STA130 Fall 2019 – T0107

Week 3: Data Wrangling

(Materials used in this presentation are provided by the U of T Statistical Sciences Department.

This presentation was prepared by Vivian Ngo.)

Github.com/vivianngo97/STA130-Fall-2019

viv.ngo@mail.utoronto.ca

Agenda

- Debrief of self-reflection
- Vocabulary
- R Live Coding
- Group Discussion
- Writing Examples
- Writing Activity

Debrief of Self-Reflection

- “Why so much group work?”
- “More demonstrations in R”
- “Writing activity is stressful”

Vocabulary

- **General terms:**
  - Proportion
  - Outlier
- **Terms for describing data wrangling:**
  - Cleaning data
  - Tidy data
  - Removing a column
  - Replacing values above/below a certain threshold
  - Taking the subset of variables
  - Filtering the data frame based on a condition (e.g. based on one of the variables/columns)
  - Sorting data based on a variable
  - Renaming the variables
  - Grouping categories
  - Defining new variables
  - Producing new data frames
  - Handling missing values (NAs)
  - Creating summary tables

R Live Coding

Group Discussion

- **Question 1:** (Based on Practice Problem 1.iv) Comment on the strengths and weaknesses of each of the visualizations and summary table you constructed in parts (i), (ii), and (iii).
- **Question 2:** Anything odd or surprising they observed; e.g. certain groups of people who were more likely to be on the Titanic or survive?

- Remember to discuss the big picture, not just what you did for homework.

Writing Examples

- **An Assessment of Oral Health on the Pine Ridge Indian Reservation. (Gallegos, JR et al.)** *Modified for this course*

We assessed the oral health of a group of local Indigenous people living in the Pine Ridge Indian Reservation.	Purpose
Based on a sample of 292 adults and children, screening personnel counted the number of decayed teeth, total teeth, and dental cavities (both filled and unfilled).	Methods. Very simple because not statistics-based. You should include more detail here.
On average, each individual had 4 decayed teeth. Half of adults had 27 or fewer teeth and 26% had an unfilled cavity. Further, 75% of children (<5 years of age) had an unfilled cavity.	Results. Notice how only things critical to their purpose and methods are included. Very concise!
Amongst the people of Pine Ridge, our study documented a high prevalence of cavities, numerous people with missing teeth, and many unmet dental needs, particularly among children. Future studies of oral health related behaviors, and access to oral health care are needed to explain the dental, periodontal, and soft tissue problems that adversely affect the people of the Pine Ridge Indian Reservation.	Conclusion (and recommendation)

Writing Examples

- Participants were 477 male, first year students at a liberal arts college. In the week before the start of classes, participants were given two surveys: one of expected college engagement, and the second of video game usage, including a measure of video game addiction. Results suggested that video game addiction is (a) negatively correlated with expected college engagement, (b) negatively correlated with college grade point average (GPA), even when controlling for high school GPA, and (c) negatively correlated with drug and alcohol violations that occurred during the first year in college.

Writing Examples

- More examples with explanations for what makes them good/ poor can be found here:
- https://www.kibin.com/essay-writing-blog/10-good-abstract-examples/

Writing Activity

- Imagine you work as a business analyst for a new cruise company, Fun Cruises. Your CEO has asked you to deliver a written summary of your most interesting research findings on the Titanic (practice problem 1). They want this by the end of the day because they're meeting a new client tonight for dinner and think this research will be of interest to them. It's already 3:30 pm and it's a Friday!
- Remember, your CEO is a busy person. They only want the most important information and they don't want to read more than half a page of text. Use visuals to help get key points across, if you can. The CEO has only limited statistical background, so make sure everything is clear and makes sense! Remember to start with the purpose - your CEO is busy and they may have forgotten exactly what your research was about! Also make sure to include a complete, but concise, summary of the methods, key results, and a conclusion. Remember, you're the statistics expert and the CEO is counting on *you* to summary this research!