DSCI354-451 Foundation: Question, Tidy, Check, Explore(CWRU, Pitt, UCF, UTRGV)

Profs: R. H. French, L. S. Bruckman, P. Leu, K. Davis, S. Cirlos

TAs: W. Oltjen, K. Hernandez, M. Li, M. Li, D. Colvin

15 September, 2022

Contents

3.2.3.1 First: Guess The Correlation Game

- Guess The Correlation Game

3.2.3.2 Tidyverse thinking

3.2.3.2.1 Hadley Wickham Talk on Tidyverse

- Watch this one for class
  - Hadley Wickham on Tidyverse
  - Hadley is from New Zealand
    - And is chief scientist at RStudio company

Alternatively here is a more extended version

- RConf-Jan 2017: Data Science in the Tidyverse And slides

3.2.3.2.2 Hadley has a great book on making R packages

- Its available to read online
  - R Packages

3.2.3.2.3 Purrr is another part of the Tidyverse

- purrr enhances R's functional programming (FP) toolkit
  - by providing a complete and consistent set of tools
  - for working with functions and vectors.

If you've never heard of FP before,

- Check Wikipedia first
  - Functional Programming
- the best place to start is
- the family of map() functions
  - which allow you to replace many for loops
  - with code that is both more succinct and easier to read.

The best place to learn about the map() functions

- is the iteration chapter in R for data science.
- Use quick formula functions in purrr::map (+ base vs tidtyverse idiom comparisons/examples)

purrr tutorial

3.2.3.3 An Example

- Chart: It's not your imagination, US gun violence is over the top this summer

  Falling into the Pit of Success

3.2.3.3.1 Jenny Bryan on Data Wrangling

- Jenny Bryan on Data Wrangling

  Basic care and feeding of data in R

3.2.3.4 Some Tidy Data Analysis Resources

3.2.3.4.1 Elements of Data Analytic Style; Jeff Leak

- https://leanpub.com/datastyle

3.2.3.4.2 R for Data Science

- By Garrett Grolemund, Hadley Wickham
- http://r4ds.had.co.nz/

3.2.3.4.3 What is Tidy Data

- A Wickham paper in your readings subdirectory
- http://vita.had.co.nz/papers/tidy-data.pdf

3.2.3.5 Data Analysis Question, Tidying, Checking, Exploratory Data Analysis

3.2.3.5.1 Answering the question

1. Did you specify the type of data analytic question

```
- (e.g. exploration, association, causality)
  - exploration
  - association
  - causality
  - inferential (mechanistic)
- before touching the data?
```

So here is another John Tukey quote.

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

3.2.3.5.2 The Data Analysis Flow Chart

1. Types of Data Analyses

- Descriptive:
  - A descriptive data analysis seeks to summarize the measurements
    - in a single data set without further interpretation.
- 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.
- 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.
- 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.
- Causal:
  - A causal data analysis seeks to find out what happens to one measurement
    - if you make another measurement change.

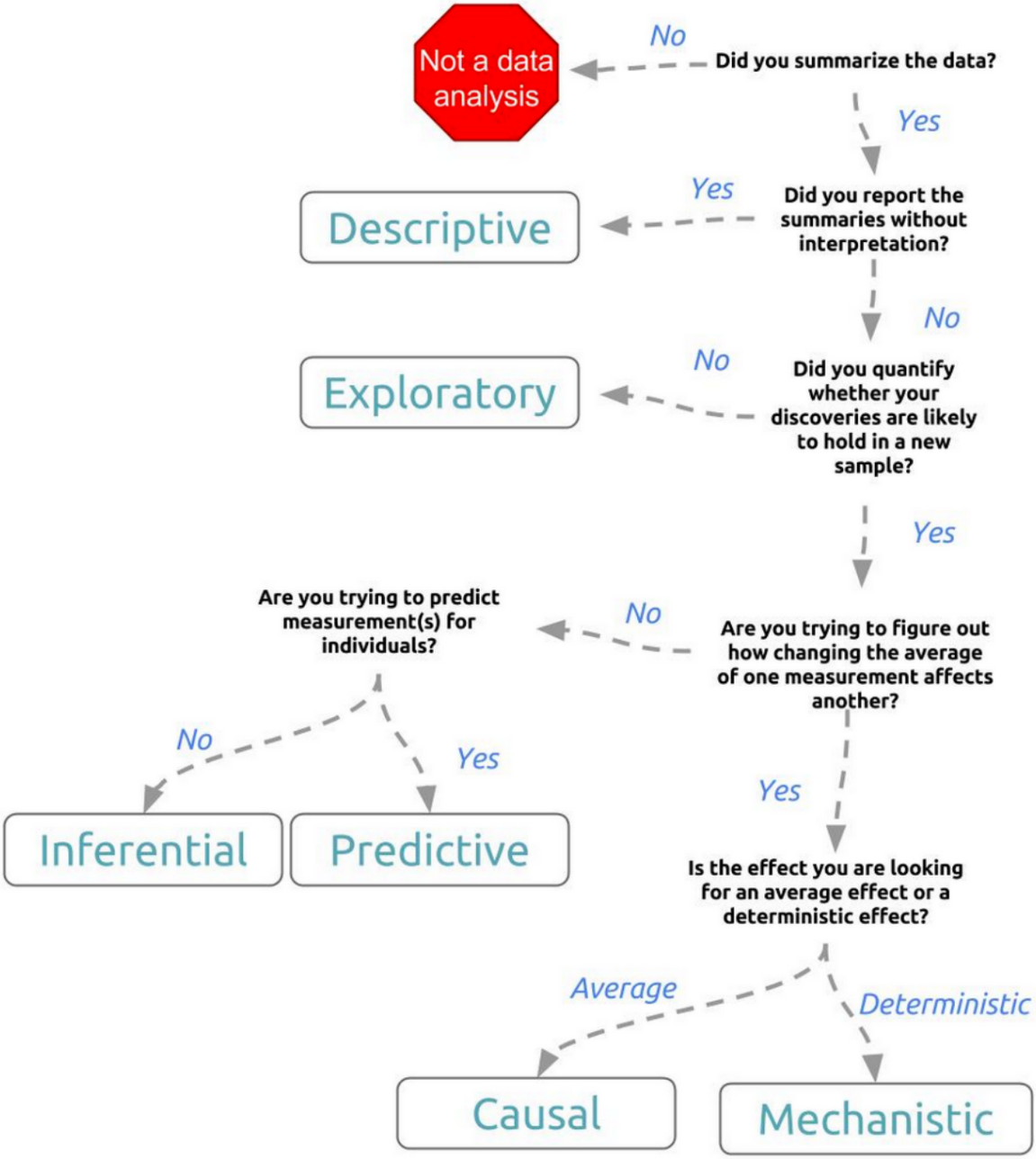

Figure 2.1 The data analysis question type flow chart

Figure 1: the da question flow chart

- 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.

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?

3.2.3.5.3 Common Mistakes

- 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 from The data analytic question
- if there is a relationship between two variables.

A great website to hunt for spurious correlations is

- http://tylervigen.com

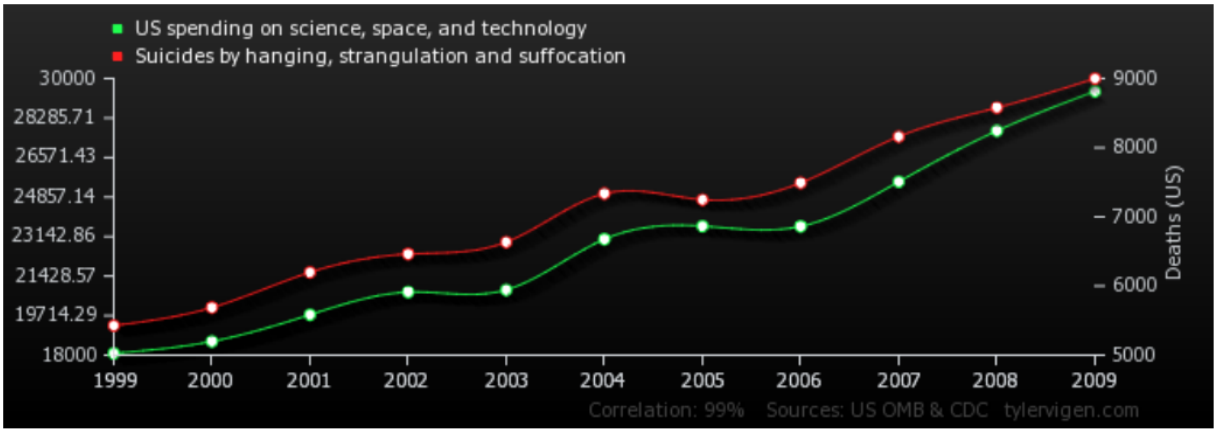

Figure 2.2 A spurious correlation

Figure 2: spurious correlations

Particular caution should be used when applying words

- such as “cause” and “effect” when performing inferential analysis.

Inference is not about causation; its infering relationships between variables.

- 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 predictive 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 (Not Good) 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.
- but this is useless

For a predictive model;

- you need to split your data into training and test datasets,
- and evaluate how well it predicts the test dataset.
- 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.
- 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.

As Ronald Coase said:

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

3.2.3.6 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.
  - 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 data 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.

3.2.3.6.1 These are the components of a processed data set:

- The raw data.
- A tidy data set.
- A code(data) book describing each variable
  - and its values in the tidy data set.
- An explicit and exact recipe you used
  - to go from raw to tidy and a databook.

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(data) book?
6. Did you record all parameters, units, and functions applied to the data?

3.2.3.6.2 Raw data: It is critical that you include

- the rawest form of the data that you have access to.

Raw data is relative:

- The raw data will be different
  - to each person that handles the data.
- One person's raw data,
  - may be some previous persons tidy data!

3.2.3.6.3 Tidy data:

- The general principles of tidy data are laid out by Hadley Wickham in
  - this paper http://vita.had.co.nz/papers/tidy-data.pdf
  - and this video https://vimeo.com/33727555.

The paper and the video are both focused on the tidyverse R packages.

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 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

3.2.3.6.4 Include a row at the top of each data table/spreadsheet

- that contains full row names.
- If you are sharing your data with the collaborator

3.2.3.6.5 They should be shared as csv files, not in Excel

- Since Excel files can have buried macros,
  - you lost control of the data analysis process.

Also one csv table per file, no workbooks

- No highligting cells.
- csv files are for data only; no code.

- or one Excel file per table.
  - but xls or xlsx files are binary and fragile
  - ascii csv files are more robust

3.2.3.7 The code (or data) book:

- 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

3.2.3.7.1 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.

3.2.3.7.2 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.

3.2.3.7.3 If there is no script,

- be very detailed about
  - parameters,
  - versions, and
  - order of software

3.2.3.7.4 Common Mistakes

- Combining multiple variables into a single column
- Merging unrelated data into a single file
- An instruction list that isn't explicit

3.2.3.8 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?

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 http://Data.gov
- 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 ave 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.

3.2.3.8.1 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. In the code book you should explain why censored values are missing.
2. Avoid coding categorical or ordinal variables as numbers.
3. Always encode every piece of information about your observations using text.
4. Identify the missing value indicator

There are a number of different ways

- that missing values can be encoded in data sets.
- The common choices are "NA". Don't use numbers.

5. Check for clear coding errors
6. Check for label switching
7. 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.

You should check that for each patient

- in the two files the sex is recorded the same.
- This is part of data validation.

8. Check the units (or lack of units)

Define the units.

3.2.3.8.2 Common Mistakes

- Failing to check the data at all
- Encoding factors as quantitative numbers
- Not making sufficient plots
- Failing to look for outliers or missing values