Contents

[2. The data analytic question 6](#_Toc448603299)

[2.1 Define the data analytic question first 6](#_Toc448603300)

[2.2 Descriptive 6](#_Toc448603301)

[2.3 Exploratory 6](#_Toc448603302)

[2.4 Inferential 6](#_Toc448603303)

[2.5 Predictive 7](#_Toc448603304)

[2.6 Causal 7](#_Toc448603305)

[2.7 Mechanistic 7](#_Toc448603306)

[2.8 Common mistakes 7](#_Toc448603307)

[2.8.1 Correlation does not imply causation 7](#_Toc448603308)

[2.8.2 Overfitting 7](#_Toc448603309)

[2.8.3 n of 1 analysis 7](#_Toc448603310)

[2.8.4 Data dredging 7](#_Toc448603311)

[3. Tidying the data 8](#_Toc448603312)

[3.1 The component of a data set 8](#_Toc448603313)

[3.2 Raw data 8](#_Toc448603314)

[3.3 Raw data is relative 8](#_Toc448603315)

[3.4 Tidy data 8](#_Toc448603316)

[3.5 Include a row at the top of each data table/spreadsheet that contains full row names. 9](#_Toc448603317)

[3.6 If you are sharing your data with the collaborator in Excel, the tidy data should be in one Excel file per table. 9](#_Toc448603318)

[3.7 The code book 9](#_Toc448603319)

[3.8 The instruction list or script must be explicit 9](#_Toc448603320)

[3.9 The ideal instruction list is a script 9](#_Toc448603321)

[3.10 If there is no script, be very detailed about parameters, versions, and order of software 9](#_Toc448603322)

[3.11 Common mistakes 9](#_Toc448603323)

[3.11.1 Combining multiple variables into a single column 9](#_Toc448603324)

[3.11.2 Merging unrelated data into a single file 9](#_Toc448603325)

[3.11.3 An instruction list that isn’t explicit 9](#_Toc448603326)

[4. Checking the data 9](#_Toc448603327)

[4.1 How to code variables 10](#_Toc448603328)

[4.2 In the code book you should explain why censored values are missing 10](#_Toc448603329)

[4.3 Avoid coding categorical or ordinal variables as numbers 10](#_Toc448603330)

[4.4 Always encode every piece of information about your observations using text 10](#_Toc448603331)

[4.5 Identify the missing value indicator 10](#_Toc448603332)

[4.6 Check for clear coding errors 10](#_Toc448603333)

[4.7 Check for label switching 10](#_Toc448603334)

[4.8 If you have data in multiple files, ensure that data that should be identical across files is identical 10](#_Toc448603335)

[4.9 Check the units (or lack of units) 11](#_Toc448603336)

[4.10 Common Mistakes 11](#_Toc448603337)

[4.10.1 Failing to check the data at all 11](#_Toc448603338)

[4.10.2 Encoding factors as quantitative numbers 11](#_Toc448603339)

[4.10.3 Not making sufficient plots 11](#_Toc448603340)

[4.10.4 Failing to look for outliers or missing values 11](#_Toc448603341)

[5. Exploratory Analysis 11](#_Toc448603342)

[5.1 Interactive analysis is the best way to explore data 11](#_Toc448603343)

[5.2 Plot as much of the actual data as you can 11](#_Toc448603344)

[5.3 Exploratory graphs and tables should be made quickly 11](#_Toc448603345)

[5.4 Plots are better than summaries 11](#_Toc448603346)

[5.5 For large data sets, subsample before plotting 11](#_Toc448603347)

[5.6 Use color and size to check for confounding 12](#_Toc448603348)

[5.7 For multi-panel plots of the same data type fix the axis 12](#_Toc448603349)

[5.8 For multi-panel plots match the comparison axis 12](#_Toc448603350)

[5.9 Use log transforms to “spread out” data with varying orders of magnitude 12](#_Toc448603351)

[5.10 Use log transforms for ratio measurements 12](#_Toc448603352)

[5.11 When comparing two measurements of the same thing – use Bland Altman plots 12](#_Toc448603353)

[5.12 Common Mistakes 12](#_Toc448603354)

[5.12.1 Optimizing style too quickly 12](#_Toc448603355)

[5.12.2 False pattern recognition 12](#_Toc448603356)

[5.12.3 Failure to explore and jumping to statistical tests 12](#_Toc448603357)

[5.12.4 Failure to look at patterns of missing values and the impact they might have on conclusions. 12](#_Toc448603358)

[6. Statistical modeling and inference 13](#_Toc448603359)

[6.1 When possible, perform exploratory and confirmatory analysis on separate data sets. 13](#_Toc448603360)

[6.2 Define the population, sample, individuals and data 13](#_Toc448603361)

[6.3 Identify reasons your sample may not represent the population 13](#_Toc448603362)

[6.4 Identify potential confounders 13](#_Toc448603363)

[6.5 Check the distribution of missing data 14](#_Toc448603364)

[6.6 Check for outliers 14](#_Toc448603365)

[6.7 Confirm that estimates have reasonable signs and magnitudes 14](#_Toc448603366)

[6.8 Be careful of very small or very large samples 14](#_Toc448603367)

[6.9 When performing multiple hypothesis tests, correct for multiple testing 14](#_Toc448603368)

[6.10 Smooth when you have data measured over space, distance, or time 14](#_Toc448603369)

[6.11 Know your real sample size 14](#_Toc448603370)

[6.12 Common errors 14](#_Toc448603371)

[6.12.1 Failing to account for dependencies 14](#_Toc448603372)

[6.12.2 Focusing on p-values over confidence intervals 14](#_Toc448603373)

[6.12.3 Inference without exploration 14](#_Toc448603374)

[6.12.4 Assuming the statistical model fit is good 15](#_Toc448603375)

[6.12.5 Drawing conclusions about the wrong population 15](#_Toc448603376)

[6.12.6 Not addressing uncertainty 15](#_Toc448603377)

[6.12.7 Identifying real correlations you don’t care about 15](#_Toc448603378)

[7. Prediction and machine learning 15](#_Toc448603379)

[7.1 Split the data into training and validation sets 15](#_Toc448603380)

[7.2 Identify reasons your sample may not represent the population 15](#_Toc448603381)

[7.3 More data usually beats better algorithms 15](#_Toc448603382)

[7.4 Features are more important than the algorithm 15](#_Toc448603383)

[7.5 Define your error measure first 15](#_Toc448603384)

[7.6 Avoid overfitting with cross validation 16](#_Toc448603385)

[7.7 If the goal is prediction accuracy, average many prediction models together 16](#_Toc448603386)

[7.8 Prediction is about tradeoffs 16](#_Toc448603387)

[8. Causality 16](#_Toc448603388)

[8.1 Causal data analysis of non-randomized experiments is often difficult to justify 16](#_Toc448603389)

[8.2 Even randomized studies may be difficult to interpret causally 16](#_Toc448603390)

[8.3 For randomized studies use exploratory analysis to confirm the randomization “worked” 16](#_Toc448603391)

[8.4 Causal data analyses seek to identify average effects between often noisy variables 16](#_Toc448603392)

[8.5 Unless you have performed a randomized experiment or used causal techniques avoid using causal language 17](#_Toc448603393)

[8.6 Common Mistakes 17](#_Toc448603394)

[8.6.1 Causality creep 17](#_Toc448603395)

[9. Written analyses 17](#_Toc448603396)

[9.1 The elements of a written analysis 17](#_Toc448603397)

[9.2 Lead with the question you are answering 17](#_Toc448603398)

[9.3 Describe the experimental design 17](#_Toc448603399)

[9.4 Describe the data set 17](#_Toc448603400)

[9.5 When describing a statistical model use equations or pseudocode 17](#_Toc448603401)

[9.6 Specify the uncertainty distribution 17](#_Toc448603402)

[9.7 For each parameter of interest report an estimate and interpretation on the scale of interest 18](#_Toc448603403)

[9.8 For each parameter report a measure of uncertainty on the scientific scale 18](#_Toc448603404)

[9.9 Summarize the importance of reported estimates 18](#_Toc448603405)

[9.10 Report potential problems with the analysis 18](#_Toc448603406)

[9.11 Do not report every analysis you performed 18](#_Toc448603407)

[9.12 Every statistical or machine learning method should be referenced 18](#_Toc448603408)

[10. Creating figures 18](#_Toc448603409)

[10.1 Information should be communicated as often as possible, and on common scales 18](#_Toc448603410)

[10.2 Low information density should be avoided 18](#_Toc448603411)

[10.3 Gratuitous flourishes should be avoided 18](#_Toc448603412)

[10.4 Color and size may both be used to communicate information 18](#_Toc448603413)

[10.5 When there are many values of a third variable use faceting 19](#_Toc448603414)

[10.6 Axis values should be large, easy to read, in plain language 19](#_Toc448603415)

[10.7 Include units in figure labels and legends 19](#_Toc448603416)

[10.8 Use figure legends 19](#_Toc448603417)

[10.9 Figure legends in the figure are preferred 19](#_Toc448603418)

[10.10 Figure titles should communicate the message of the plot 19](#_Toc448603419)

[10.11 Label multi-panel plots with numbers or letters 19](#_Toc448603420)

[10.12 Add text to the plot itself to communicate a message 19](#_Toc448603421)

[10.13 Figure captions should be self-contained 19](#_Toc448603422)

[10.14 Common Errors 19](#_Toc448603423)

[10.14.1 Using a color palette that colorblind people can’t see 19](#_Toc448603424)

[10.14.2 Using colors that look too similar to each other 19](#_Toc448603425)

[10.14.3 Not making scatterplots when you should 19](#_Toc448603426)

[10.14.4 Failing to take logs. 19](#_Toc448603427)

[10.14.5 Using a plot of x versus y when a plot of (x+y) versus (x-y) is more informative 19](#_Toc448603428)

[10.14.6 Failing to ensure that multiple panels have the same vertical scale 19](#_Toc448603429)

[10.14.7 Failing to consider the point of the graph but rather just using some standard form 20](#_Toc448603430)

[10.14.8 Bar plots with antennas 20](#_Toc448603431)

[10.14.9 Being inconsistent about color choice across multiple panels or figures 20](#_Toc448603432)

[10.14.10 Aligning things horizontally when comparisons would be better made vertically, or vice versa 20](#_Toc448603433)

[10.14.11 Any of the other issues in Karl Broman’s presentation on displaying data badly 20](#_Toc448603434)

[11. Presenting data 20](#_Toc448603435)

[11.1 Tailor your talk to your audience 21](#_Toc448603436)

[11.2 Order your talk in story format 21](#_Toc448603437)

[11.3 Use large fonts 21](#_Toc448603438)

[11.4 Include contact information early 22](#_Toc448603439)

[11.5 All figures should have large axis labels in plain English 22](#_Toc448603440)

[11.6 Be sure to attribute all images and text you borrow 22](#_Toc448603441)

[11.7 In general use a solid background and opposite color font 22](#_Toc448603442)

[11.8 Minimize text on slide 22](#_Toc448603443)

[11.9 Explain every figure in your talk in detail 22](#_Toc448603444)

[11.10 Use equations to make ideas concrete, but use them sparingly 22](#_Toc448603445)

[11.11 Be willing to say “I don’t know” 22](#_Toc448603446)

[11.12 Distinguish your response type when answering questions 22](#_Toc448603447)

[11.13 Never be aggressive 22](#_Toc448603448)

[11.14 Finish on time 23](#_Toc448603449)

[11.15 Where you should put your talk 23](#_Toc448603450)

[12. Reproducibility 23](#_Toc448603451)

[12.1 Have a data analysis script 23](#_Toc448603452)

[12.2 Record versions of software and parameters 23](#_Toc448603453)

[12.3 Organize your data analysis 23](#_Toc448603454)

[12.4 Use version control 24](#_Toc448603455)

[12.5 Set a seed 24](#_Toc448603456)

[12.6 For large data sets, save intermediate results and especially how you got them 24](#_Toc448603457)

[12.7 Have someone else run your analysis 24](#_Toc448603458)

[12.8 Common mistakes 24](#_Toc448603459)

[12.8.1 Not using a script for your analysis 24](#_Toc448603460)

[12.8.2 Not recording version numbers or parameters used 24](#_Toc448603461)

[12.8.3 Not sharing data or code 24](#_Toc448603462)

[12.8.4 Using reproducibility as a weapon 24](#_Toc448603463)

[13. A few matters of form 24](#_Toc448603464)

# 2. The data analytic question

## 2.1 Define the data analytic question first

“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.”

## 2.2 Descriptive

A summarization of the measurements in a single data set without further interpretation, like the US Census.

## 2.3 Exploratory

Exploratory analyses build on a descriptive analysis by searching for discoveries, trends, correlations or relationships between the measurements of multiple variables to generate ideas or hypotheses, like when amateur astronomers discovered a four-planet solar system using public data from the Kepler telescope.

Can make discoveries, but can rarely confirm them.

## 2.4 Inferential

Inferential analyses quantify whether an observed pattern will likely hold beyond the data set in hand, like a study of whether air pollution correlates with life expectancy at the state level. The goal of such an analysis is to both determine the strength of the relationship in both the specific data sets and to determine whether that relationship will hold in future data.

In non-randomized experiments, it is normally only possible to observe whether a relationship between two measurements exists. It is often impossible to determine the how or the why of this discovered relationship.

## 2.5 Predictive

Predictive analyses use a subset of measurements from an inferential analysis (the features) in order to predict another measurement (the outcome) on a single person or unit, like when FiveThirtyEight.com uses polling data to predict how people will vote on election day.

May only show that you can predict one measurement from another, and not necessarily why that choice of prediction works.

## 2.6 Causal

Causal analyses figure out what would happen to one measurement if you make another measurement change, like a randomized clinical trial in which fecal transplants are being tested to see if they do or don’t reduce the likelihood of an infection.

Causal analyses identify both the magnitude and direction of relationships between variables.

## 2.7 Mechanistic

Mechanistic analyses are similar to causal analyses however, causal analyses are for averaging (like if you smoke you are much more likely to get cancer) whereas mechanistic analyses are for deterministic situations where if one measurement is changed, the other measurement is always and exclusively changed to a specific and deterministic behavior (like with wing design changes changing air flow over a wing to decrease drag).

## 2.8 Common mistakes

### 2.8.1 Correlation does not imply causation

Interpreting an inferential analysis as causal.

Unless a randomized study is performed, it is difficult to infer why there is a relationship between two variables.

### 2.8.2 Overfitting

Interpreting an exploratory analysis as predictive.

### 2.8.3 n of 1 analysis

Descriptive versus inferential analysis.

Hard to make an exploration- let along an inference- out of a very small sample size.

### 2.8.4 Data dredging

Interpreting an exploratory analysis as inferential.

# 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 component of a data set

Converting data from its raw form into a directly analyzable form is the first step of any data analysis.

The components of a data set include:

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 to 3.

## 3.2 Raw data

It is critical to include the rawest form of the data that you have access to.

Raw data can be:

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 can tell if data is raw by determining that:

1. No software has been run on it,
2. No numbers manipulation has been performed on it,
3. No data has been removed from any data set, and,
4. No summarization has been made.

Any manipulation at all of the raw data invalidates its format as the rawest form of data.

## 3.3 Raw data is relative

Raw data will be different to each person handling it. I.e., a blood pressure machine does an internal calculation to get its measurement that you may not be privy to.

Obtain the rawest form of data as possible. That is to say, some pre-processing is usually inevitable.

## 3.4 Tidy data

Four general principles of tidy data include:

1. Each measured variable should be in its own column.
2. Each different observation of that variable should be in its own row.
3. There should be one table for each “kind” of variable.
4. If you have multiple tables, they should include a column in the table that allows them to be linked.

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

If age of diagnosis is being measured, AgeatDiagnosis is appropriate whereas ADx is not.

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

Single worksheet, no macros, no highlighting.

## 3.7 The code book

At a minimum, this 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.

## 3.8 The instruction list or script must be explicit

Make sure the rest of the world can reproduce your results from the raw data all the way to the final results.

## 3.9 The ideal instruction list is a script

Uses the raw data as an input and produces the tidy data as an output.

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

Provide pseudocode if unable to script yourself. Confirm system OS, software used, and if there were multiple runs.

## 3.11 Common mistakes

### 3.11.1 Combining multiple variables into a single column

Combining two variables (sex and age) into a single variable.

### 3.11.2 Merging unrelated data into a single file

Person’s health versus person’s finances.

### 3.11.3 An instruction list that isn’t explicit

Not reporting parameters or versions of software.

# 4. Checking the data

Data munging, or processing, will be needed for almost all data sets you have access to.

## 4.1 How to code variables

-Continuous: These variables are anything measured on a quantitative scale that could be any fractional number. An example would be something measured in kilograms.

-Ordinal: These data have a fixed and small (less than 100) number of possible values (levels) that are ordered. An example would be survey choices.

-Categorical: Same as ordinal data but without ordering. An example would be gender.

-Missing: Data that is missing and that you do not know the mechanism. Use a single column code for all missing values.

-Censored: Data that is missing and that you know the mechanism on some level. For example, a measurement being below the detection limit, or a patient being lost to follow-up. When being coded, if no data, code like missing. However, add a new column, “VariableNameCensored”, with values of, “TRUE” if censored or, “FALSE” if not.

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

REPORT this! Statistical models treating missing and censored data are completely different.

## 4.3 Avoid coding categorical or ordinal variables as numbers

For example, with gender, enter male or female.

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

For example, if Excel is being used and colored text or cell background coloring is done to indicated information about the observation, when the information is exported as raw data, ALL of that is lost. Every piece of data should be actual text that is exportable. For instance, with possibly questionable data, DO NOT highlight the data in question but rather make an additional column that denotes whether or not it is questionable.

## 4.5 Identify the missing value indicator

Use a value like, “NA” or, “.”. Avoid numeric values, as they will skew results.

## 4.6 Check for clear coding errors

Variables which take 0, 1, or 2 should not take 9. Similarly, the gender should all be one specific value per gender (i.e. Male for men or males, female for women or females).

## 4.7 Check for label switching

Same individual should not be listed as male in one table and female in another.

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

If measurements must be recorded twice, make sure they are the same.

## 4.9 Check the units (or lack of units)

Make sure units make sense- for instance, a 180 inch person was most likely measured in centimeters.

## 4.10 Common Mistakes

### 4.10.1 Failing to check the data at all

Do not fall into this temptation: ALWAYS check your data.

### 4.10.2 Encoding factors as quantitative numbers

If the scale is qualitative, do not use numbers in encoding.

### 4.10.3 Not making sufficient plots

Tabular summaries alone are not enough- a data visualization for each and every potential problem in the data set is the safest way to go.

### 4.10.4 Failing to look for outliers or missing values

Do not assume that all measurements follow the appropriate distribution.

# 5. Exploratory Analysis

Exploratory analysis often involves summarizing and visualizing data before formal modeling.

Graphs are made for data exploration in order:

* To understand properties of data.
* To inspect qualitative features rather than a huge table of data.
* To discover new patterns or associations.

## 5.1 Interactive analysis is the best way to explore data

To understand data, play around with it and explore it: make plots, make tables, identify quirks, identify outliers, identify missing data patterns and identify problems with the data.

Use random sampling when possible: “make big data as small as possible as quickly as possible”.

## 5.2 Plot as much of the actual data as you can

Plotting more data allows you to find outliers, confounders, missing data, relationships and correlations.

## 5.3 Exploratory graphs and tables should be made quickly

These are for communication with yourself: speed and accuracy are important, but polish not so much.

## 5.4 Plots are better than summaries

Because correlation and regression lines can be identical for varied plots, plots are better than summaries.

## 5.5 For large data sets, subsample before plotting

Most trends will persevere in a random subsampling.

## 5.6 Use color and size to check for confounding

With scatterplots, use color or size of points to check for a confounding relationship. I.e., in a scatterplot where the more you study corresponds to the lesser score you get on a test, the points had better be sized or colored by skill so that you can figure out the confounding piece.

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

Don’t make it more difficult to interpret by placing the data on varied axes, which will place the points on different scales.

## 5.8 For multi-panel plots match the comparison axis

Y-axis values compared horizontally, x-axis values compared vertically.

## 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. To avoid these zeroes, do a log(data + c), where c is a small, positive constant.

## 5.10 Use log transforms for ratio measurements

This will often make the distribution more symmetric.

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

For seeing if two different ways of measuring the same quantity agree: plot x+y versus x-y.

## 5.12 Common Mistakes

### 5.12.1 Optimizing style too quickly

The goal is to understand a data set, not to make it beautiful.

### 5.12.2 False pattern recognition

Do not identify and interpret a pattern without breaking it down by checking for confounders and/or alternative explanations.

### 5.12.3 Failure to explore and jumping to statistical tests

Do not look for statistical significance without an exploration- particularly common mistake when using automated software.

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

Missing data is often ignored by statistical software. For example, a relationship cannot be established between geographical location and income in a city while missing all the data from suburban locations.

# 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.

This subsample is identified with probability.

Statistical modeling and inference are used to try to generalize what we see in the population.

There are two steps to inference, in order:

1. Obtain a best estimate for what is expected in the population.

2. Quantify the uncertainty regarding the estimate we are making.

Uncertainty can be due to many things, including:

1. The fact that the population is sampled,

2. The technology used to measure the data, and,

3. Natural variation between the individuals being measured.

At least these three things should be accounted for by a good statistical model.

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

When exploring data for new relationships without having an a priori model or a hypothesis, split the data into two random subsamples, and perform the exploration on one subsample but the confirmation on the other. A typical split is 70% for discovery and 30% for confirmation.

## 6.2 Define the population, sample, individuals and data

Example: Clinical trial’s population is people on diabetes drugs; sample is a subset of the trial enrollees; individuals (the sampling units) are trial enrollees; data are the measurements on the sampling units.

## 6.3 Identify reasons your sample may not represent the population

Since it is often impossible to sample purely randomly from a population, misrepresentation will creep in: report all possible reasons for that.

## 6.4 Identify potential confounders

If measuring shoe size and literacy a correlation will be found. This is due to age- a confounder- affecting schooling and thus literacy.

Confounders are variables associated with both variables you are trying to relate.

## 6.5 Check the distribution of missing data

Determine if missing values are associated with any of the variables you have in your data set. This can cause distortions between two variables when the presence of one missing variable is correlated with the second.

## 6.6 Check for outliers

These can lead to misleadingly large relationships in summary statistics.

## 6.7 Confirm that estimates have reasonable signs and magnitudes

If a year’s increase in education seems to lead to a 1 million dollar increase in yearly salary, check your data.

## 6.8 Be careful of very small or very large samples

This is mostly an issue for estimates of uncertainty- small samples make them hard to quantify, whereas large ones make them almost obsolete.

## 6.9 When performing multiple hypothesis tests, correct for multiple testing

Classic hypothesis tests call results significant 5% of the time by design.

False discovery rate, as an error rate, shows the rate at which significant discoveries are false and are most commonly used when there are potentially many true discoveries.

The family wise error rate shows the probability of making even one false significant call among the tests you perform, and this is standardly controlled using the Bonferroni Correction approach.

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

Regression is a form of smoothing. So are smoothing spines, moving averages, and loess.

## 6.11 Know your real sample size

Do not be tricked by false size information, such as, possibly, the amount of people in a social network.

## 6.12 Common errors

### 6.12.1 Failing to account for dependencies

Data measured across time across space is likely dependent. Plot each variable across time AND across space, to detect dependencies.

### 6.12.2 Focusing on p-values over confidence intervals

p-values are not good enough alone for any convincing analysis- always include and interpret confidence intervals, credible intervals, or another measurement of inference on a scientific scale.

### 6.12.3 Inference without exploration

Always tidy, check and explore data first. This helps with identification of dataset specific conditions that may violate your model assumptions.

### 6.12.4 Assuming the statistical model fit is good

Once a model is fit to data, ALWAYS evaluate how well that model describes the data.

### 6.12.5 Drawing conclusions about the wrong population

Inference to the wrong population or having your population change WILL bias your results.

### 6.12.6 Not addressing uncertainty

Not reporting uncertainty measures equates to claiming you know the exact value of that parameter. Saying the average height of a US citizen is 5’8” without stating a measure of uncertainty claims that you measure the height of all US citizens and know their exact values.

### 6.12.7 Identifying real correlations you don’t care about

Ice cream sales plotted against murder rates will find a (real) correlation (due to higher murder rates in higher temperature areas, where ice cream is often sold). This correlation is real, but not cared about.

# 7. Prediction and machine learning

The central idea with prediction is to take a sample from a population and create a training set.

Some training set variables are called features and the others outcomes.

The goal of prediction is to build an algorithm that automatically takes features from a new sampling unit and spits out their estimated outcome.

## 7.1 Split the data into training and validation sets

70% training set which, once finalized, leads to a 30% validation set to which the model is applied once and only once to estimate the real error rate of your algorithm.

## 7.2 Identify reasons your sample may not represent the population

If the sample used to create the training set differs from the population, predictive functions WILL fail.

## 7.3 More data usually beats better algorithms

Collecting more appropriate data for a prediction algorithm will normally improve it more than optimization.

## 7.4 Features are more important than the algorithm

Because many out of the box prediction algorithms perform very similarly on most data sets, the best way to improve accuracy is often to pick better data and variables.

## 7.5 Define your error measure first

Always define your error measure BEFORE starting to model.

Accuracy, sensitivity and specificity often used for binary outcomes.

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

## 7.6 Avoid overfitting with cross validation

Do not tune your model too closely to the observed data. Use cross-validation, or the building of two random training sets- one to build models with and one to pick a model with.

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

In general, the prediction algorithms the most frequently win prediction competitions blend multiple models together by averaging.

## 7.8 Prediction is about tradeoffs

Interpretability versus accuracy

Speed versus accuracy

Simplicity versus accuracy

Scalability versus accuracy

# 8. Causality

Gold standard here is combining specific experimental designs (e.g. randomized studies) with standard statistical analysis techniques.

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

Non-randomized studies frequently produce false, disprovable analyses.

## 8.2 Even randomized studies may be difficult to interpret causally

Common difficulties include:

1. Patient dropout,

2. Unblinding of trials,

3. Treatments that are difficult to take or adhere to, so just the intent to treat the person must be used as the treatment itself.

## 8.3 For randomized studies use exploratory analysis to confirm the randomization “worked”

Make tables, plots, and include missed variables while ensuring same distribution amongst groups.

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

Smoking increases likelihood of developing cancer, but smokers do not always get cancer.

## 8.5 Unless you have performed a randomized experiment or used causal techniques avoid using causal language

Causality creep is using words like, “cause”, “effect” and “leads to an increase” when performing an inferential or predictive analysis.

## 8.6 Common Mistakes

### 8.6.1 Causality creep

Do not look for correlations or associations between measurements in a non-randomized study.

# 9. Written analyses

Data analysis requires communication as much as statistics.

Written reports must tell a clear, precise, and compelling story that laymen could read.

Focus on how text, figures, equations and code add to or detract from that story.

## 9.1 The elements of a written analysis

Always include:

* A title
* An introduction/motivation
* A description of the statistics or machine learning models used
* Results WITH measures of uncertainty
* Conclusions WITH potential problems
* References

## 9.2 Lead with the question you are answering

Before explanations, lead with a scientific or business application question which you are answering.

## 9.3 Describe the experimental design

Explain data origins, collection, and technologies and systems used in collection.

## 9.4 Describe the data set

Explain processing and produced tidy data.

## 9.5 When describing a statistical model use equations or pseudocode

Every model MUST be mathematically specified, whether in-text (for statistical audiences) or in an appendix (for laymen).

## 9.6 Specify the uncertainty distribution

Declare the distribution of errors.

What assumptions about errors are made?

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

For weight, report an estimated change of 3 pounds in weight per 1 inch in height, not a=3.

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

Use standard deviations, confidence intervals or credible intervals as your measurements typically.

## 9.9 Summarize the importance of reported estimates

Report why the estimate was calculated and what the quantitative value means.

## 9.10 Report potential problems with the analysis

Report outliers, missing data, confounders, etc.

## 9.11 Do not report every analysis you performed

Especially not exploratory ones.

Ask it if contributes to the story or tells a crucial fact that cannot be left out.

## 9.12 Every statistical or machine learning method should be referenced

Give credit to model developer.

# 10. Creating figures

Goal is that when viewed, and with an appropriately detailed caption, they can stand alone without further explanation as a unit of information.

## 10.1 Information should be communicated as often as possible, and on common scales

Pie charts are bad- angles are notoriously hard to compare.

Bar charts are much better.

Humans are best at perceiving position along a single axis with a common scale.

## 10.2 Low information density should be avoided

Plots with a single point are bad- just report this as a data value.

Similarly, be wary of graphs primarily composed of whitespace.

## 10.3 Gratuitous flourishes should be avoided

Please the eye, but don’t make eye candy

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

Both simultaneously, or one at a time.

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

Facet the plot by making multiple panels, one per value of the third variable, in order to view correctly.

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

Not variable names from the code book.

## 10.7 Include units in figure labels and legends

ALWAYS include units on figure labels.

## 10.8 Use figure legends

For size and/or color

## 10.9 Figure legends in the figure are preferred

When possible, put legends in the plot, without a separate figure.

## 10.10 Figure titles should communicate the message of the plot

Communicate the main take home message, don’t just describe the data.

## 10.11 Label multi-panel plots with numbers or letters

Do this natively so that each plot can be referred to individually.

## 10.12 Add text to the plot itself to communicate a message

Feel free to label trends or features.

## 10.13 Figure captions should be self-contained

They should explain the whole figure: x- and y-axes, units, colors, sizes, expected observations.

## 10.14 Common Errors

### 10.14.1 Using a color palette that colorblind people can’t see

No red or green

### 10.14.2 Using colors that look too similar to each other

Red and pink, blue and light blue, etc.

### 10.14.3 Not making scatterplots when you should

Use scatterplots- the most informative- in favor of box- or bar-plots.

### 10.14.4 Failing to take logs.

Use log transforms if they are called for, such as when data is highly skewed.

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

Use Bland-Altman plots to see if two or more ways of measuring the same quantity agree.

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

Make sure ALL axes are on the same scale.

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

Do not automatically use the customary method.

### 10.14.8 Bar plots with antennas

Dynamite plots have low information density.

### 10.14.9 Being inconsistent about color choice across multiple panels or figures

Same variable, same color choice.

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

Line panels up via the axis used in comparison.

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

<https://www.biostat.wisc.edu/~kbroman/presentations/IowaState2013/graphs_combined.pdf>

Excel obfuscation.

To display data badly:

* Display as little information as possible
* Obscure what you do show (with chart junk)
* Use pseudo-3d and color gratuitously
* Make a pie chart (preferably in color and 3d)
* Use a poorly chosen scale
* Ignore sig figs

Displaying data well:

* Be accurate and clear
* Let the data speak
  + Show as much information as possible while not obscuring the message
* Science not sales
  + Avoid unnecessary frills
* In tables, ALL digits meaningful (don’t drop ending 0’s)

# 11. Presenting data

Giving data science talks helps one to:

* Meet people
* Get excited about your ideas/software/results
* Help others understand what you’re excited about

Entertain, but remember, this is NOT a TED talk.

## 11.1 Tailor your talk to your audience

Small group meeting)

Goal: Update people

Talk about: Short intro, brief update, long discussion on where you’re going and help needed.

Short talk at a conference)

Goal: Entertain people, get people to read your paper/use your software

Talk about: Short intro, brief solution, links

Long format formal talk)

Goal: Entertain people, get people to read your paper/use your software, make them understand your problem AND solution

Talk about: Intro, solution, results, connection to broader ideas.

Job talk)

Goal: Get a job, entertain people, make them understand your problem AND solution

Talk about: You, intro to ONE problem, how solved, results, summary of actions taken and planned.

## 11.2 Order your talk in story format

Always lead with a brief, layman’s terms explanation of the scientific problem that lead to your data.

Explain data and measurement technology.

Explain features used to model data.

Explain the model.

Present results by telling a story.

## 11.3 Use large fonts

Go HUGE.

## 11.4 Include contact information early

On the title slide give a way to contact you.

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

Readable from back of room.

## 11.6 Be sure to attribute all images and text you borrow

Links and courtesy of statements.

## 11.7 In general use a solid background and opposite color font

No pictures or similar color fonts to background color.

## 11.8 Minimize text on slide

1 bullet point per slide MAX- textless is better.

## 11.9 Explain every figure in your talk in detail

What is it supposed to communicate? Axes? Audience should look for what?

If time prevents you from fully explaining a figure, omit it.

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

-Explain data first

-Use words, not symbols

-No more than two subscripts if possible

-When explaining:

-First explain equation’s point.

-Then explain each symbol.

-Finally, explain how the model relates them.

## 11.11 Be willing to say “I don’t know”

Say those words if you have to.

## 11.12 Distinguish your response type when answering questions

Did you look into the question and know an answer, or are you giving a complete guess? Or are you somewhere in between?

## 11.13 Never be aggressive

Even if an audience member is aggressive, NEVER fall into the temptation to respond aggressively.

## 11.14 Finish on time

Big crowd winner.

## 11.15 Where you should put your talk

<http://speakerdeck.com>

<http://www.slideshare.net>

Latter is better for mobile viewing.

# 12. Reproducibility

Reproducibility is the quality of being able to recalculate the exact numbers calculated in a data analysis when given the code and raw data by another data analyst.

## 12.1 Have a data analysis script

R Markdown or iPython notebooks

## 12.2 Record versions of software and parameters

Record versions of software at a minimum.

Get the devtools::session\_info() in R.

## 12.3 Organize your data analysis

Have SEPARATE folders at EACH level 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 ALL components

-Final data analysis products (presentations and/or write-ups)

## 12.4 Use version control

Like popular Github.

## 12.5 Set a seed

Use a seed for random number generation.

In R- set.seed(x) where x is any positive integer.

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

Create a script- with version numbers and parameters included- to calculate summary measures and save intermediate data.

## 12.7 Have someone else run your analysis

Make sure they get your same results.

## 12.8 Common mistakes

### 12.8.1 Not using a script for your analysis

Written documentation introduces much more error proneness.

### 12.8.2 Not recording version numbers or parameters used

Record:

1. Type of computer

2. Versions of ALL used software

3. ALL parameters used to perform analysis

### 12.8.3 Not sharing data or code

Always link these.

### 12.8.4 Using reproducibility as a weapon

Do not humiliate people publicly when finding issues with their work- contact them to help!!!

# 13. A few matters of form

-Estimates are followed by parentheses.

-P-value reporting should not include numbers below machine precision.

-P-value permutation reporting should avoid reported zeroes.

-Do not report estimates with over precision.

-When programming- variable names are lower case, words separated by underscores and as explicit as possible in the data frames being analyzed.

-In written analysis, report variable names in plain language, not as variable names.