Tutorial 1

STAT5002 - Introduction to statistics

March 7, 2018

Summary

This week we discussed the ways that data can be collected and potential sources of bias. We also examined various approaches for exploring and summarising data numerically and graphically and their limitations. Remember, when choosing a graphical or numerical summary think...

- Why am I using this summary?
- What properties of the data will this summary highlight?
- Is this summary appropriate for communicating what I want to communicate?

When using a summary think...

- What is this summary showing me?
- What is this summary not showing me?

And remember,

- Always think critically about how your data was collected, when exploring raw or summarised data, and, when someone offers an interpretation of data.

Question 1: Simpson's Paradox

Three groups of students were asked to sit two tests. The (standardised) test scores were recorded as `x` and `y` in the data file `Simpsons.csv`. The group information was recorded as `group`, which is an indication of the high school years completed by the students.

(a) Read the data `Simpsons.csv` into `R` using the function `read.csv`. What does `header = TRUE` do? (Hint: change it to `FALSE` to see what happens.) Also run the `summary()` function on the data.

(b) Extract `x = dat$x` and `y = dat$y`, and then create a scatter plot of these two variables. What can you conclude from this data? Is the trend between `x` and `y` positive or negative? In other words, if a student scored well in the first test (`x`), is the student more likely to achieve a good score or a poor score in the second test (`y`)? (If you know about correlation, then you can also try to run `cor(x,y)`).

(c) Calculate the mean, median, standard deviation, and IQR of `x` and `y`. (Hint: use in-built functions for these.)

(d) Create histograms and boxplots for `x` and `y` separately. Are there any outliers that could potentially inflate/deflate the values of above statistics?

(e) Check the documentations on the `median` function by typing `?median` into `R`. Then, calculate the median of `x` and `y` using `type = 1` and `type = 7`. Which one is "more" valid?

(f) Create `group = dat$group` and then make another scatter plot for `x` and `y` again. But this time, use `plot(x, y, col=group)`. What is the trend between `x` and `y` for each `group` now? Should your conclusion in (b) be altered?

(g) Repeat part (c), but now, calculate these statistics within each group. You can either use `mean(x[group == 1])` etc. Alternatively, you could look up `?tapply` and perform these calculations faster. What can you conclude from these statistics when split by groups?

(h) Create a boxplot for `x` and `y`, but now, try to split the boxplot by the `group` variable.

(i) What did you learn from this exercise? Why was the conclusion reversed when the extra `group` variable was taken into account? Is pooling data from different sources always a good idea? How should we analyse data when a confounding variable is present?

(j) Reading exercise: read the Wikipedia article on "Simpson's paradox".

Question 2

A government agency wanted to know on average, how many glasses of water does a person drink a day. To do this they randomly called the home phones of 20 people. As part of their survey they asked if people fell into one of 3 age groups that represent the lower (L), middle (M) and upper (U) third of the population.

Glasses	8	7	3	4	3	3	2	4	5	8	9	2	4	5	6	3	2	4	9	100
Age	*L*	*M*	*U*	*U*	*U*	*M*	*U*	*M*	*M*	*L*	*L*	*M*	*M*	*U*	*U*	*U*	*U*	*U*	*M*	*L*

a. Characterise and explore the data using numerical and graphical summaries. What do you notice?
b. Calculate the mean and median number of glasses of water that were drunk by the survey participants.
c. Calculate the mean and median number of glasses of water that were drunk by survey participants in each age group.
   - Which of these are more appropriate estimates of the usual number of glasses drunk by a person from a particular age group in the general population?
d. Calculate an appropriate estimate of the number of glasses of water drunk by a person from the general population.
   - Why is this an appropriate estimate?
   - Use one or two plots to help communicate your answer.

Question 2

Download one or more datasets from http://www.maths.usyd.edu.au/u/UG/JM/StatsData.html or elsewhere.

a. Use numerical and graphical summaries to explore the datasets.
   - How many variables in the dataset?
   - How many observations?
   - What types of variables are measured?
   - Do any variables have interesting behaviours?
     - Outliers?
     - Skewed?
     - Multimodal?
     - Missing values?
   - Is there any possible interpretation for what you find?
b. Try the following graphical summaries on your dataset. What are the advantages and disadvantages of the different plots?

```
# If needed explore the documentation for each function i.e.
?stripchart
# or
help('stripchart')

install.packages('vioplot') # Install the vioplot library
library('vioplot') # Load the vioplot library

plot(table(x))
barplot(table(x))
hist(x)
```

```
plot(density(x))
boxplot(x)
stripchart(x)
vioplot(x)
```

c. Many of the plotting functions allow you to put two plots overlayed in the one window. Can you combine these with any of the plots above? Do these change any of your previous interpretations or your assessment of advantages and disadvantages?

```
points(density(x))
stripchart(x,add = TRUE, vertical = TRUE)
```

d. Many functions have parameters that alter their behaviour. Do these alter your interpretations?

```
stripchart(list(x,x), vertical = TRUE, method = 'jitter',jitter = 1)
points(density(x),type = 'line')
hist(x, breaks = 100)
```

e. Can you use any of these functions to compare two variables? Does this alter your findings?