STA130 Winter 2020

(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-Winter-2020

viv.ngo@mail.utoronto.ca

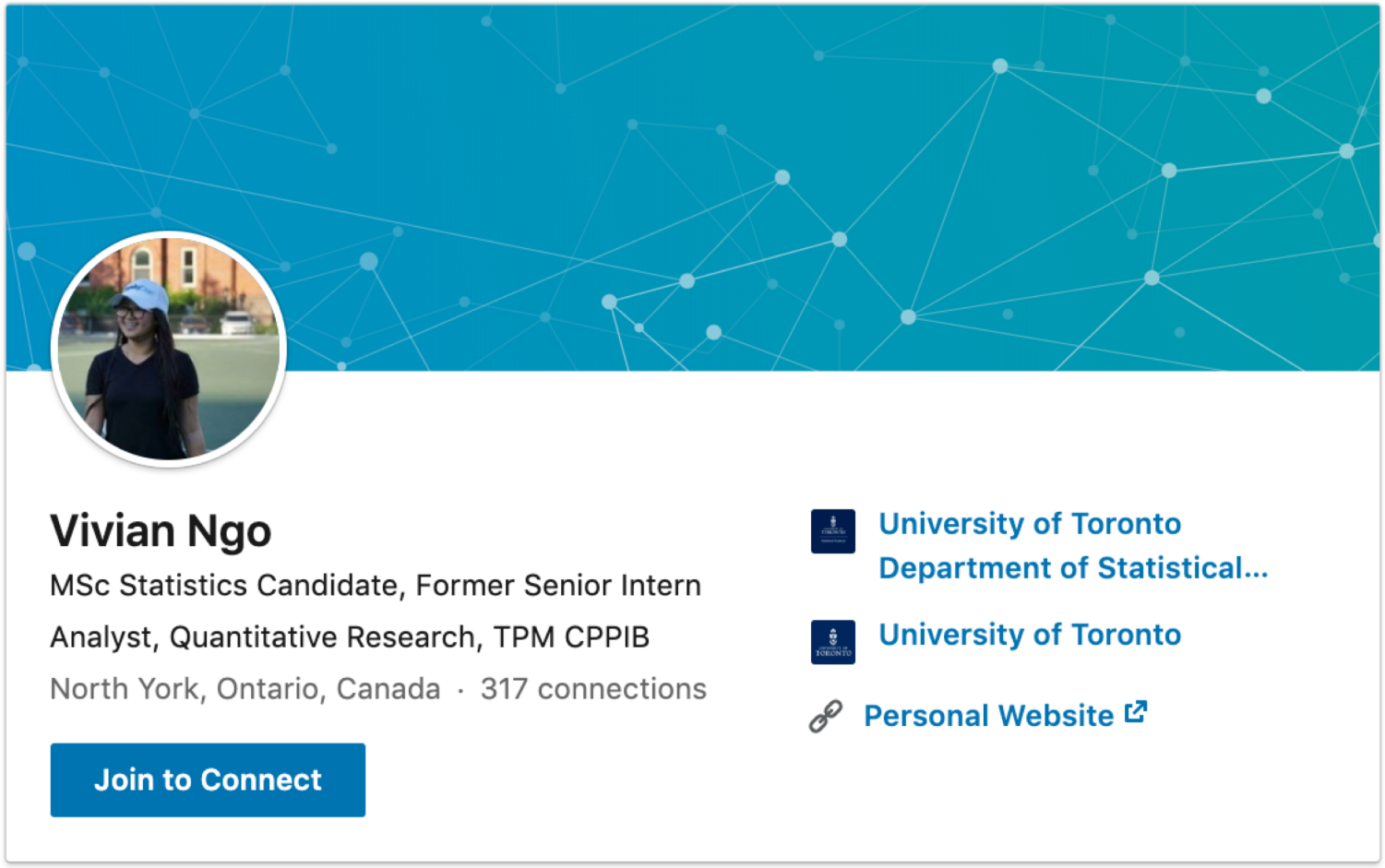
Agenda

- Introductions
- Vocabulary
- Group discussion
- Writing examples
- Writing exercise

Introduction

- TA: Vivian Ngo
- Tutorial Section: 0208
- Time: 2-4pm
- Room: LM123

About me


Vivian Ngo
MSc Statistics Candidate, Former Senior Intern
Analyst, Quantitative Research, TPM CPPIB
North York, Ontario, Canada · 317 connections
Join to Connect
University of Toronto
Department of Statistical...
University of Toronto
Personal Website

About tutorials

- Tutorials worth 20%, attendance is MANDATORY
- Each week, you have the opportunity to gain 6 points: 1 point attendance, 1 point practice problem, and 4 points for the writing/presentation exercise.
- The 3rd week will be a half tutorial, the second half is reserved for the mentorship program which is worth 3% of your grade. Additional details are on Quercus and will be discussed during class in week 3. Any questions about the program should be sent to Megan (her email is on Quercus).
- Have questions completed and submitted to Quercus, **no emailed homework will be accepted**
- Tutorial is **NOT** the place for troubleshooting R code. You should be prepared to discuss your results during tutorial. Instead, go to OH or post questions to the discussion board ahead of tutorial. OH are on Tues, Wed and Thurs.
- We use RCloud
- Respectfully participate in group work and class discussions
- Practice English writing & oral presentation skills, particularly for non-statistician audiences
- Show up on time, tutorial starts 10 past the hour and goes until the hour.
- Tutorial is a safe and friendly environment to ask questions and practice communication skills
- QUESTIONS?

Ice breaker!

Vocabulary

- **Bar graphs, histograms:**
  - Where are the data centered (towards the left, right, middle)
  - How much spread (relative to what?)
  - Shape: symmetric, left-skewed, right-skewed
  - The tails of the distribution (heavy-tailed or thin-tailed)
  - Modes: where, how many, unimodal, bimodal, multimodal, uniform
  - Outliers, extreme values
  - Frequency (which category occurred the most or least often; data concentrated near a particular value or category)
- **Scatterplots (bivariate or pairwise scatterplots):**
  - Strong / weak relationship
  - Linear / nonlinear relationship
  - Direction of association (positive or negative)
  - Outliers (deviation from what?)
  - Any visible clusters forming
  - Each dot represents …

Visualizations

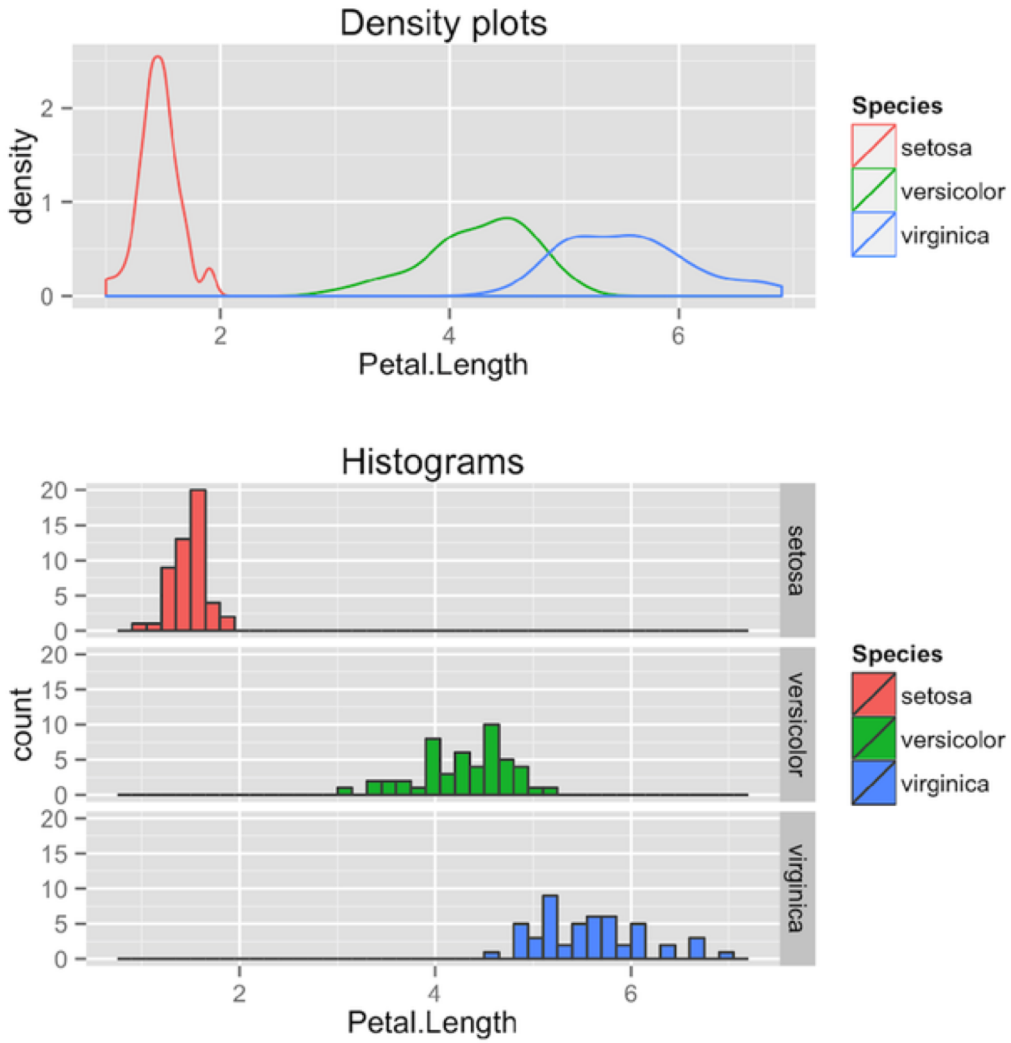
- What are the most effective types of graphs to summarize information in categorical or quantitative variables?
- What does the distribution tell you about for each types of data (categorical or quantitative)?
- How do you describe a histogram or a scatterplot? (refer to this week's vocabulary list)
- Be mindful of how you present your findings – are you potentially misleading the audience? Is it reader friendly?
- Think about the source of the data – this may have important implications for your interpretations!

Group Discussion

- What do you notice about the number of bins a histogram has, its shape and precision?
- In Question 1d, you could have presented both book cover types (hard or paper) in the same plot or presented them on separate plots. What are some considerations for which presentation you may want to choose (e.g. what are the pros and cons of each one)?
- If presenting two plots side by side, what are some things to consider to ensure they are comparable and reader-friendly?
- In questions 2 and 3, you saw examples of survivor bias. What is this? How does it impact, for example, the mean survival time calculated in Question 3?

Writing example

Iris dataset from R



Writing example

- Finding a way to lead a reader through a visual
- Describe what the graphs are telling us (x-axis, y-axis labels should be clear, etc)
- Come up with a “story” of main results
- Provide figures to support the “story”

Writing example

A possible writing template:

- Give some context to the variables you are graphing based on what you know about the dataset (units and types of variables involved should be clear).

Either:

- Give the most striking features of the graphs (contrast or similarity).
- Synthesize these features and make a conclusion based on these features.

Or:

- Make a statement or conclusion based on your impression.
- Explain each of the features of the graphs (contrast or similarity) that support your statement or conclusion.

Writing example #1

- The *petal length of* Iris setosa distributes differently from Iris versicolor and Iris virginica. The density plot/histogram of petal length of Iris setosa has a sharp peak while the other two have a flatter distribution.

Writing example #2

- We looked at the petal length of *Iris*. Specifically, Iris versicolor and Iris virginica, despite having different centres, have similar spread in terms of their petal length. Interestingly, the shape of distribution also differs between the species. We conclude that the petal length of Iris setosa, Iris versicolor and Iris virginica are different in terms of their centre, spread and shape.

Writing Example #3

- The *petal length in c.m.* of *50 samples from each of three species of Iris* (*Iris setosa, Iris virginica and Iris versicolor*) was investigated/examine/summarized. The graph suggested that *distribution of petal length is species dependent*. In particular, petal length of Iris setosa is shown to be less variable than Iris versicolor and Iris virginica. However, despite Iris versicolor having on average longer petals than Iris virginica, the range of petal length is similar for these two species. Further, the shape of distribution also differs according to species, with Iris setosa more or less symmetric about its centre, and versicolor and virginica skewed to the left and right, respectively.

Writing Activity

- **Write a short paragraph to describe coherently the graphs you produced and structure these graphs to tell an interesting “story” about the data used in Question 1.**
- You should use at least 2 graphs or plots from question 1 to support your story.