STA130 TUT0110 Week1

TA: Gloria

shiyi.hou@mail.utoronto.ca

Self Introduction

- I am a fourth-year student, I am in Statistics Specialist Program.
- This is my second time TA-ing this course.
- Among all the statistics topics, I am mostly intrigued by Machine Learning.

Ice-breaker

- Attendance is mandatory. Show up on time.
- ELL and Writing center are available on campus.
- Mentorship Program: 4 events in total, count towards 3% of total mark.

Tutorial expectations

- Complete your homework before you coming to Friday's tutorial.
  - Submit on Quercus; no emailed homework will be accepted.
- We don't troubleshoot R code during tutorials. However, you could go to OH or post questions to Piazza ahead of tutorial.
- Recommend using RStudio instead of R.
- Tutorial is a safe and friendly environment to ask questions and practice your communication skills.

Agenda

- Statistics vocabularies
- Describing visualizations
- Group discussion
- Example on written activity
- Writing exercise

Vocabularies

BAR GRAPHS

HISTOGRAM

- 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)

Vocabularies

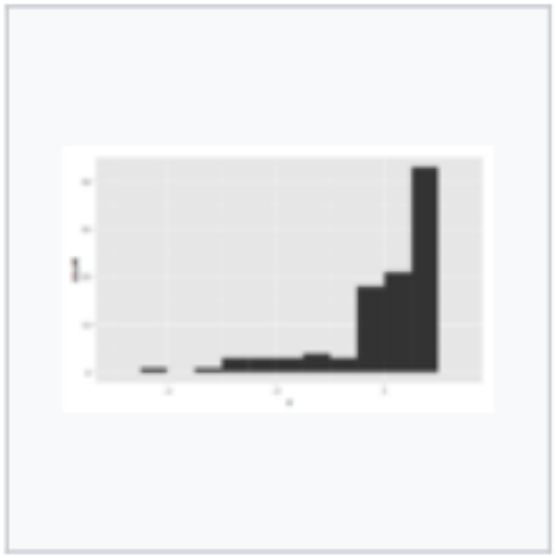
SCATTER PLOT

- Strong / weak relationship
- Linear / nonlinear relationship
- Direction of association (positive or negative)
- Outliers (deviation from what?)
- Any visible clusters forming
- Each dot represents …

Skewness

Symmetric, unimodel

Right-skewed



Left-skewed

Mode

bimodel

Multimodel

Symmetric

Examples from homework Q1

- What types of figures (that we've learned so far) would be appropriate for this question? Why or why not?
- What type of distribution does the number of pages have?
- Is there an association between number of pages and weight? Number of pages and type of cover (hard or paper)?

Discussion

- question 1 from the homework
  - 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?

Discussion

- If presenting two plots side by side, what are some things to consider to ensure they are comparable and reader-friendly?

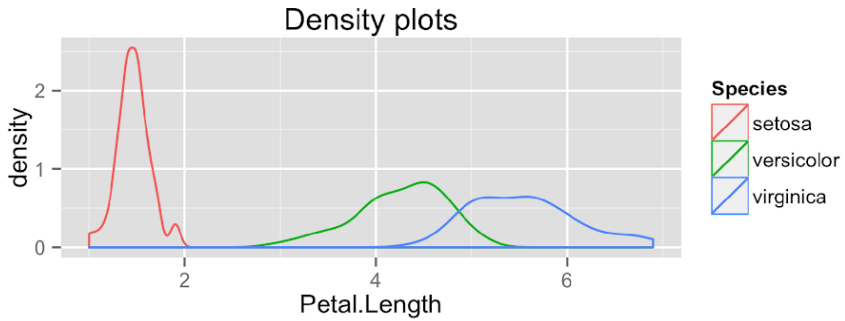
Discussion

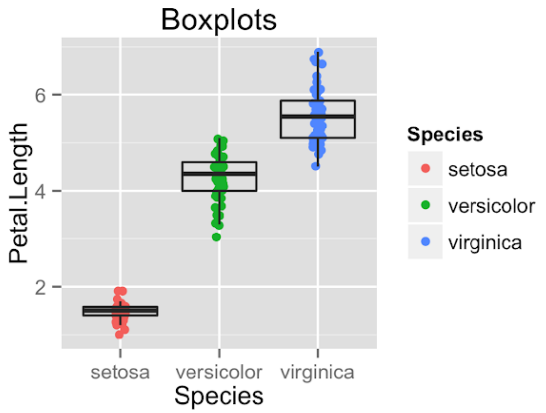
- In questions 2 and 3, you saw examples of survivor bias.
  - What is this?
  - How does it impact? (For example, the mean survival time calculated in Question 3)

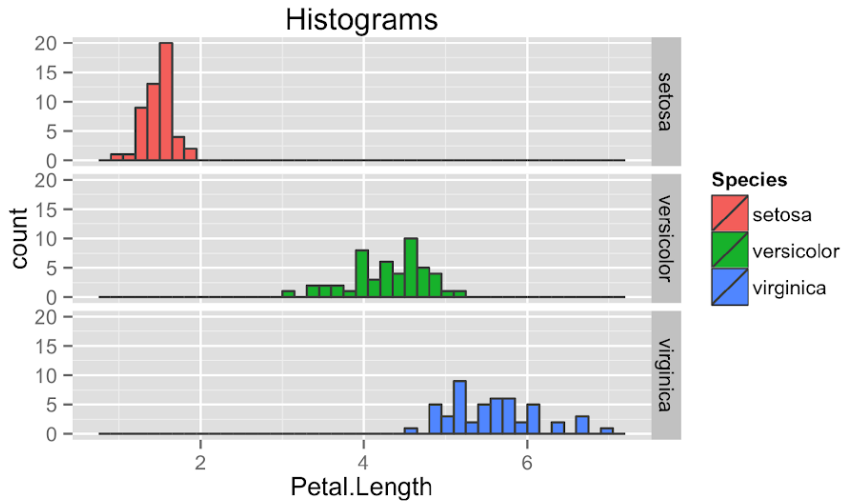
How to synthesizing information from visualizations

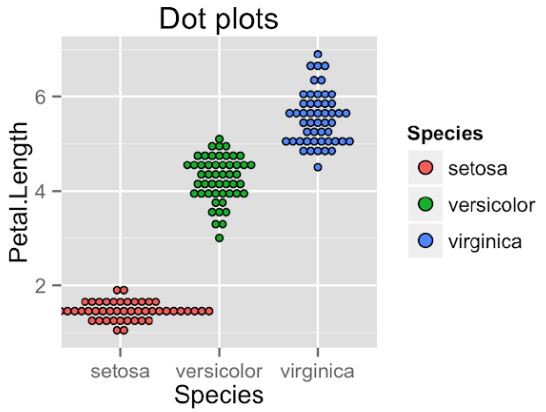
- think about the most logical order in which to lead the reader through the visual
  - Describe what their graphs are telling us; i.e., what type of relationship(s) are apparent, the x- and y-axis labels should be clear, etc.
  - Come up with a "story" of your main results, use a few of your graphs. You can make one not originally asked in the question, if it will help and is appropriate; Consider the logical order that you 'tell your story' – how will you most effectively tell your 'story'.
  - Make sure you provide figures to support your 'story'.

Example: synthesizing information from visualizations









What do you think of this example?

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.

- *The **context is missing** and **description** of graphs **is not specific** and too vague (distributes differently in terms of what?).*
- *The **most striking features** of the graphs **were not mentioned***
- *The **conclusion is missing** or **not supported** by any statements.*
- ***No transitions.***

What about this example?

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.

- *There was **some context,** but the **connection** to all variables used was **not clear.***
- *The **organization is logical** and **appropriate use of transitions.***
- *The **feature statements** contain some information but **too vague.***
- *Descriptions of graphs **somewhat support** the conclusion.*

What about this example?

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.

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 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**
  - use at least 2 graphs or plots from question 1 to support their story.
  - Submit on Quercus

JANUARY 10, 2020 TUTORIAL

Writing activity rubric: STA130 - Writing Activity Rubric.pdf

At the **end of tutorial**, submit your written exercise here: Tutorial 1 Writing Activity