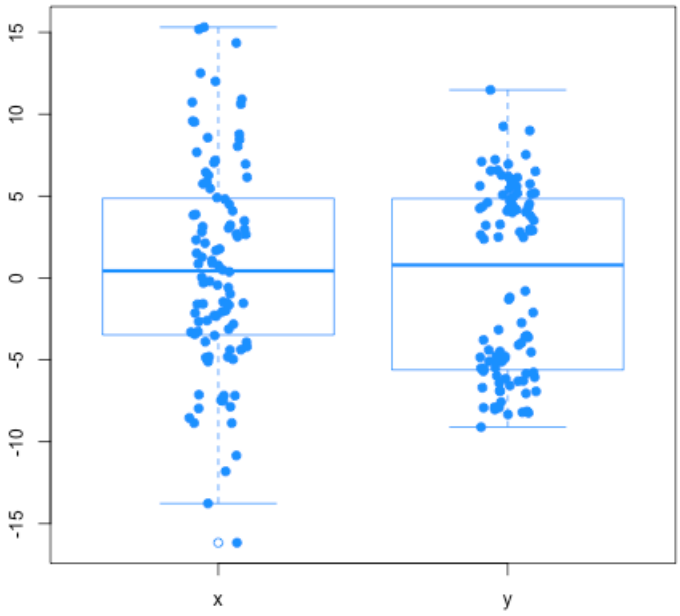

Figure 5.2 Boxplots that look similar with points overlayed

Plotting more of the data allows you to identify outliers, confounders, missing data, relationships, and correlations much more easily than with summary measures.

5.3 Exploratory graphs and tables should be made quickly

Unlike with figures that you will distribute, you are only communicating with yourself. You will be making many

graphs and figures, the goal is speed and accuracy at the expense of polish. Avoid spending time on making axis labels clear or spending time on choosing colors. Find a color palette and sizing scheme you like and stick with it.

5.4 Plots are better than summaries

You can explore data by calculating summary statistics, for example the correlation between variables. However all of these data sets have the exact same correlation and regression line[2].

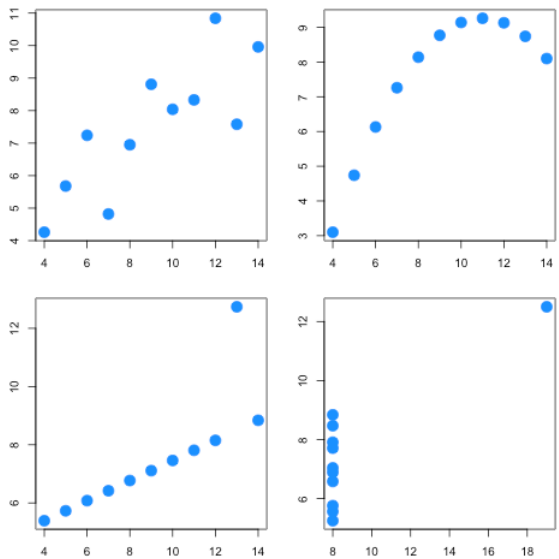

Figure 5.3 Data sets with identical correlations and regression lines

[2] http://en.wikipedia.org/wiki/Anscombe%27s_quartet

This means that it is often better to plot, rather than summarize, the data.

5.5 For large data sets, subsample before plotting

In general, most trends you care about observing will be preserved in a random subsample of the data.

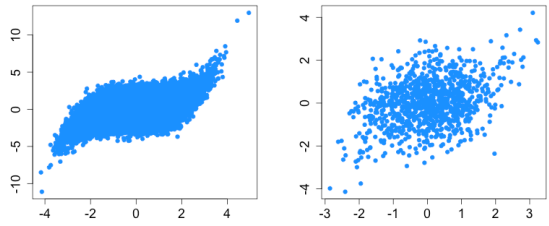

Figure 5.4 A large data set and a random subsample

5.6 Use color and size to check for confounding

When plotting the relationship between two variables on a scatterplot, you can use the color or size of points to check for a confounding relationship. For example in this plot it looks like the more you study the worse score you get on the test:

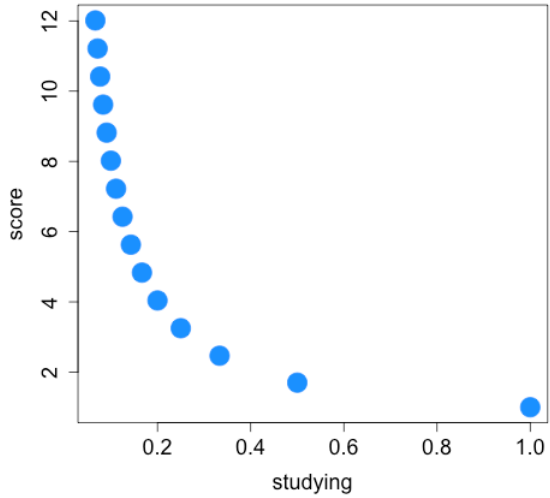

Figure 5.5 Studying versus score

but if you size the points by the skill of the student you see that more skilled students don't study as much. So it is likely that skill is confounding the relationship

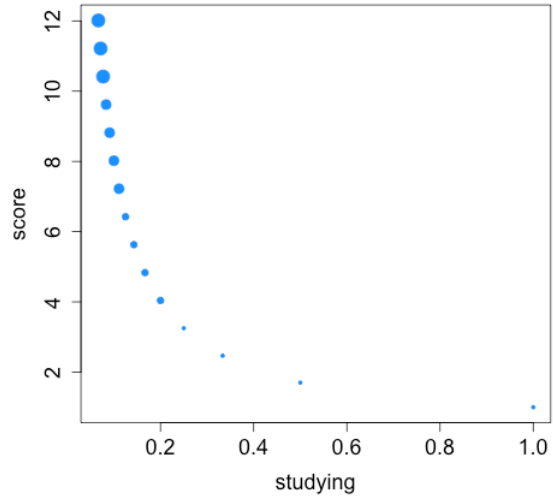

Figure 5.6 Studying versus score with point size by skill level

5.7 For multi-panel plots of the same data type fix the axis

When making multiple plots of the same data type for comparison purposes, having varying axes will place the points on different scales and make them difficult to interpret.

5.8 For multi-panel plots match the comparison axis

If you are comparing two plots on their y-axis values, then the plots should be side by side. If you are comparing the two plots

on their x-axis values the plots should be placed vertically on top of each other.

5.9 Use log transforms to "spread out" data with varying orders of magnitude

Data measured across multiple scales will often be highly skewed, with many values near zero. One way to "spread the values out" is to take the log transform of the data. To avoid zeros you can take `log(data + c)` where `c` is a small positive constant. But be careful about interpreting the spread of the data on the log scale.

5.10 Use log transforms for ratio measurements

Taking the log of ratio-valued data will often make the distribution more symmetric. Since the log of one is also zero, values of zero on the log scale can be interpreted as equality of the two values in the ratio.

5.11 When comparing two measurements of the same thing - use Bland Altman plots

If you have two ways of measuring the same quantity and you want to see if they agree, use a Bland-Altman plot[3]. So instead of plotting `x` versus `y`, plot `x+y` versus `x-y`. Then differences between the two variables will be on the y axis - so you hope all the values will be near zero (Figure 5.7, adapted from gismoskater97[4]). It also makes it easier to see if there are magnitude effects, for example when small values of x and y are more similar than large values of x and y.

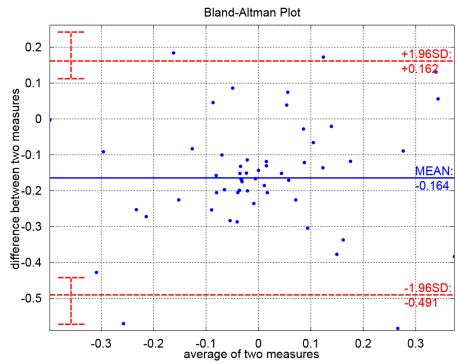

Figure 5.7 A Bland-Alman Plot with confidence intervals

[3] http://en.wikipedia.org/wiki/Bland%E2%80%93Altman_plot

[4] https://en.wikipedia.org/wiki/File:Bland-Altman_Plot.svg

5.12 Common mistakes

5.12.1 Optimizing style too quickly

The goal is to quickly understand a data set, during exploratory data analysis speed is more important than style, so tools that make beautiful graphics but take time should be avoided.

5.12.2 False pattern recognition

One of the most common mistakes in exploratory data analysis is to identify and interpret a pattern without trying to break it down. Any strong pattern in a data set should be checked for confounders and alternative explanations.

5.12.3 Failing to explore data and jumping to statistical tests

A common failure, particularly when using automated software, is to immediately apply statistical testing procedures and to look for statistical significance without exploring the data first.

5.12.4 Failing to look at patterns of missing values and the impact they might have on conclusions.

Missing data are often simply ignored by statistical software, but this means that if the missing data have informative patterns, then analyses will ultimately be biased. As an example,

suppose you are analyzing data to identify a relationship between geography and income in a city, but all the data from suburban neighborhoods are missing.

6. Statistical modeling and inference

The central goal of statistical modeling is to use a small subsample of individuals to say something about a larger population. The reasons for taking this sample are often the cost or difficulty of measuring data on the whole population. The subsample is identified with probability (Figure 6.1).

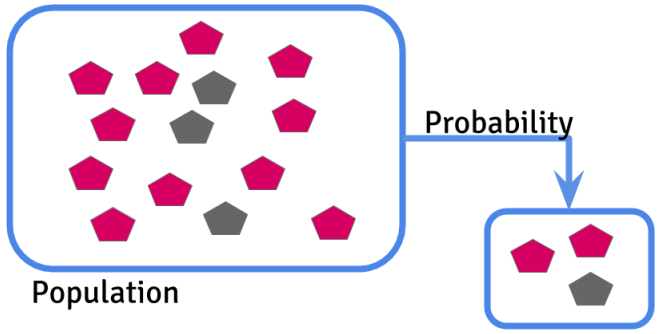

Figure 6.1 Probability is used to obtain a sample

Statistical modeling and inference are used to try to generalize what we see in the sample to the population. Inference involves two separate steps, first obtaining a best estimate for what we expect in the population (Figure 6.2).

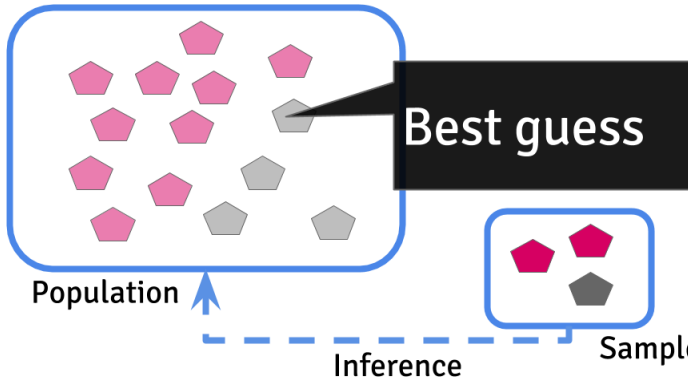

Figure 6.2 The first step in inference is making a best estimate of what is happening in the population

Inference is also used to quantify how uncertain we are about the quantity we are estimating (Figure 6.3). This uncertainty could be due to multiple sources including the fact that we only have a sample of the population, the technology we used to measure the data, or natural variation between the individuals being measured. Good statistical models account for all of these sources of variation.

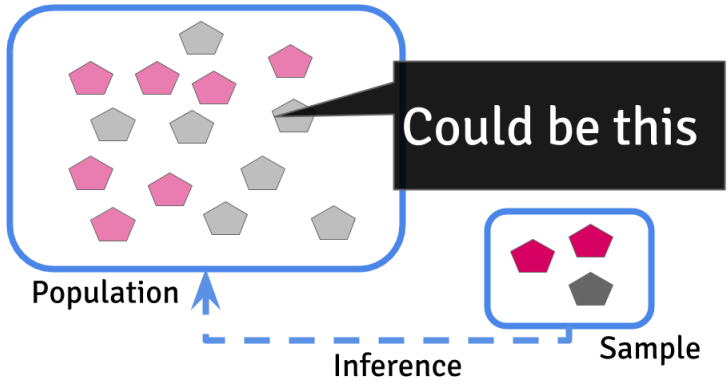

Figure 6.3 The second step in inference is estimating how uncertain we are about our estimate

6.1 When possible, perform exploratory and confirmatory analysis on separate data sets

If you plan to explore a data set for new relationships without an a priori model or hypothesis, split your data into two random subsamples. Perform exploration on the first part of the data and confirm that any relationships you detect appear in the second part at the end of your analysis. A typical split is 70% for discovery and 30% for validation. This splitting is ideal it may lead to loss of power and may not always be possible.

6.2 Define the population, sample, individuals and data

In a clinical trial, for example, you might say that the population is all people who might take a drug for diabetes. The sample is the subset of people that we will enroll in the trial, the individuals (sometimes called sampling units) are the people in the trial, and the data are the measurements on the people.

6.3 Identify reasons your sample may not represent the population

If you take a completely random sample from the population it will be representative or it will be possible to understand how it is not representative of the population. However, it is often impossible to sample randomly from a population. So people make samples based on convenience, cost, and time. Be sure to report all reasons the sample may not represent the population.

6.4 Identify potential confounders

If you measure the shoe size and literacy of a random subsample of people from the United States, they will be correlated. The reason is that young people are less literate than old

people and also have smaller shoes. Age is related to both literacy and shoe size and is a confounder for that relationship. When you observe a correlation or relationship in a data set, consider the potential confounders - variables associated with both variables you are trying to relate.

6.5 Check the distribution of missing data

Determine whether missing values are associated with any of the variables you have in your data set. Associations detected between two variables may be distorted if the presence of missing values for one variable is correlated with the second.

6.6 Check for outliers

Values that are outside of the common range for a single variable may lead to apparently large relationships in summary statistics like correlations or regression coefficients.

6.7 Confirm that estimates have reasonable signs and magnitudes

If you observe that for every increase of one year in education you get an estimate of $1,000,000 more dollars of yearly income, it is likely there is a problem with the data set or your analysis. Similarly if people who exercise more are dramatically more obese than those who do not, it suggests there may be a data problem.

6.8 Be careful of very small or very large samples

In both very small and very large sets of data, estimates of uncertainty should be used with caution. In very small data sets there is not enough data to make an accurate measure of uncertainty. In very large data sets, the measure of uncertainty will be accurate, but the question may no longer be how uncertain we are about estimates - we will be very certain about them. In this case one can focus exclusively on whether the value of the estimate is meaningful, or meaningfully different from some other value - rather than potential uncertainties.

6.9 When performing multiple hypothesis tests, correct for multiple testing

Classic hypothesis tests are designed to call results significant 5% of the time, even when the null is true (e.g. nothing is going on). One really common choice for correcting for multiple testing is to use the false discovery rate[1] to control the rate at which things you call significant are false discoveries. People like this measure because you can think of it as the rate of noise among the signals you have discovered. This error rate is most widely used when there are potentially many true discoveries.

Another common error rate is the family wise error rate[2]

[1] http://en.wikipedia.org/wiki/False_discovery_rate

[2] http://en.wikipedia.org/wiki/Familywise_error_rate

which is the probability of making even one false significant call among all the tests you perform. The standard approach to controlling the family wise error rate is the Bonferroni correction.

More test results will be called significant with the false discovery rate than with the family wise error rate, but the significant results may include more false positives.

6.10 Smooth when you have data measured over space, distance, or time

This is one of the oldest ideas in statistics - regression is a form of smoothing. I personally like locally weighted scatterplot smoothing a lot. Some common smoothers are smoothing splines[3], moving averages, and loess[4].

6.11 Know your real sample size

It can be easy to be tricked by the size of a data set. Imagine you have an image of a simple black circle on a white background stored as pixels. As the resolution increases the size of the data increases, but the amount of information may not (hence vector graphics). Similarly in genomics, the number of reads you measure (which is a main determinant of data size) is not the sample size, it is the number of individuals. In social networks, the number of people in the network may not

[3] http://en.wikipedia.org/wiki/Smoothing_spline

[4] http://en.wikipedia.org/wiki/Local_regression

be the sample size. If the network is very dense, the sample size might be much less. In general the bigger the sample size the better and sample size and data size aren't always tightly correlated.

6.12 Common errors

6.12.1 Failing to account for dependencies

If data are measured across time, across space they will likely be dependent. Before performing inference each variable should be plotted versus time to detect dependencies, and similarly for space. Similarly, identifying potential confounders should occur before model fitting.

6.12.2 Focusing on p-values over confidence intervals

P-values can be a useful measure of statistical significance if used properly. However, a p-value alone is not sufficient for any convincing analysis. A measure of inference on a scientific scale (such as confidence intervals or credible intervals) should be reported and interpreted with every p-value.

6.12.3 Inference without exploration

A very common mistake is to move directly to model fitting and calculation of statistical significance. Before these steps, it is critical to tidy, check, and explore the data to identify dataset specific conditions that may violate your model assumptions.

6.12.4 Assuming the statistical model fit is good

Once a statistical model is fit to data it is critical to evaluate how well the model describes the data. For example, with a linear regression analysis it is critical to plot the best fit line over the scatterplot of the original data, plot the residuals, and evaluate whether the estimates are reasonable. It is ok to fit only one statistical model to a data set to avoid data dredging, as long as you carefully report potential flaws with the model.

6.12.5 Drawing conclusions about the wrong population

When you perform inference, the goal is to make a claim about the larger population you have sampled from. However, if you infer to the wrong population or if the population changes after you take your sample, all of your results will be biased (Figure 6.4). A recent example is when Google Flu trends created a prediction algorithm to predict flu cases based on search terms. When people started searching differently, the prediction algorithm broke down[5].

[5] http://gking.harvard.edu/files/gking/files/0314policyforumff.pdf

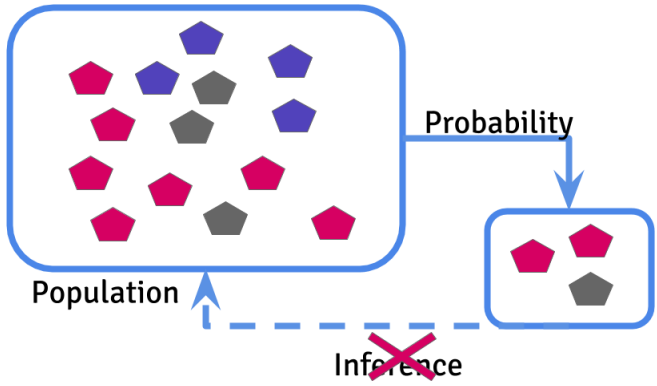

Figure 6.4 If you infer to the wrong population bias will result.

6.12.6 Not addressing uncertainty

If you report an estimate without a measure of uncertainty, then you are claiming that you know the exact value of that parameter. For example, if you report the average height of men in the United States is 5'8" based on your analysis with no measure of uncertainty, then you are claiming you measured the height of every man in the United States and know the exact value.

6.12.7 Identifying real correlations you don't care about

In many cases, it is possible to identify two variables that are really correlated due to a confounder or some spurious relationship. A common example is the relationship between ice cream sales and murder in large cities. This correlation is real and reproducible across cities, but is due to the fact that

high temperatures lead to both more murders and more ice cream sales. While this correlation is real, it isn't one you care about. This is related to the problem of not reporting confounders.

7. Prediction and machine learning

The central idea with prediction is to take a sample from a population - like with inference - and create a training data set. Some variables measured on the individuals in the training set are called features and some are outcomes. The goal of prediction is to build an algorithm or prediction function that automatically takes the feature data from a new individual and makes a best guess or estimate about the value of the outcome variables (Figure 7.1).

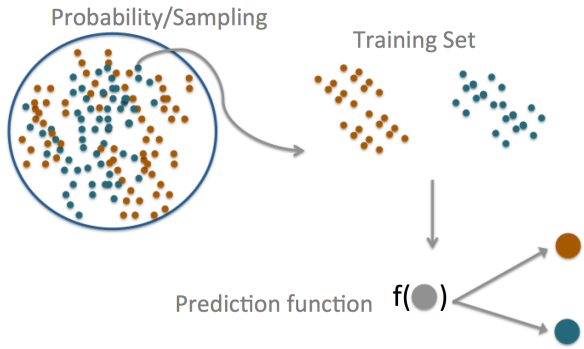

Figure 7.1 How prediction works

7.1 Split the data into training and validation sets

Before performing any analysis at all, split the data into training and validation sets. A typical split is 70% in training and 30% in validation. Then put the validation data set aside and ignore it until the very end. Once you have finalized all parts of your model, apply it once and only once to the validation data set to estimate the real error rate of your prediction algorithm.

7.2 Identify reasons your sample may not represent the population

If you take a completely random sample from the population it will be representative. However, it is often impossible to sample randomly from a population. So people make samples based on convenience, cost, and time. If the sample used to create the training set differs from the population then prediction functions will fail. An example is when Google Flu trends used search terms to try to predict flu outbreaks. When people started using search terms differently and the population changed, the prediction algorithm failed[1].

[1] http://gking.harvard.edu/files/gking/files/0314policyforumff.pdf

7.3 More data usually beats better algorithms

In general collecting more, or more appropriate, data for a prediction algorithm will improve accuracy much more than improving the prediction algorithm itself. This has been called "the unreasonable effectiveness of data"[2].

7.4 Features are more important than the algorithm

In general many out of the box prediction algorithms perform very similarly on most data sets[3]. The best way to improve accuracy is often to pick better data and variables.

7.5 Define your error measure first

You may directly try to optimize a specific error rate, say by minimizing root mean squared error. Or you may choose to weight different types of errors differently. But you should define your error measure before starting to model.

Accuracy, sensitivity, and specificity are typical for binary outcomes (Figure 7.2, adapted from Wikipedia Sensitivity and Specificty[4] for binary outcomes]).

[2] https://www.youtube.com/watch?v=yvDCzhbjYWs

[3] http://arxiv.org/pdf/math/0606441.pdf

[4] http://en.wikipedia.org/wiki/Sensitivity_and_specificity

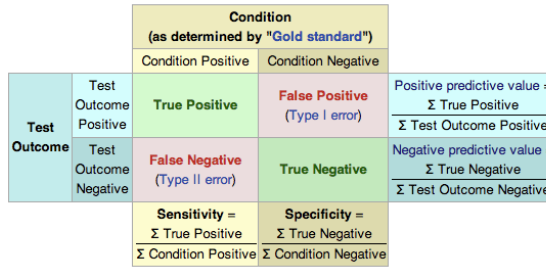

Figure 7.2 Common binary error rates

Root mean squared error is typical for continuous outcomes. That is the square root of the sum of the squared difference between the prediction and the true value.

7.6 Avoid overfitting with cross validation

To avoid tuning your model too closely to the observed data use cross-validation[5] or subsampling. The basic idea is to split your training set up into two sets randomly, build different models on the first piece, and try them on the second set, picking the model that performs best.

7.7 If the goal is prediction accuracy, average many prediction models together.

In general, the prediction algorithms that most frequently win prediction competitions blend multiple models together

[5] http://en.wikipedia.org/wiki/Cross-validation_%28statistics%29

by averaging. The idea is that by averaging (or majority voting) multiple good prediction algorithms you can reduce variability without giving up bias. One of the earliest descriptions of this idea was of a much simplified version based on bootstrapping samples and building multiple prediction functions - a process called bagging[6] (short for bootstrap aggregating). Random forests[7], another successful prediction algorithm, is based on a similar idea with classification trees.

7.8 Prediction is about tradeoffs

- Interpretability versus accuracy
- Speed versus accuracy
- Simplicity versus accuracy
- Scalability versus accuracy

In some areas one of these components may be more important than others. The Netflix prize awarded $1,000,000 to the best solution to predicting what movies people would like. But the winning solution was so complicated it was never implemented by Netflix because it was impractical[8].

[6] http://statistics.berkeley.edu/sites/default/files/tech-reports/421.pdf

[7] http://en.wikipedia.org/wiki/Random_forest

[8] https://www.techdirt.com/blog/innovation/articles/20120409/03412518422/

8. Causality

The gold standard for causal data analysis is to combine specific experimental designs such as randomized studies with standard statistical analysis techniques. If correctly performed, the experimental design makes it possible to identify how variables affect each other on average.

8.1 Causal data analysis of non-randomized experiments is often difficult to justify.

Causal data analysis in observational studies requires both more technical statistical analyses and belief that untestable and often unrealistic assumptions are met. Phenomena that are not easily studied with randomized studies, such as pregnancy or carcinogenesis, frequently produce spurious causal data analyses that don't hold up to scrutiny.

8.2 Even randomized studies may be difficult to interpret causally

Some common difficulties in clinical trials are: (1) patient dropout preferentially by treatment, (2) unblinding of trials -

when the experimenters or subjects discover what treatment they are part of, (3) treatments that are difficult to take or adhere to, so just the intent to treat a person must be used as the treatment itself. All of these issues make causal analysis difficult even in randomized studies.

8.3 For randomized studies use exploratory analysis to confirm the randomization "worked"

Make tables and plots to ensure all non-randomized variables measured on each subject have approximately the same distribution between the groups, including missing values.

8.4 Causal data analyses seek to identify average effects between often noisy variables.

Decades of data show a clear causal relationship between smoking and cancer. If you smoke, it is a sure thing that your risk of cancer will increase. But it is not a sure thing that you will get cancer. The causal effect is real, but it is an effect on your average risk.

8.5 Unless you have performed a randomized experiment or use causal techniques avoid causal language

Causality creep is the idea that causal language is often used to describe inferential or predictive analyses. Avoid using words like “cause”, “effect”, or “leads to an increase” if you have not performed a causal analysis.

8.6 Common mistakes

8.6.1 Causality creep

A common mistake is to perform an analysis that looks for a correlation or association between measurements in a non-randomized study, then interprets that correlation as a causal relationship. For example, if you identify the well-known correlation between ice cream scales and murder, interpreting the correlation as “ice cream sales lead to increased homicide rates”.

9. Written analyses

Data analysis is as much about communication as it is about statistics. A written data analysis report should communicate your message clearly and in a way that is readable to non-technical audiences. The goal is to tell a clear, precise and compelling story. Throughout your written analysis you should focus on how each element: text, figures, equations, and code contribute to or detract from the story you are trying to tell.

9.1 The elements of a written analysis

A written analysis should always include

- A title
- An introduction or motivation
- A description of the statistics or machine learning models you used
- Results including measures of uncertainty
- Conclusions including potential problems
- References

9.2 Lead with the question you are answering

Before explaining the data, models, or results, lead with the question that you are trying to answer. The question should be a scientific or business application - not a statistical or computational question. A good example would be, “Can we use tweets about stocks to predict stock prices?”

9.3 Describe the experimental design

Explain where the data came from, how they were collected, and relevant information about the technologies and systems used to collect the data.

9.4 Describe the data set

Explain what processing you did to the data, and the tidy data you produced. It is common to lead with a table summarizing the variables in the tidy data set, including sample sizes, number of variables, averages and variances or standard deviations for each variable. This component of an analysis is critical to identify data versioning issues.

9.5 When describing a statistical model use equations or psuedocode

Every model must be mathematically specified. This model may be specified in the main text of the writing if the audience is statistical or in an appendix if the audience is non-technical.

Modeling procedures should be completely specified using equations or algorithms with explicit definitions of each input and output. Each term and index in the equation should be explained in plain language. When possible, use letters and symbols that are abbreviations of the variables in your model. For example, if modeling the relationship between weight and height you might write the model $W = a + b\,H + e$ and then explain that W stands for weight, a is the weight for a person with 0 height, b is the increase in weight units for a person with one additional height unit, H is the height of the person, and e is measurement error or noise.

9.6 Specify the uncertainty distribution

You should declare the distribution of the error terms like e and the assumptions you are making about those errors - specifically whether they are independent or dependent and whether they have common variance or are heteroskedastic[1].

[1] http://en.wikipedia.org/wiki/Heteroscedasticity

9.7 For each parameter of interest report an estimate and interpretation on the scale of interest

When reporting an estimate do not say that we estimated *a=3*, instead report that we estimated a change of *3* pounds in weight for one inch in height.

9.8 For each parameter report a measure of uncertainty on the scientific scale

For every estimate you report you should report a measure of uncertainty on the scale of interest. Report that a 95% confidence interval for the estimated change in weight for a unit change in height is *1* and *5* inches. Typical measures of uncertainty are standard deviations, confidence intervals, or credible intervals.

9.9 Summarize the importance of reported estimates

When reporting an estimate, also report why you calculated the estimate and what the quantitative value means.

9.10 Report potential problems with the analysis

If you fit a model and you observe that there may be missing data, or that outliers may be driving a particular estimate, report that as part of the analysis.

9.11 Do not report every analysis you performed

Many analyses, particularly exploratory analyses, will not be useful to explaining your result and interpretation. Before including an analysis in the final report, ask whether it contributes to the story or explains a crucial fact about the data set that can't be left out.

9.12 Every statistical or machine learning method should be referenced

When using models it is important to give credit to the person who developed them. It is also important so that the reader can go back to the original source of the model and understand it.

There is a convention that some statistical methods are not cited like maximum likelihood or least squares - but at minimum a reference to a textbook should be made when using even common techniques.

10. Creating figures

Figures ar graphs you are likely to share with other people or include in reports. They are used to communicate with others what we observed or discovered in the data set. The goal for figures is that when viewed with an appropriately detailed caption, they can stand alone without any further explanation as a unit of information. There are several basic principles behind creating useful expository graphs.

10.1 Information should be communicated as often as possible with position, and on common scales.

The reason why statisticians don't like pie-charts is that comparing angles is notoriously hard. In Figure 10.1 which slice is bigger A or D?

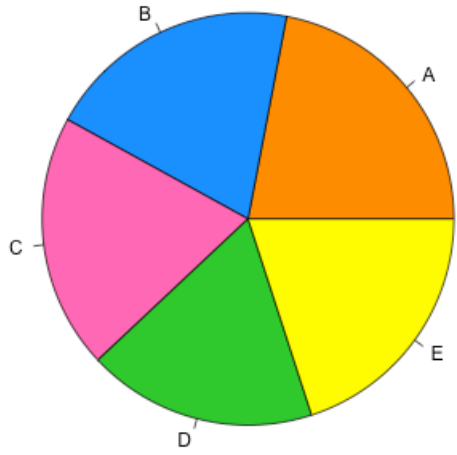

Figure 10.1 A pie chart

If we make this as a barchart it is much easier to see that A is bigger than D (Figure 10.2).

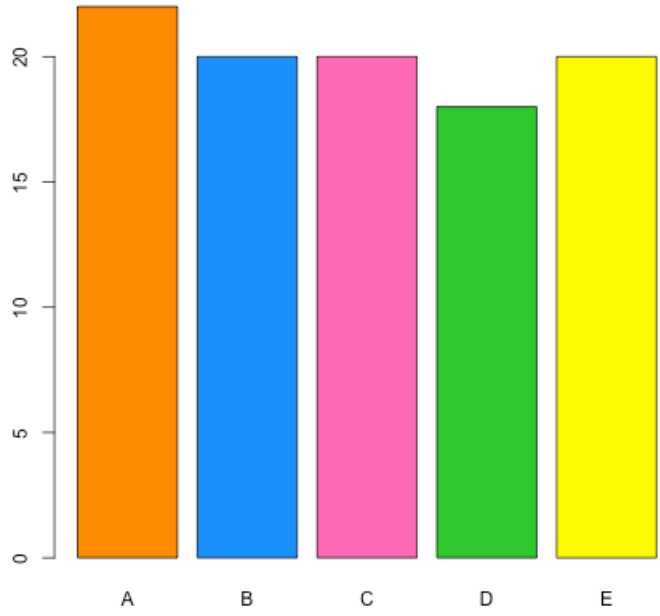

Figure 10.2 A bar chart

This is actually a more general phenomenon. There are a large number of visual tasks you could use when making plots (Figure 10.3, adopted from Graphical perception: Theory, Experimentation, and Applications to the Development of Graphical Models[1]). It turns out that humans are best at perceiving position along a single axis with a common scale.

[1] http://www.jstor.org/discover/10.2307/2288400?uid=3739704&uid=2&uid=4&uid=3739256&sid=21101742782357

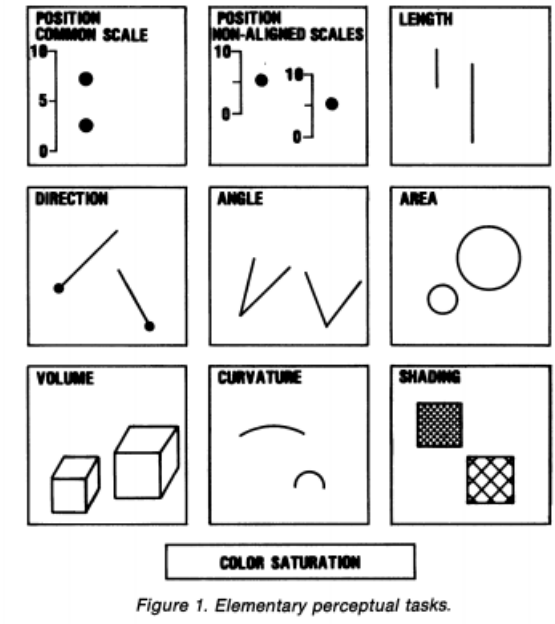

Figure 10.3 Visual tasks

10.2 Low information density should be avoided

The most extreme example of this is a plot with a single point. It is better to just report the data value. But any graph that is primarily white space and doesn't illustrate a key feature of the data should be avoided.

10.3 Gratuitous flourishes should be avoided

When creating figures for papers, it is a good idea to make them visually pleasing with a nice color palette and good choices for the shape of points and fonts. However, gratuitous flourishes like 3-d barcharts, 3-d pie charts, or animation for no reason should be avoided.

10.4 Color and size may both be used to communicate information.

Color and size may both used, be used for communication. A common approach is to plot a scatterplot of two variables, where the points are colored or sized by a third. Color and size should be used to communicate information, although the choices can also be made on the basis of style and with the goal of making figures interesting.

10.5 When there are many values of a third variable use faceting

In some cases, there are too many values of a third variable to view in a single plot. When this happens, facet the plot by making multiple panels, one panel each for different values of the third variable.

10.6 Axis labels should be large, easy to read, in plain language

For figures and expository graphs, axis labels should be in large type and easily readable. In addition, the axis labels should not be variable names. For example instead of using the value "x$hght" or "x[1,]", which may be the default from software, use "Height".

10.7 Include units in figure labels and legends

When labeling a variable in a figure, always include units, for example "Height (cm)".

10.8 Use figure legends

When size or color are used to distinguish points, include a figure legend that labels any scales or colors used in plain language.

10.9 Figure legends in the figure are preferred

When possible, directly label point clouds, distributions, or other features of a plot in the plot itself, without using a separate figure legend.

10.10 Figure titles should communicate the message of the plot

A title should be used to communicate the main take home message of a plot. For example if you plot the relationship between studying and grades, your figure title might read, "More studying is associated with improved grades". Eliminate titles that simply describe the data, for example, "Studying versus grades".

10.11 Label multi-panel plots with numbers or letters

If you make a plot with multiple panels always include numbers or letters for each panel so each component of the plot can be individually referred to without resorting to descriptions like, "the scatterplot in the upper right hand corner".

10.12 Add text to the plot itself to communicate a message

To make a figure stand alone it is often helpful to label specific trends or features of the data that are relevant to the person looking at the graph. For example:

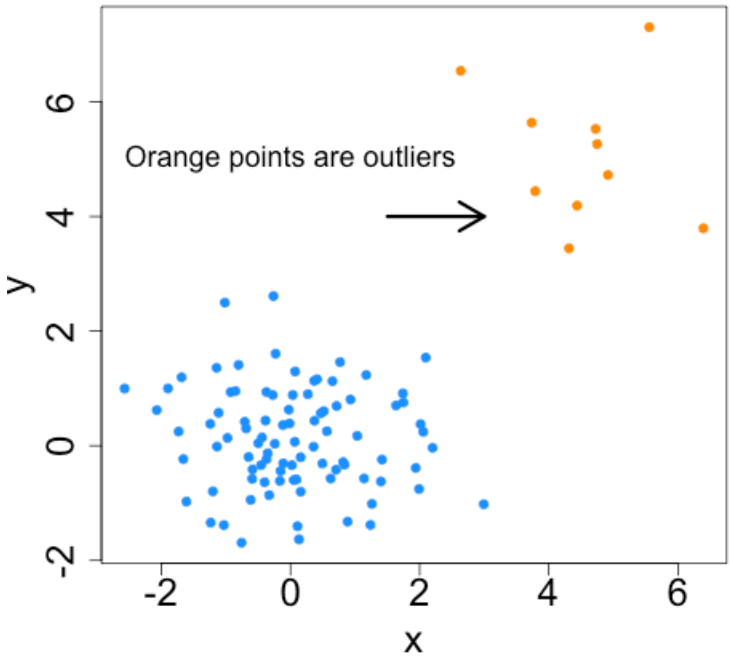

Figure 10.4 An in graph label

10.13 Figure captions should be self-contained

A person should be able to understand from your figure caption everything they need to know about your graph. So you should explain at minimum what the x and y axis are, what their units are, what all colors and sizes in the graph mean, and in particular any trends you expect them to observe in the figure.

10.14 Common errors

10.14.1 Using a color palette that colorblind people can't see

A common mistake is to use red and green to indicate different things on the same plot. There are many web tools[2] available for uploading figures you have made to check whether they are visible to color-blind people.

10.14.2 Using colors that look too similar to each other

In general avoid using colors that appear very similar (red and pink, blue and light blue) for making plots. Try to use colors that are clearly different.

10.14.3 Not making scatterplots when you should

Scatterplots are one of the most informative type of plot since it shows each individual data value. When possible, scatterplots should be used in favor of boxplots or barplots. Even when using these more summarized versions of the data, overlaying the data points can be useful.

10.14.4 Failing to take logs.

When data are highly skewed and you make a plot without taking the log transform first, a large amount of the data

[2] http://www.vischeck.com/

will be clustered in the corner of the plot so most of the patterns in the data will be obscured (Figure 10.5, adopted from Displaying data badly[3].

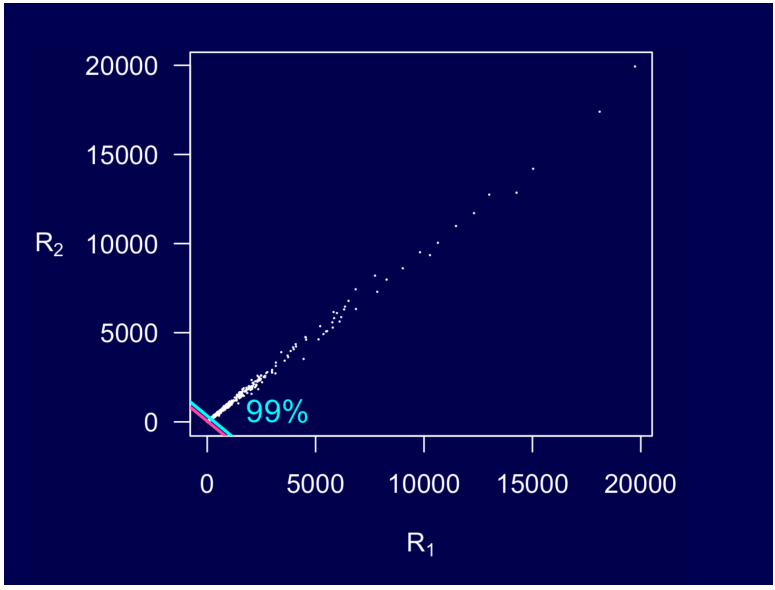

Figure 10.5 Without logs 99% of the data are in the lower left hand corner in this figure from

10.14.5 Using a plot of x versus y when a plot of (x+y) versus (x-y) is more informative

If you have two ways of measuring the same quantity and you want to see if they agree, use a Bland-Altman plot[4]. So instead of plotting `x` versus `y`, plot `x+y` versus `x-y`. Then differences

[3][Displayingdatabadly](https://www.biostat.wisc.edu/~kbroman/presentations/IowaState2013/graphs_combined.pdf)

[4]http://en.wikipedia.org/wiki/Bland%E2%80%93Altman_plot

between the two variables will be on the y axis - so you hope all the values will be near zero. It also makes it easier to see if there are magnitude effects, for example when small values of x and y are more similar than large values of x and y.

10.14.6 Failing to ensure that multiple panels have the same vertical scale

When you have multiple panels make sure all axes are on the same scale otherwise differences will not be correctly perceived.

10.14.7 Failing to consider the point of the graph but rather just using some standard form

In some disciplines certain plots are used by custom. However, each plot in a paper should communicate a message. If they do not, even if they are "customary" they should be eliminated.

10.14.8 Bar plots with antennas

These plots are sometimes called dynamite plots[5] and are very low information density. Instead, scatterplots should be used.

[5] http://emdbolker.wikidot.com/blog:dynamite

10.14.9 Being inconsistent about color choice across multiple panels or figures

If you use color to label points in a figure according to a variable, then every figure or panel in your paper should use those same colors to label points by the same variable.

10.14.10 Aligning things horizontally when comparisons would be better made vertically, or vice versa

When using multiple panels, the panels should be lined up according to the axis for comparison. So if the points are to be compared on the x-axis values they should be aligned vertically and if they are to be compared on the y-axis values they should be aligned horizontally.

10.14.11 Any of the other issues in Karl Broman's presentation on displaying data badly

There are many other common visualization mistakes. A large number of these are summarized in Karl Browman's excellent presentation on displaying data badly[6].

[6] https://www.biostat.wisc.edu/~kbroman/presentations/IowaState2013/graphs_combined.pdf

11. Presenting data

Giving data science talks can help you:

- Meet people
- Get people excited about your ideas/software/results
- Help people understand your ideas/software/results

The importance of the first point can't be overstated. The primary reason you are giving talks is for people to get to know you. Being well known and well regarded can make a huge range of parts of your job easier. So first and foremost make sure you don't forget to talk to people before, after, and during your talk.

Point 2 is more important than point 3. As a data scientist, it is hard to accept that the primary purpose of a talk is advertising, not data science. See for example Hilary Mason's great presentation Entertain, don't teach[1]. Here are reasons why entertainment is more important:

That being said, be very careful to avoid giving a TED talk. If you are giving a data science presentation the goal is to communicate specific ideas. So while you are entertaining, don't forget why you are entertaining.

[1] http://www.hilarymason.com/speaking/speaking-entertain-dont-teach/

11.1 Tailor your talk to your audience

It depends on the event and the goals of the event. Here is a non-comprehensive list:

- **Small group meeting**:
  - **Goal**: Update people you work with on what you are doing and get help.
  - **What to talk about**: Short intro on your problem, brief update on what you've tried, long discussion about where you are going/what you need help on.
- **Short talk at conference**:
  - **Goal**: Entertain people, get people to read your paper/blog or use your software.
  - **What to talk about**: Short intro on your problem, brief explanation of solution, links to software
- **Long format formal talk**:
  - **Goal**: Entertain people, get people to read your software, make them understand your problem/-solution
  - **What to talk about**: Intro to your problem, how you solved it, results, and connection to broader ideas
- **Job talk**:
  - **Goal**: Get a job, entertain people, make them understand your problem/solution
  - **What to talk about**: Brief overview of who you are, intro to your (single) problem, how you solved it, results, summary of what you have done/plan to do.

11.2 Order your talk in story format

The biggest trap in giving a talk is assuming that other people will follow you because you follow the talk. It is key to be aware of what your audience knows and doesn't know. In general it is always better to assume your audience knows less than you think they do. People like to feel smart. I have rarely heard complaints about people who went too basic in their explanations, but frequently hear complaints about people being lost. That being said, here are some structural tips. Your mileage may vary.

- **Always lead with a brief, understandable to everyone statement of the scientific problem that lead to your data**.
- Explain the data and measurement technology before you explain the features you will use
- Explain the features you will use to model data before you explain the model
- When presenting results, make sure you are telling a story.

The last point is particularly important. Usually by the results section people are getting a little antsy. So a completely disparate set of results with little story behind them is going to drive people bonkers. Make sure you explain up front where the results are going (e.g. "Our results will show our method is the fastest/most accurate/best ever."), then make sure that your results are divided into sections by what point they are making and organized just like they would be if you were telling a story.

11.3 Use large fonts

Fonts can never be too big. Go huge. Small fonts will be met with anger

11.4 Include contact information early

Your title slide should have a contactable form of you (twitter handle, email address, etc.)

11.5 All figures should have large axis labels in plain English.

Figure axes should be readable from the back of the room.

11.6 Be sure to attribute all images and text you borrow

Any figure you borrow off the internet should have a web link to the source. Any figure you borrow off a paper should have a web link to the paper. Any time you use someone else's slide you should put "Slide courtesy of So and so" with a link to their page.

11.7 In general use a solid background and opposite color font

Unless you have experience in design, using pictures for backgrounds or similar color backgrounds and fonts are dangerous for presentations.

11.8 Minimize text on slides

Whenever possible break multiple bullet points up into individual slides with only one bullet point on them. If you can, go textless with only images. That way you won't fall into the trap of reading your slides.

11.9 Explain every figure in your talk in detail

If you have a figure in your talk you should present it in the following way.

- Explain what the figure is supposed to communicate (e.g. "this figure shows our method has higher accuracy")
- Explain the figure axes (e.g. "the y-axis is sensitivity the x-axis is 1-specificity")
- Explain what trends the audience should look for (e.g. "curves that go straight up at zero and across the top of the plot are best")

If you don't have time to fully explain a figure, than simplify it, or omit it.

11.10 Use equations to make ideas concrete, but use them sparingly

If you are giving a data talk you will often have some equations. That is ok. But the way you present them is critically important to giving an understandable talk. Here are a few important points:

- Before presenting an equation explain the data
- Whenever possible use words instead of symbols in equations (*Expression = Noise + Signal* is better than $E = N + S$ is better than $Y = X + E$)
- Use no more than two subscripts whenever possible
- When explaining an equation
  - First explain the point of the equation ("we are trying to model expression as a function of noise and signal")
  - Then explain what each symbol is ("E is expression, N is noise, etc.")
  - Then explain how the model relates them ("E is a linear function of signal and noise")

11.11 Be willing to say "I don't know"

Inevitably you will get hard questions during your talk. The most important point is not to panic and not to get defensive. It is way better to just say *I don't know*, then to get upset.

When you get asked a really hard question you should:

- Take a second and a deep breath. If necessary, ask the person to repeat the question.
- Answer the question the best you can come up with on the spot
- Say *I don't know* if you don't know. If you say *I don't know*, then you can give your best guess and explain it is a guess.

11.12 Distiguish your response type when answering questions

The key is to distinguish what kind of response you are giving. Are you giving a response where you know the answer because you actually looked into that? Are you giving a complete guess where you have no idea? Or, what is more likely, are you somewhere in between?

11.13 Never be aggressive

Almost everywhere you give a talk there will be a person who is upset/intense/aggressive. **Do not fall into the temptation**

to be aggressive in return. Answer their first questions politely just like everyone else. If they keep asking really aggressive/lots of questions, you are within your rightst to say: "You are bringing up a lot of good issues, I'd be happy to discuss with you after the presentation more in depth, but in the interest of time I'm going to move on to the next part of my talk". It is ok to keep things moving and to finish on time.

11.14 Finish on time

Do it. People will love you for it. If you are the last speaker in a session and others have gone long, adapt and go shorter. The value you gain by making your audience happy >>>> the extra 5 minutes of details you could have explained.

11.15 Where you should put your talk

If you have weblinks in your talk you need to post it online. There is no way people will write down any of the links in your talk. Two good places to put your talks are

- https://speakerdeck.com/[2]
- http://www.slideshare.net/[3].

Slideshare is easier to view on mobile phones and iPads, which is the most likely place someone will read your talk.

[2] https://speakerdeck.com/

[3] http://www.slideshare.net/

This chapter builds on and expands the book author's talk guide[4].

[4] https://github.com/jtleek/talkguide

12. Reproducibility

Reproducibility involves being able to recalculate the exact numbers in a data analysis using the code and raw data provided by the analyst. Reproducibility is often difficult to achieve and has slowed down the discovery of important data analytic errors[1]. Reproducibility should not be confused with "correctness" of a data analysis. A data analysis can be fully reproducible and recreate all numbers in an analysis and still be misleading or incorrect.

12.1 Have a data analysis script

The primary way to ensure an analysis is reproducible is to create code scripts that exactly generate all numbers and figures in an analysis. R markdown[2] and iPython notebooks[3] are both useful for organizing scripts. These tools integrate code, plain text, and figures into one document and are examples of what is called, "literate programming"[4]. An R markdown document includes text, code, and figures and can easily be converted into pdf or HTML format (Figure 12.1, adapted from R markdown documentation[5]).

[1] http://www.cbsnews.com/news/deception-at-duke-fraud-in-cancer-care/

[2] http://rmarkdown.rstudio.com/

[3] http://ipython.org/notebook.html

[4] http://en.wikipedia.org/wiki/Literate_programming

[5] http://rmarkdown.rstudio.com/

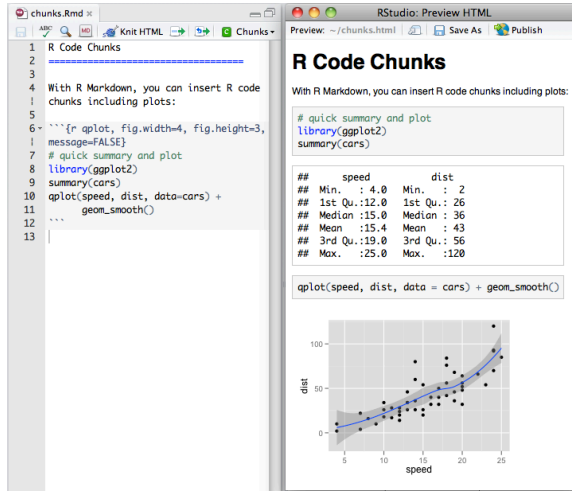

Figure 12.1 An example of an R markdown document

12.2 Record versions of software and parameters

If you are using open-source software such as R or Python, then packages will often be updated and their exact behavior may change. At minimum you should record what versions of the software you used to perform an analysis. In R you can get this information with the `sessionInfo()` command or the `devtools::session_info()` command, which is more informative.

12.3 Organize your data analysis

A good organizational structure for data analysis files is to keep separate folders for:

- Data
  - raw data
  - processed data
- Figures
  - Exploratory figures
  - Final figures
- R code
  - Raw or unused scripts
  - Data processing scripts
  - Analysis scripts
- Text
  - README files explaining what all the components are
  - Final data analysis products like presentations/writeups

12.4 Use version control

When performing analysis it is a good idea to keep track of the version of the software you are writing. One very popular approach is to use Github[6]. You can put your R scripts, R markdown documents, and small data sets on Github, then use the version control system Git to manage your analysis. Github is an example of a distributed version control system where a version of the files is on your local computer and also available at a central server (Figure 12.2, adapted from git documentation[7]). Multiple people can then work on the analysis code and push changes to the central server.

[6] https://github.com/

[7] http://git-scm.com/book/en/v2/Getting-Started-About-Version-Control

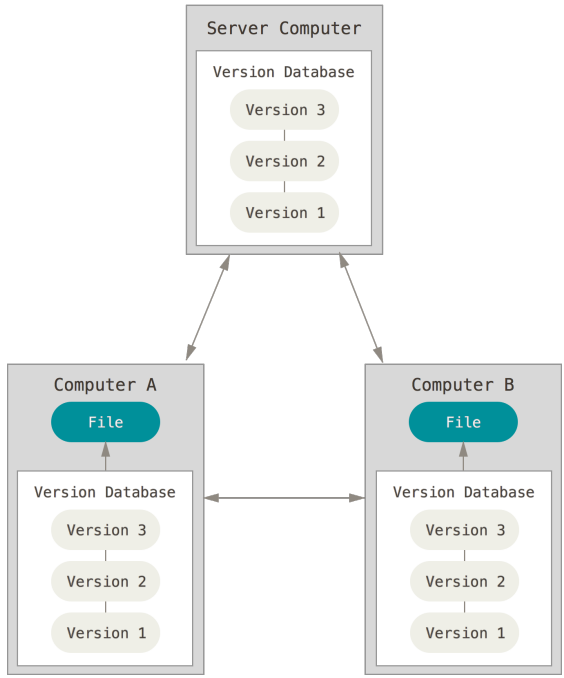

Figure 12.2 Distributed version control

Using version control will help you keep track of changes to your analysis and their impact on the results.

12.5 Set a seed

When your calculations involve random sampling - for example if you use the bootstrap or permutations - then you need to set a random number seed. This will ensure that the pseudorandom numbers generated by the computer will always be the same when different people run your analysis. In R you can do this with the command `set.seed(13323)` where 13323 can be replaced with any positive integer.

12.6 For large data sets, save intermediate results and especially how you got them

If you are working with large data sets, create a script to calculate summary measures. Include all the same details about version numbers and parameters. Then save the intermediate data set in a tidy format in a csv or tab-delimited file for easy sharing.

12.7 Have someone else run your analysis

The best way to check reproducibility is to have someone else, ideally on a different type of computer, run your data analysis scripts. They should be able to run the analysis and get the same figures and results without having to ask you questions.

This chapter builds on and expands the book author's data sharing guide[8].

12.8 Common mistakes

12.8.1 Not using a script for your analysis

If you describe your analysis in written documentation, it is much easier to make mistakes of reproducibility.

[8] https://github.com/jtleek/datasharing

12.8.2 Not recording version numbers or parameters used

It is important to record: (1) the type of computer used, (2) the version of all software used, and (3) all parameters you used when performing an analysis.

12.8.3 Not sharing data or code

For every analysis you perform you should include a link to the code and data you used to perform the analysis.

12.8.4 Using reproducibility as a weapon

If you reproduce someone else's analysis and identify a problem, bug or mistake you should contact them and try to help them resolve the problem rather that pointing the problem out publicly or humiliating them.

13. A few matters of form

- Report estimates followed by parentheses.

  The increase is 5.3 units (95% CI: 3.1, 4.3 units)

- When reporting P-values do not report numbers below machine precision. P-values less than 2 x 10e-16 are generally below machine precision and inaccurate.

  Reporting a P-value of 1.35 x 10e-25 is effectively reporting a P-value of 0 and caution should be urged. A common approach is to report censored P-values such as P < 1 x 10e-8.

- When reporting permutation P-values avoid reporting a value of zero.

  P-values should be calculated as (K + 1)/(B + 1)[1] *where B is the number of permutations and K is the number of times the null statistic is more extreme than the upper bound.*

- Do not report estimates with over-precision.

[1] http://www.statsci.org/webguide/smyth/pubs/permp.pdf

If measurements are only accurate to the tenths digit, do not report an estimate of 6.8932

- When programing variable names should be lower case, with words separated by underscores, and as explicit as possible in data frames you are analyzing.

  The date of visiting a website might be named date_of_visit.

- In written analysis variable names should always be reported in plain language, not as variable names.

  The date of visiting a website would be described as "the date of visit variable".

14. The data analysis checklist

This checklist provides a condensed look at the information in this book. It can be used as a guide during the process of a data analysis, as a rubric for grading data analysis projects, or as a way to evaluate the quality of a reported data analysis.

14.1 Answering the question

1. Did you specify the type of data analytic question (e.g. exploration, association causality) before touching the data?
2. Did you define the metric for success before beginning?
3. Did you understand the context for the question and the scientific or business application?
4. Did you record the experimental design?
5. Did you consider whether the question could be answered with the available data?

14.2 Checking the data

1. Did you plot univariate and multivariate summaries of the data?
2. Did you check for outliers?
3. Did you identify the missing data code?

14.3 Tidying the data

1. Is each variable one column?
2. Is each observation one row?
3. Do different data types appear in each table?
4. Did you record the recipe for moving from raw to tidy data?
5. Did you create a code book?
6. Did you record all parameters, units, and functions applied to the data?

14.4 Exploratory analysis

1. Did you identify missing values?
2. Did you make univariate plots (histograms, density plots, boxplots)?
3. Did you consider correlations between variables (scatterplots)?
4. Did you check the units of all data points to make sure they are in the right range?
5. Did you try to identify any errors or miscoding of variables?
6. Did you consider plotting on a log scale?
7. Would a scatterplot be more informative?

14.5 Inference

1. Did you identify what large population you are trying to describe?

2. Did you clearly identify the quantities of interest in your model?
3. Did you consider potential confounders?
4. Did you identify and model potential sources of correlation such as measurements over time or space?
5. Did you calculate a measure of uncertainty for each estimate on the scientific scale?

14.6 Prediction

1. Did you identify in advance your error measure?
2. Did you immediately split your data into training and validation?
3. Did you use cross validation, resampling, or bootstrapping only on the training data?
4. Did you create features using only the training data?
5. Did you estimate parameters only on the training data?
6. Did you fix all features, parameters, and models before applying to the validation data?
7. Did you apply only one final model to the validation data and report the error rate?

14.7 Causality

1. Did you identify whether your study was randomized?
2. Did you identify potential reasons that causality may not be appropriate such as confounders, missing data, non-ignorable dropout, or unblinded experiments?
3. If not, did you avoid using language that would imply cause and effect?

14.8 Written analyses

1. Did you describe the question of interest?
2. Did you describe the data set, experimental design, and question you are answering?
3. Did you specify the type of data analytic question you are answering?
4. Did you specify in clear notation the exact model you are fitting?
5. Did you explain on the scale of interest what each estimate and measure of uncertainty means?
6. Did you report a measure of uncertainty for each estimate on the scientific scale?

14.9 Figures

1. Does each figure communicate an important piece of information or address a question of interest?
2. Do all your figures include plain language axis labels?
3. Is the font size large enough to read?
4. Does every figure have a detailed caption that explains all axes, legends, and trends in the figure?

14.10 Presentations

1. Did you lead with a brief, understandable to everyone statement of your problem?
2. Did you explain the data, measurement technology, and experimental design before you explained your model?

3. Did you explain the features you will use to model data before you explain the model?
4. Did you make sure all legends and axes were legible from the back of the room?

14.11 Reproducibility

1. Did you avoid doing calculations manually?
2. Did you create a script that reproduces all your analyses?
3. Did you save the raw and processed versions of your data?
4. Did you record all versions of the software you used to process the data?
5. Did you try to have someone else run your analysis code to confirm they got the same answers?

14.12 R packages

1. Did you make your package name "Googleable"
2. Did you write unit tests for your functions?
3. Did you write help files for all functions?
4. Did you write a vignette?
5. Did you try to reduce dependencies to actively maintained packages?
6. Have you eliminated all errors and warnings from R CMD CHECK?

15. Additional resources

15.1 Class lecture notes

- Johns Hopkins Data Science Specialization[1] and Additional resources[2]
- Data wrangling, exploration, and analysis with R[3]
- Tools for Reproducible Research[4]
- Data carpentry[5]

15.2 Tutorials

- Git/github tutorial[6]
- Make tutorial[7]
- knitr in a knutshell[8]
- Writing an R package from scratch[9]

[1] https://github.com/DataScienceSpecialization/courses
[2] http://datasciencespecialization.github.io/
[3] https://stat545-ubc.github.io/
[4] http://kbroman.org/Tools4RR/
[5] https://github.com/datacarpentry/datacarpentry
[6] http://kbroman.org/github_tutorial/
[7] http://kbroman.org/minimal_make/
[8] http://kbroman.org/knitr_knutshell/
[9] http://hilaryparker.com/2014/04/29/writing-an-r-package-from-scratch/

15.3 Leek group guides

- To data sharing[10]
- To giving talks[11]
- To developing R packages[12]

15.4 Books

- An introduction to statistical learning[13]
- Advanced data analysis from an elementary point of view[14]
- Advanced R programming[15]
- OpenIntro Statistics[16]
- Statistical inference for data science[17]

[10] https://github.com/jtleek/datasharing
[11] https://github.com/jtleek/talkguide
[12] https://github.com/jtleek/rpackages
[13] http://www-bcf.usc.edu/~gareth/ISL/
[14] http://www.stat.cmu.edu/~cshalizi/ADAfaEPoV/
[15] http://adv-r.had.co.nz/
[16] https://www.openintro.org/stat/textbook.php
[17] https://leanpub.com/LittleInferenceBook