STA130 Week 2 Tutorial

Kaixuan (Bryan) Hu

Mentorship program

- Second half of the tutorial on next week.
- Your mentors will come here to discuss about this.
- 3% of your final mark.

STA130 MENTORSHIP PROGRAM

This program familiarizes you with key areas of university life. You will complete four requirements which count towards 3% of your grade in the course; career development, social development, personal development and 1:1 meeting with your assigned mentor.

CAREER DEVELOPMENT

Explore where your degree can take you and expand your professional skills.

Worth 1% of your grade.

SOCIAL DEVELOPMENT

Get to know your classmates outside of class and discover what UofT and the city have to offer.

Worth 1% of your grade.

PERSONAL DEVELOPMENT

Discover and expand on your interests beyond the classroom.

Worth 0.5% of your grade.

1:1 MEETING

Benefit from a senior student's experience through one-to-one time with your assigned mentor

Worth 0.5% of your grade.

KEY PROGRAM DEADLINES

- Monday, January 21st 2019 - Program Coordinator (Megan Whitehead-Douglas) will visit your classroom to provide more information about the mentorship program
- Friday, January 25th 2019 - Meet your mentor during your tutorial
- Friday, March 29th 2019 - Last day to complete all four requirements of the program and submit supporting documents (signed career/ personal tracking form) to your mentor.

For concerns or questions about the program, email Megan.Whitehead@utoronto.ca

Measures of central tendency: Mean and Median

- Mean: common way to measure the center of a distribution of a random variable.
- How do we compute mean given overserved value of a random variable: $x_1, x_2, \ldots, x_n$?
- You just take the average of them: $\overline{x} = \frac{x_1 + \cdots + x_n}{n}$
- Median:
- If you order your data in a nondecreasing way : $x_1 < \cdots < x_k < \cdots x_n$ then median is x_k if n is odd, or is the average of the two middle points if n is even.

Variance and Standard deviation

- Measure the spread (or variation) of a distribution of a numerical variable.
- How do we compute variance?

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- You may curious about why the denominator is n-1 instead of n?

For those of you who are interested:

- The formula shown above is actually not the formula for computing variance, rather it is the formula for computing the sample variance.
- What is the difference?
- Variance is a mathematical concept defined as the expectation squared (i.e. to the power of 2) distance between of random variable and its mean. $\mathrm{Var}(\mathrm{X}) = \mathrm{E}(\mathrm{X} - \mu)^2$
- And sample variance is something that we used to estimate the numerical value of the variance.

$$s^2 = \frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}$$

- We put n-1 in the denominator to make s^2 an *unbiased* estimator of variance. (i.e. a more accurate estimate of the true variance, numerically)

- For standard deviation, we just take the square root of the variance.
- Compute standard deviation in R: By using sd() function
- Example:
- Suppose we have a data set: student_height
- We do: **sd(student_hight)** & **(sd(student_hight))^2** to compute the (sample) standard deviation and variance numerically.

Data frame

- Data frames are the two-dimensional version of a list. They are the most useful storage structure for data analysis
- Think of it as a table to store data.
- How do we create a data frame in R?
- By using data.frame() function
- Give data_frame() any number of vectors, each separated with a comma; each vector will be a column in the new data frame.

Example on working with data frame

- Student number & name has 1-1 correspondence.

```
student_num <- c(1, 2, 3, 4)
name <- c("Nadia", "Shiyi", "Yizhe", "Wei")
mydat <- data.frame(obsnum = student_num, student_name = name)
mydat
```

```
##   obsnum student_name
## 1      1        Nadia
## 2      2        Shiyi
## 3      3        Yizhe
## 4      4          Wei
```

Example on working with data frame

(Hint: if you want to investigate the content of a R object, just type in its name in the console.)

- Combined with group_by() function:

```
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)
```

student <dbl>	courses <chr>	grade <dbl>
1	STA130	82
1	MAT137	83
1	ECO100	77
1	CSC148	84
1	PHL100	79
2	STA130	83
2	MAT137	74
2	ECO100	85
2	CSC148	77
2	PHL100	72

```
marks
```

```
# A tibble: 10 x 3
   student courses grade
     <dbl> <chr>   <dbl>
 1       1 STA130     82
 2       1 MAT137     83
 3       1 ECO100     77
 4       1 CSC148     84
 5       1 PHL100     79
 6       2 STA130     83
 7       2 MAT137     74
 8       2 ECO100     85
 9       2 CSC148     77
10       2 PHL100     72
```

- If we don’t do group_by on marks and student,
- Then the summarise() function couldn’t distinguish different students. It treat the date frame marks as 1 single column.
- Let’s see the difference:

```
> summarise(marks, ave = mean(grade))
# A tibble: 1 x 1
    ave
  <dbl>
1  79.6
```

```
> summarise(marks_grouped, ave = mean(grade))
# A tibble: 2 x 2
  student   ave
    <dbl> <dbl>
1       1  81
2       2  78.2
```

- what do you notice?

- Homework question 1 (b)
- Create a data frame called *data* that contains the family id number and the numbers of kids in each family.

```
library(tidyverse)
data <- summarise(group_by(Galton, family),
numkids = mean(nkids))
data<-data.frame(data)
```

- In general, how we use group_by():

group_by(<data_frame> , <column of the variable you want to work with>)

Data Wrangling:

- Essentially, cleaning the data you have.
- Some general steps we want to perform:
- Create a dummy variable
- Removing the column
- Replace values above/below a certain threshold by NAs
- Taking the subset of variables
- Filtering the data frame based on a condition (e.g. based on one of the variables/columns)
- Renaming the variables
- Grouping the categories
- Merging two data frames based on a common variable/column
- Reshaping the data frame

Vocabulary for this week

- Data frame
- Vector (how do we create a vector in R?)

(by using c() function)

- Average
- Standard deviation
- Variance
- Missing data

Terms for describing data transformation and manipulation:

- Performed log-transformation on x
- Took the inverse of x
- Taking the difference between x and y
- Summing variables x, y, and z
- Taking the average of x and y
- Tabulate variable x and y (i.e. make a table)
- A quantitative variable x stratified by a categorical variable y with k-levels.

Group discussion on Question 3

- **Describe or explain to each other:**
- Describe what you did to create the variables.
- Explain why you did it this way.
- Compare graphs or summary statistics on the created variables.
- What were your main findings?

In-class writing exercise

- **Explain what you have discussed in your groups for question 3.**
- For example:
- 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?
- You can use any figures they've created, if useful.

Possible writing template:

- 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.

Example for Question 1:

- *Good example, for Question 1:*
- 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.

- *Mediocre example:*
- 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.