STA130 Week 1 Tutorial

Kaixuan (Bryan) Hu

January 11, 2019

Introduction

- Tutorial: 20% of the final grade.
- Attendance is mandatory.
- 1 point attendance, 1 point practice problem, and 4 points writing/presentation exercise.
- 3rd week will be a half tutorial, the second half is reserved for the mentorship program which is worth 3% of the final grade.

- 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 Thursdays 11am-2pm.
- Participate in group work and class discussions
- You will get practice for writing/ oral presentation skills during the tutorial
- Show up on time, tutorial starts at 10:10AM. If they need to miss a tutorial or leave early for a test, you should let me know ahead of time.
- Notice that showing up late or leaving more than 15 minutes early will result in an attendance score of zero.
- Tutorial is a safe and friendly environment to practice your communication skills (critical skills for the future careers, in statistics or otherwise)

Outline for today

- Learn how to describe data visualizations aided by the vocabulary list provided below.
- Later you will be assigned to small groups for group discussion regarding question 2 from the homework

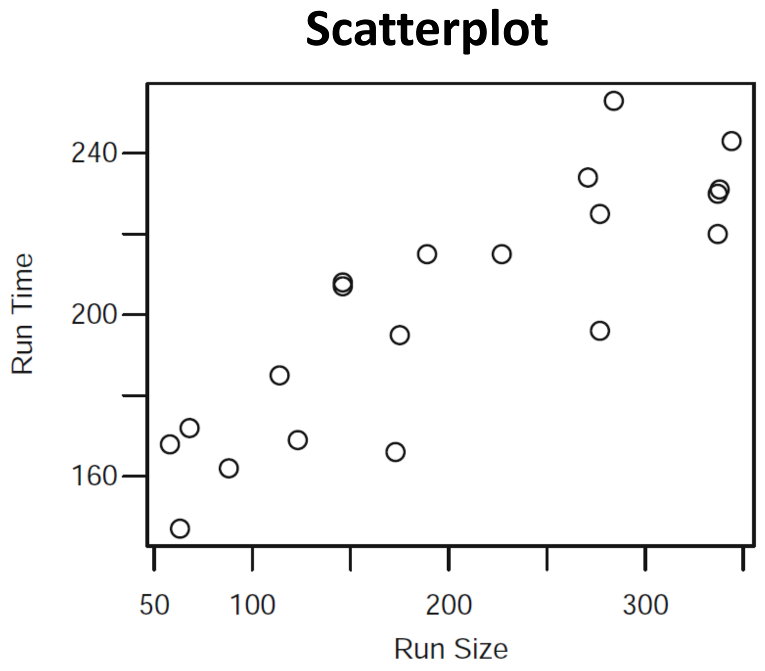
Data Visualization

- What a visualization does and how to describe a visualization?

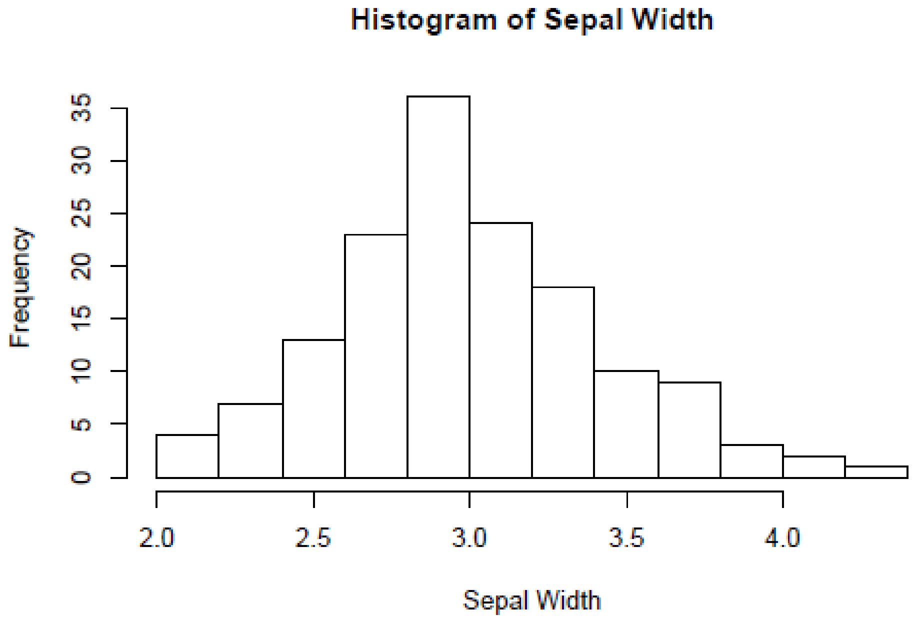
Basically, Exploring a dataset using graphs and come up with a story to describe the dataset.

- What are the most effective types of graphs to summarize information in categorical or quantitative variables?

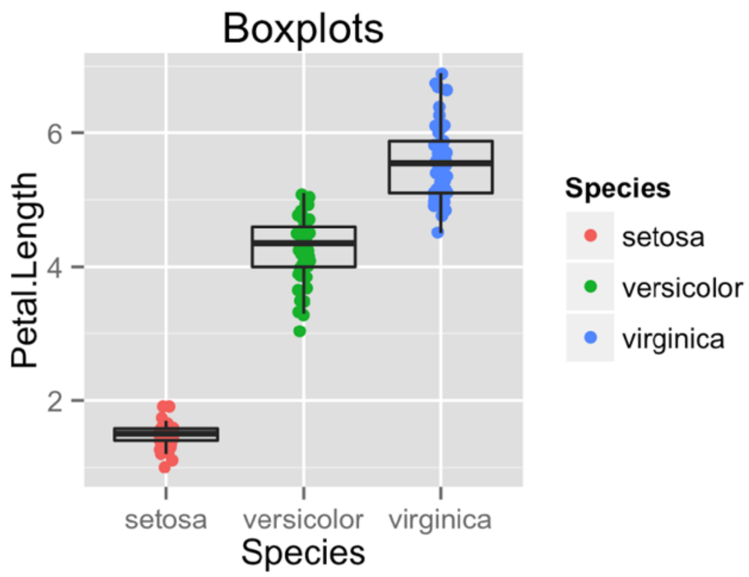
Typically, **histogram**, **scatterplot, Boxplots** are the most commonly used graphs.



Scatterplot



Histogram of Sepal Width



Boxplots

How do you describe them?

Key things to look at:

- Look for **distribution**

What does the distribution tell you about for each types of data?

e.g. Normally distributed; Symmetric; Skewed right; Skewed left;
Bimodal (i.e. double peak) distributed; Multimodal (i.e. multiple peak) distributed.

Use Statistics terminology! (Vocabularies when describing distributions of variables or relationships between two variables)

- Look for **relationships between two variables (e.g. trend)**

e.g. Linear (positive or negative) / non-linear relationship;

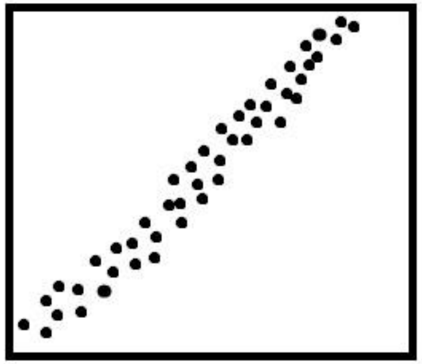
- Boxplots, histograms:

1. Where it is centred (towards the left, right, middle)
2. How much spread? (relative to what?)
3. The tails relative to a normal distribution (fat-tailed or heavy-tailed & thin-tailed)
4. Modes (i.e. peak): where, how many? (unimodal, bimodal, multimodal, uniform)
5. Symmetric; Skewness (left-skewed, right-skewed)
6. Outliers; Extreme values
7. Frequency (which category occurred the most or least often; data concentrated near a particular value or category)

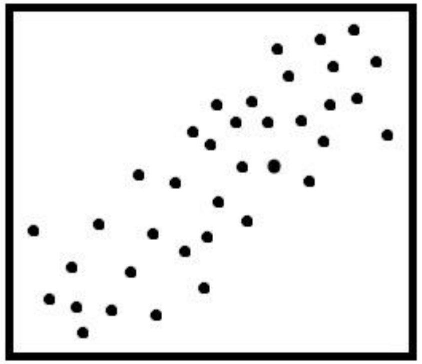
- Scatterplots:

Used to compare the relationship between two variables

1. Strong / weak relationship
2. Linear (positive or negative) / non-linear relationship
3. Outliers (deviation from what?)
4. Any visible clusters forming.



strong positive linear association

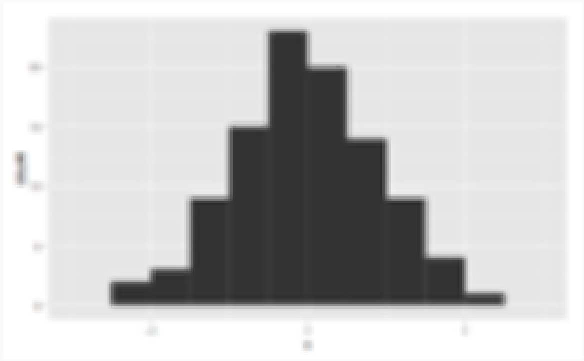


weak positive linear association

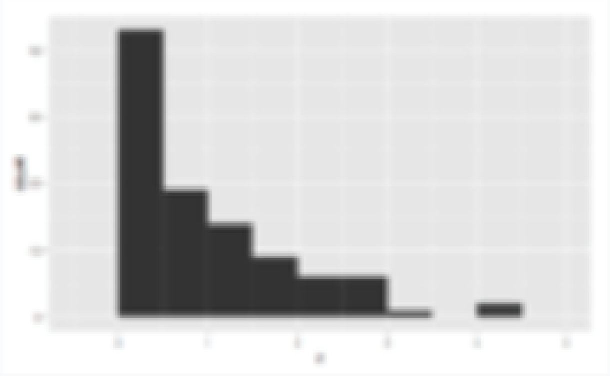
One of the **objective** of this course is to get you familiar with the statistical language, so it is helpful to understand and memorize the word or phrase shown above.

Try it by yourself

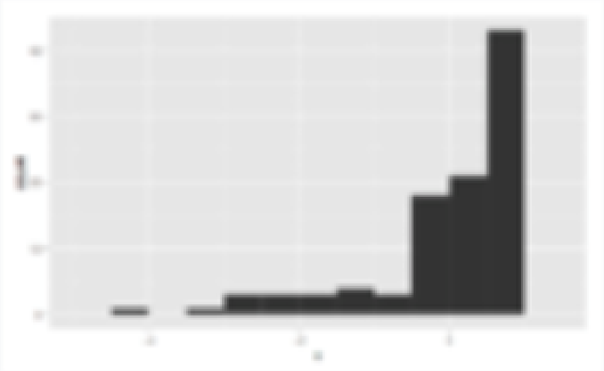
- Describe the following graphs using one or more correct (and precise) vocabularies.



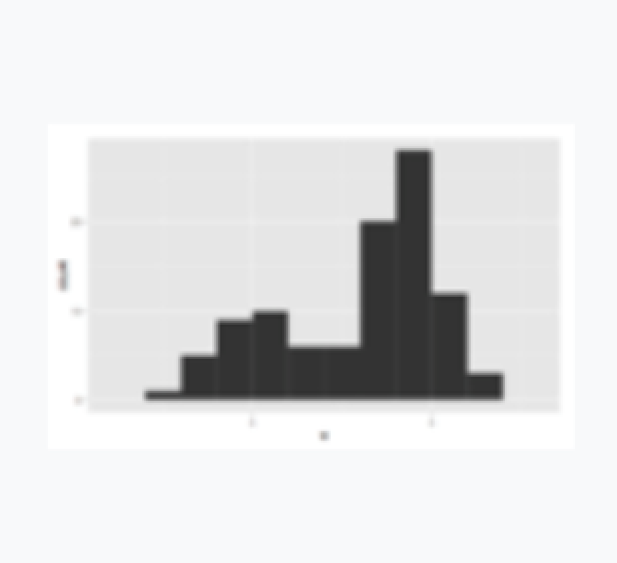
Symmetric or unimodal



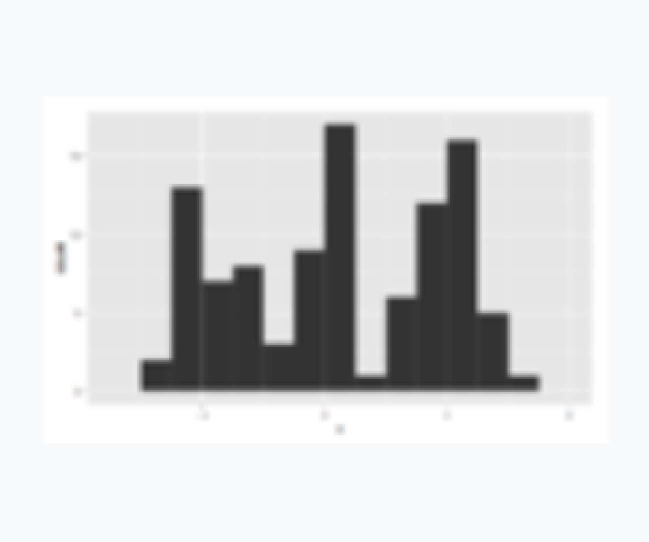
Skewed right



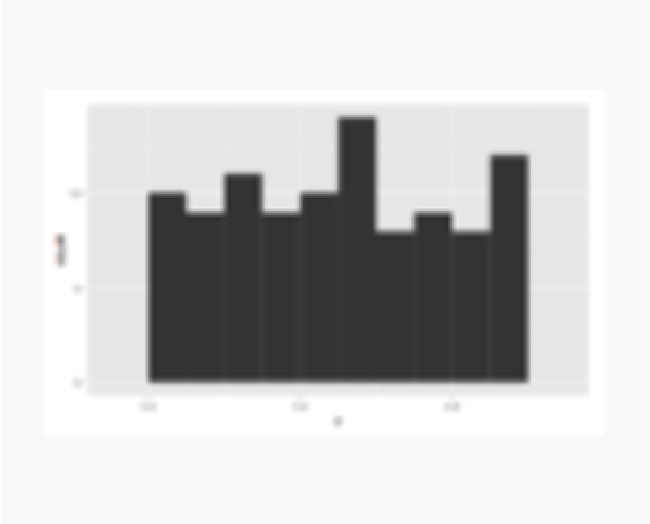
Skewed left



Bimodal



multimodal



Symmetric

Group Discussion

- Form a groups of 3-4 to discuss about question 2 from the homework.
- And explain to each other:
  - Describe what their graphs are telling us; i.e., what type of relationships do you notice between parents' height and that of their children?
  - Come up with a "story" of your main results, use a few of your graphs (you can make ones not originally asked in the question, if it will help and is appropriate).
  - Explain what logical order is the most effective to tell their stories.
  - Tell the story to each other.

synthesizing information from visualizations

- You should think about the most logical order in which to lead the reader through the visual.
- *Possible writing templates:*
- 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).
- Give the most striking features of the graphs (contrast or similarity). Synthesize these features and make a conclusion based on these features.
- 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 exercise

- Write a short paragraph to describe coherently the graphs you produced from question 2 and structure these graphs to tell a story.
- Use at least 3 graphs from question 2 to support the story.
- HAND IN for evaluation via Quercus by the end of tutorial.

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.

Vocabulary list

Histograms, boxplots, bar graphs:

1. Where it is centred (towards the left, right, middle)
2. How much spread? (relative to what?)
3. The tails relative to a normal distribution (fat-tailed or heavy-tailed & thin-tailed)
4. Modes (i.e. peak): where, how many? (unimodal, bimodal, multimodal, uniform)
5. Symmetric; Skewness (left-skewed, right-skewed)
6. Outliers; Extreme values
7. Frequency (which category occurred the most or least often; data concentrated near a particular value or category)

Scatterplots:

1. Strong / weak relationship
2. Linear (positive or negative) / non-linear relationship
3. Outliers (deviation from what?)
4. Any visible clusters forming.