STA130 T0209

Week 1: Data Visualization

(Materials used in this presentation are provided by the U of T Statistics Department)

Agenda

- Introduction
- Visualizations and Vocabulary list
- Group discussion
- Writing example
- Short writing exercise

Introduction

- L0201, Nathalie Moon, Monday 2-4pm
- T0209, Vivian Ngo, Friday 2-4pm, AP124

- TA: Vivian Ngo
- Statistics Specialist, Mathematics Major
- www.linkedin.com/in/vivian-ngo-a30bb5a8

Visualizations and Vocabulary list

Vocabularies when describing distributions of variables or relationships between two variables:

Bar graphs, histograms:

1. Where it is centered (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 and thin-tailed)
4. Modes: where, how many, unimodal, bimodal, multimodal, uniform
5. symmetric, 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 (bivariate or pairwise scatterplots):

1. strong / weak relationship
2. linear (positive or negative) / nonlinear relationship
3. outliers (deviation from what?)
4. any visible clusters forming
5. Each dot represents ...

Visualizations and Vocabulary list

General considerations:

- 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)?
- What kind of trend do you observe in some of the variables?
- How do you describe a histogram or a scatterplot?

Visualizations and Vocabulary list

Symmetric, unimodal

Skewed right

Skewed left

Visualizations and Vocabulary list

Bimodal

Multimodal

Symmetric

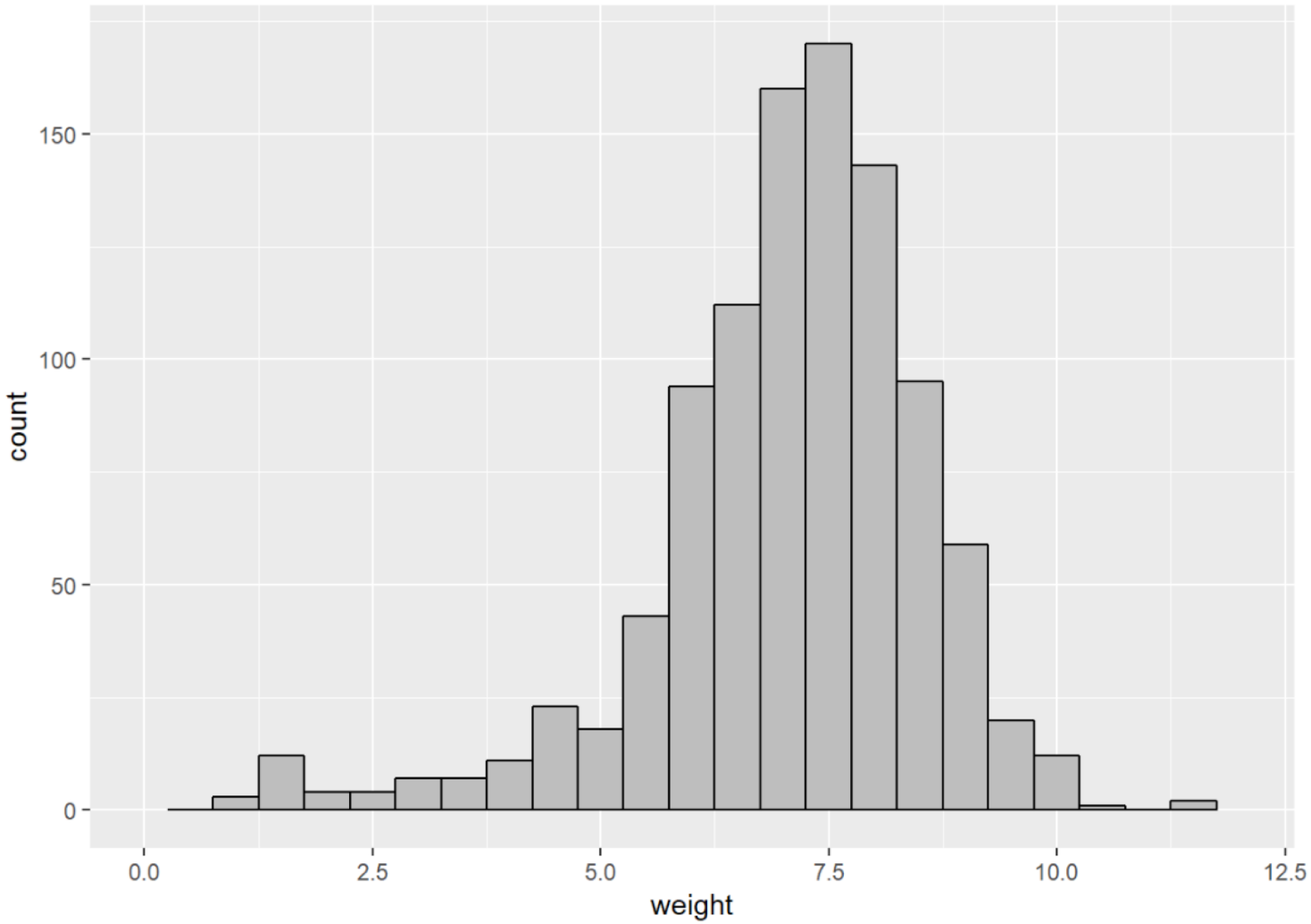
Visualizations and Vocabulary list

Data specific questions for homework question 1 (for example):

- What type of distribution does babies' weight have?
- Is there an association between the mother's age and the baby's weight?
- What types of figures would be appropriate for this question? Why or why not?

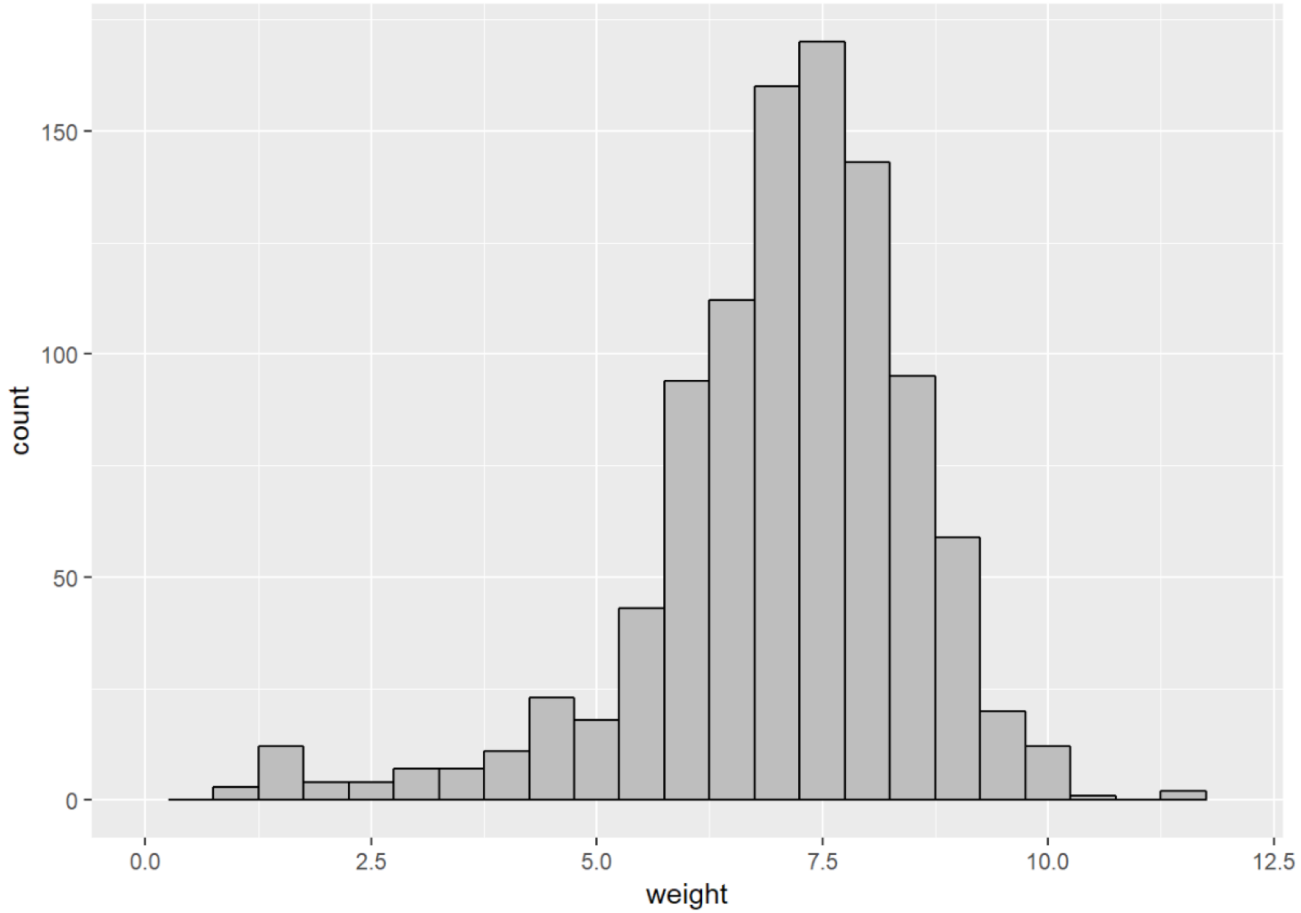
```r
# Construct your plots in this code chunk
library(tidyverse)
library(openintro)

ggplot(data = ncbirths) + aes(x = weight) + geom_histogram(fill = "grey", colour = "black", bins=25) + xlim(0,12)
```



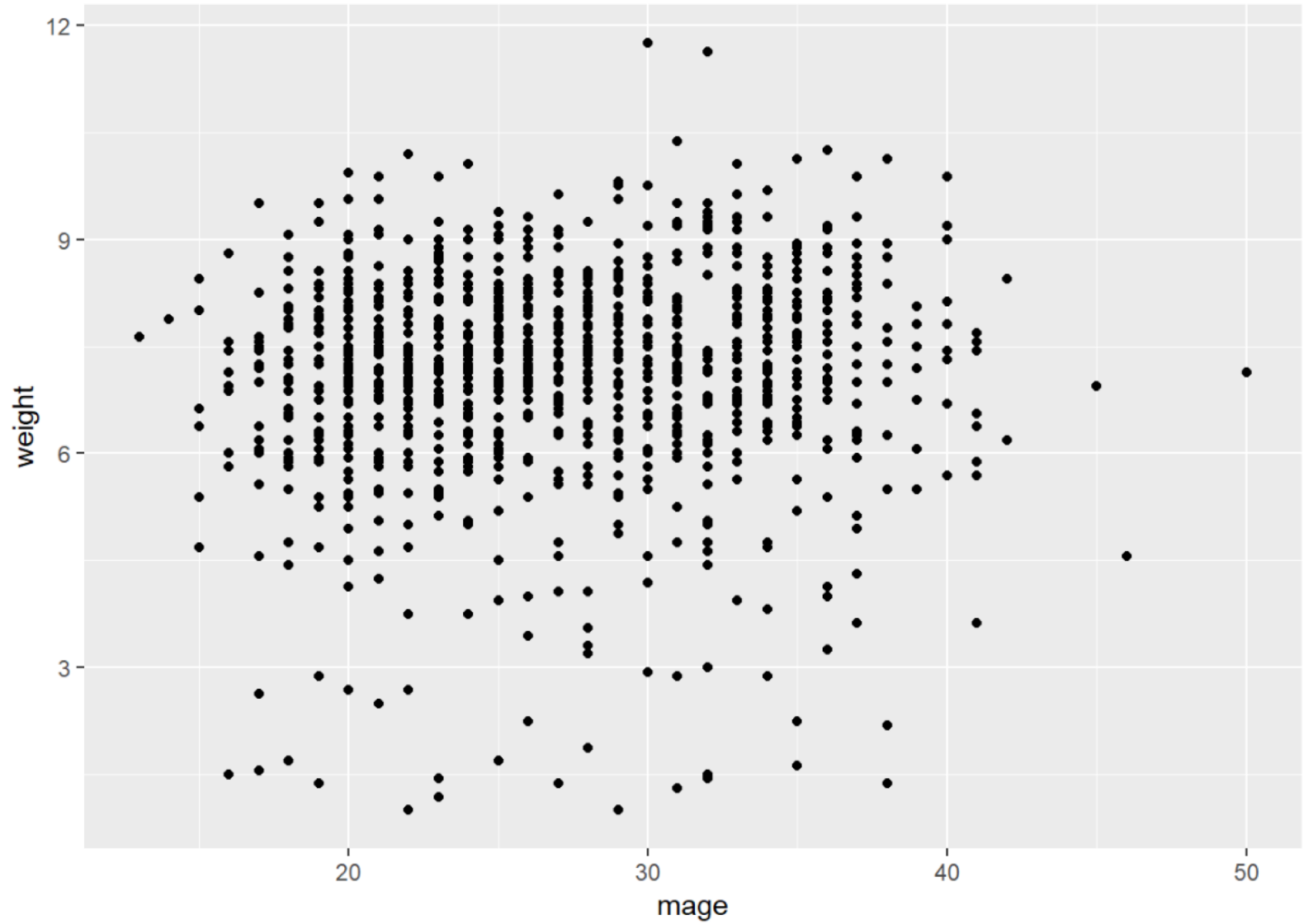
```r
# Construct your plots in this code chunk
library(tidyverse)
library(openintro)

ggplot(data = ncbirths) + aes(x = weight) + geom_histogram(fill = "grey", colour = "black", bins=25) + xlim(0,12)
```

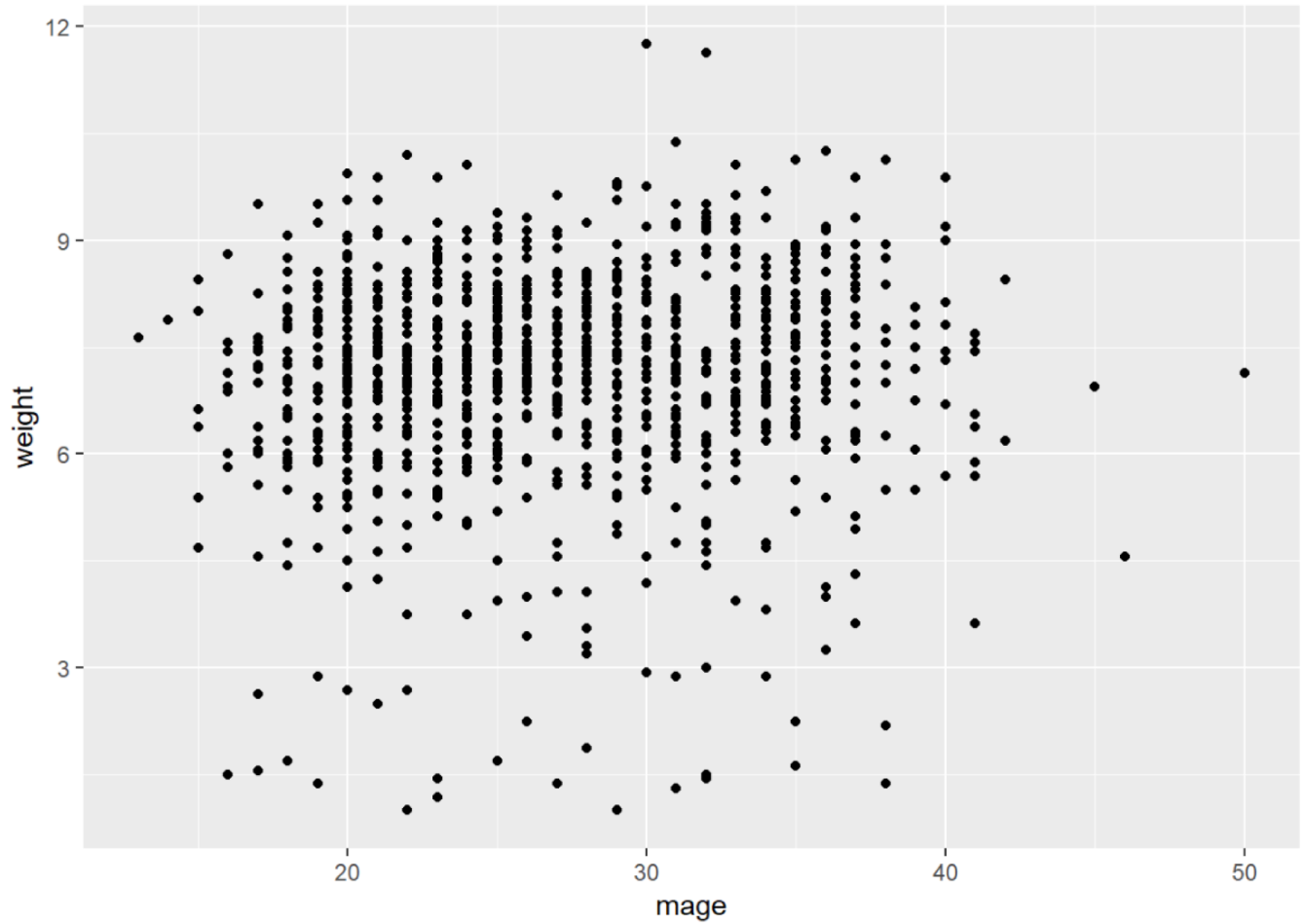


- Baby's birthweight is left skewed since the left tail is longer than the right tail. There is one prominent peak so the distribution is unimodal. The mode is around 7.5 pounds.

```
ggplot(data = ncbirths) +
  aes(x = mage, y = weight) +
  geom_point()
```



```
ggplot(data = ncbirths) +
  aes(x = mage, y = weight) +
  geom_point()
```



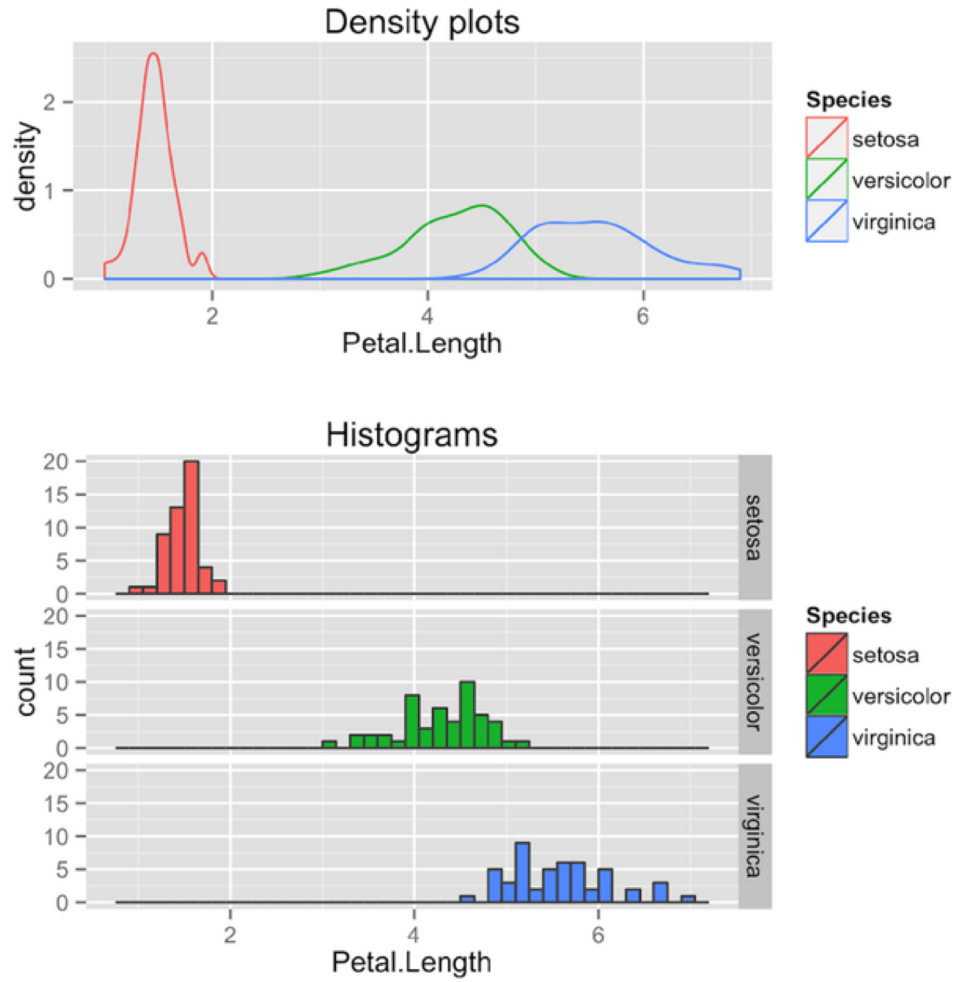
- There is no strong association between the mother's age and the baby's weight. There is no obvious pattern in the dots which would suggest a relationship between these two variables.

Group Discussion

- Homework Question #2
  - Describe what the 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.

Writing example

Iris dataset from R



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 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.

- *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 and the conclusion is missing or not supported by any statements. No transitions.*

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.

- *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.*

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.

- *Provided the context. The organization is logical. The feature statements are detailed and informative. Descriptions of graphs support the conclusion.*

Short writing exercise

- Write a short paragraph to describe coherently the graphs you produced and structure these graphs to tell a story. Use at least 3 graphs from question 2 to support the story.
- Submit on Quercus

	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.