STA490 T0209

Week 2

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

This presentation was prepared by Vivian Ngo.)

Reminders

- Tutorials start 10 minutes after the hour
- Mentors will be coming in the 2nd half of next week's tutorial

Agenda

- Homework Question 1
- Group Discussions
- Writing Example
- Writing Activity

Vocabulary for this week's material

- Data frame
- Matrix
- Vector
- Average
- Standard deviation
- Variance
- Missing data
- Types of variables; e.g. character, numeric

Homework Question 1

The `Galton` data set in the `mosaic` library contains data from Francis Galton in the 1880s.

```
## Observations: 898
## Variables: 6
## $ family <fct> 1, 1, 1, 1, 2, 2, 2, 2, 3, 3, 4, 4, 4, 4, 4, 5, 5, 5, 5...
## $ father <dbl> 78.5, 78.5, 78.5, 78.5, 75.5, 75.5, 75.5, 75.5, 75.0, 7...
## $ mother <dbl> 67.0, 67.0, 67.0, 67.0, 66.5, 66.5, 66.5, 66.5, 64.0, 6...
## $ sex    <fct> M, F, F, F, M, M, F, F, M, F, M, M, F, F, F, M, M, M, F...
## $ height <dbl> 73.2, 69.2, 69.0, 69.0, 73.5, 72.5, 65.5, 65.5, 71.0, 6...
## $ nkids  <int> 4, 4, 4, 4, 4, 4, 4, 4, 2, 2, 5, 5, 5, 5, 5, 6, 6, 6, 6...
```

a. Using the `Galton` data use R to calculate the average and variance of child's heights in the first three families. Which family has the largest variance? Explain the meaning of variance in this context. Hint: You may find it helpful to use the `filter` function (from the `dplyr` package, which is included in `tidyverse`). The `filter` function takes a data frame as an argument, and returns a data frame with only rows satisfying one or more conditions. For example, in the code below, fam1 contains only data for individuals with `family==1` from the `Galton` dataset.

Homework Question 1

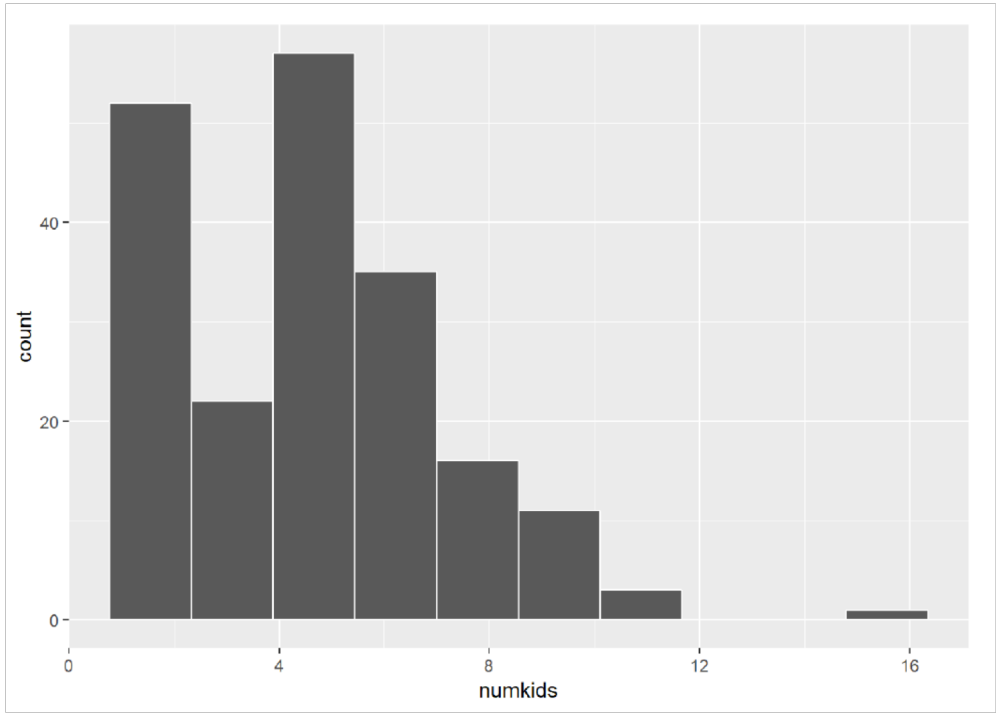
b. How many children did parents in the `Galton` data set have? Create a data frame called `data` that contains the family id number and the numbers of kids in each family. Note that the number of children is repeated for every member of the families. The data frame you create should not include the repeats. Hint: You may find it useful to use `group_by()` in combination with `summarise()` function in the `dplyr` library. We haven't covered `group_by()` in class yet, but an example on how to do this is given below. Note that both `group_by()` and `summarise()` return a data frame. Here is an example. Consider a simple data frame `marks` of the final marks for two (fictitious) students that each took five courses during their first year at UofT. The example below uses `group_by()` then `summarise()` to calculate the average mark for each student.

```
library(tidyverse)

marks <- data_frame(student = c(1, 1, 1, 1, 1, 2, 2, 2, 2, 2),
courses = c("STA130", "MAT137", "ECO100", "CSC148", "PHL100",
"STA130", "MAT137", "ECO100", "CSC148", "PHL100"),
grade = c(82, 83, 77, 84, 79, 83, 74, 85, 77, 72))

marks_grouped <- group_by(marks, student)

ave_grades <- summarise(marks_grouped, ave = mean(grade))
ave_grades
```

```
## # A tibble: 2 x 2
##   student   ave
##     <dbl> <dbl>
## 1       1  81
## 2       2  78.2
```

Homework Question 1

c. Graph the distribution of the number of kids in the `Galton` data set families. Describe (in words) the features of this distribution.



d. Just based on the graph you generated in part (c), how do you think the mean and median would compare? Justify your reasoning.

e. Compute the mean and median of the number of kids in the `Galton` data set families. Does this match what you expected to see in part (e)?

Group Discussions

- Discuss Question 3
- E.g. :
  - Describe what you did to create the variables.
  - Why did you decide to do it this way?
  - Compare graphs and summary statistics
  - What were your findings?

Writing example

- Important:
  - Introduce the variables you want to work with.
  - Define the problem you want to solve.
  - How does data wrangling fit into the problem (for example, explain how it can be solved with the newly created variable but not the original variable)?
  - Summarize the results.

Writing example

- Example 1: Because we wanted to do calculations based on family, it would be more convenient to put it in a tidy form. We calculated the number of children in each family.

Writing Example

- Example 2: For this question we used the Galton dataset, which provides data on children from in the 1880s. Because we were interested in investigating differences across families, we needed to create new variables which summed children's characteristics by family ID. For example, we were interested in determining the number of children in each family. However, this value was repeated for every member of the families included in the Galton dataset. To make it easier to generate summaries, we made a new tidy data frame that included only one row per family. Using this new data set, we were more easily able to determine the number of children in each family group. The number of kids in the Galton data set families follow a positively skewed (or equivalently, right skewed) distribution. There were many more "smaller" families than "larger"" families. The number of kids per family ranged from 1 to around 15. The distribution appears to be bimodal, with quite a few families having 1-2 children, and 4-5 children. The family with around 15 kids appears to be an outlier because it is so much higher than the number of children in the other families.

Writing Activity

- (To be submitted through Quercus by the end of tutorial)
- Explain what you have discussed for question 3.
- Some thoughts:
  - Were respondents familiar with reproducibility concerns in science? Explain.
  - Were younger respondents more or less likely to report thinking that there is a reproducibility crisis in science? Why or why not?
  - Is there variability in research reproducibility across scientific disciplines? If so, which disciplines are thought to be the most reproducible? The least?

Writing Activity Rubric

	4 (Excellent)	3 (Good)	2 (Adequate)	1 (Poor)
Context	The context and connection to the problem are clear.	Some context was provided and all variables/concepts were mentioned. Some aspects were not clear.	Very little context was provided and only some variables/ concepts were mentioned.	No context and mentioning of any variables/ concepts covering in this week's materials.
Structure	Well organized, follows a logical structure.	The organization follows some logical structure.	Some structure but difficult to follow.	There is no structure, very difficult to follow.
Conclusion	There is a clear central idea and the conclusion is correct.	A central idea or conclusion is present. The conclusion might be incorrect.	The central idea or conclusion is weak and not supported.	The central idea or conclusion is missing. Incorrect conclusion.
Transitions	The progression is logical. Effective use of transitions.	The progression is controlled. The use of transitions is mostly meaningful.	Minor disruptions in flow and weak transitions.	Weak progression and lack of transitions.
Vocabulary	Good use of statistical terms and appropriate choice of words.	Use of statistical terms and phrases mostly correct, demonstrates understanding of concepts.	Some use of statistical terms/ phrases and some understanding of concepts demonstrated.	Inaccurate or incorrect use of statistical terms or phrases and a lack of understanding statistical concepts.