STA 1013 : Statistics through Examples

Lecture 14: Review for Quiz 2

Hwiyoung Lee

October 2, 2019

Department of Statistics, Florida State University

Quiz 2

- **Oct 4 (Fri), 2019**
- Topics : **Lecture note 7** $\sim$ **Lecture note 13**
  - Exercises and Examples in the Lecture notes
  - Practice Problems
- You can use your calculator
- Bring one piece of hand written cheat sheet (both side allowed)

Measure of Center

Measure of Center

Mean

- Takes every value into account

$$\text{Mean} = \frac{\text{sum of all values}}{\text{total number of values}} = \frac{\sum_{i=1}^{n} x_i}{n} = \bar{x}$$

Median

- The middle value of the ordered data
- Half of the observations are larger and half are smaller than the Median

Mode

- The value of the data that occurs with the greatest frequency.
- The mode **may not exist**
- The mode **may not be unique**

Example : Annual Income

Bill Gates moves to town

Name	Annual Income
Tom	$ 32,000
Larry	$ 36,000
Susan	$ 39,000
Paul	$ 41,000
Marcus	$ 50,000
Randy	$ 57,000
Sandy	$ 60,000
Tim	$ 75,000
Pam	$ 80,000
Kim	$ 95,000
Bill Gates	$ 5,000,000,000

Mean ? Mean is very sensitive to the outliers

Example : Median

Example : 5, 11, 1, 13, 6, 9, 8, 3

- Sort : 1, 3, 5, 6, 8, 9, 11, 13
- Median : $(8+1)/2$ th observation
  - 4.5 th observation : average of 4 th and 5 th observation
- Q: What happens to the median if we change the 1 to -1,000?

- Mean :
- Q: What happens to the mean if we change the 1 to -1,000?

Mode

Mean, Median always exist and are always unique

1. The mode **may not exist**
   - Example : 1, 2, 3, 4, 5, 6, 7, 8, 9
2. The mode **may not be unique**
   - Example : 1, 2, 2, 2, 5, 6, 7, 7, 7, 8, 8, 9, 10, 10, 10

Weighted Mean

A weighted mean accounts for variations in the relative importance of data values.

$$\begin{aligned}\text{Weighted mean} &= \frac{\text{Sum of (each data value} \times \text{its weight)}}{\text{sum of all weights}} \\ &= \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}\end{aligned}$$

- Each data value is assigned a weight (w_i)
- Weighted means are appropriate whenever the data values vary in their degree of importance

Measure of Variation

Measure of Variation

Range

$$\text{Range} = \max - \min$$

IQR

$$\text{IQR} = Q_3 - Q_1$$

Standard Deviation

$$s = \sqrt{\frac{\text{Sum of (Data value - mean)}^2}{\text{The number of observations-1}}} = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{\mathbf{n-1}}}$$

Measure of Variation

Five number summary

$$\textbf{min}, \quad \mathbf{Q_1}, \quad \textbf{median}, \quad \mathbf{Q_3}, \quad \textbf{max}$$

Identify Outliers

Calculate the lower and upper fences:

$$\text{LF} : Q_1 - 1.5 \times \text{IQR}$$
$$\text{UF} : Q_3 + 1.5 \times \text{IQR}$$

Outliers are values that lie outside of the fences

Visualization

Histogram

Histogram is the generalized version of Stem and Leaf plot, bar chart

- The first step is to bin the data
- Count the frequencies
- Do same steps in a bar chart

Example : Weight data

Draw the histogram of the weight data given below

195.6	200.4	165.6	165.3	191.7	169.3	153.2
189.5	170.4	149.3	185.3	150.3	179.6	160.3
198.5	163.2	166.3	197.3	201.3	168.2	198.4

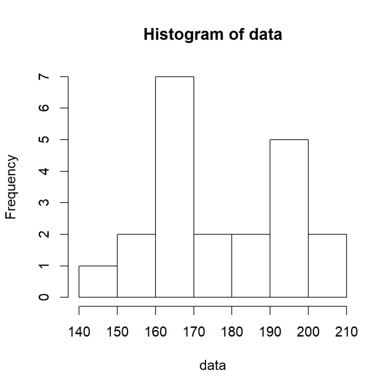
Use the following bins :

$$[140, 150), [150, 160), [160, 170), [170, 180), [180, 190), [190, 200), [200, 210)$$

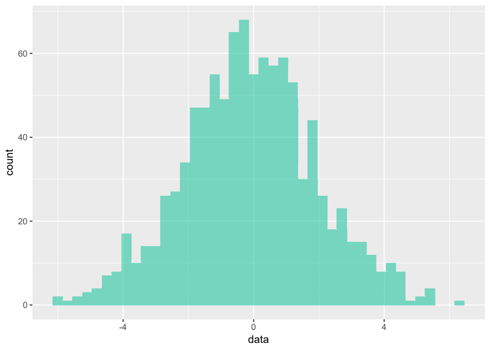
Histogram of the weight data

Binned weight data

Weight	Count
140 ~ 149.9	1
150 ~ 159.9	2
160 ~ 169.9	**7**
170 ~ 179.9	2
180 ~ 189.9	2
190 ~ 199.9	5
200 ~ 209.9	2

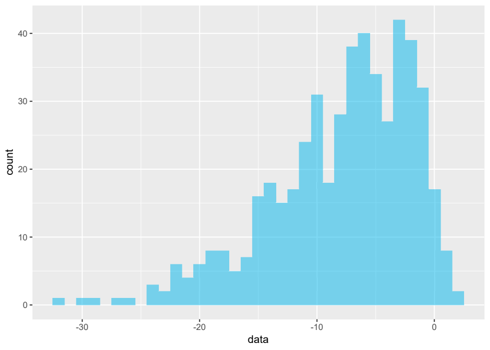


Shape of data (Symmetric)



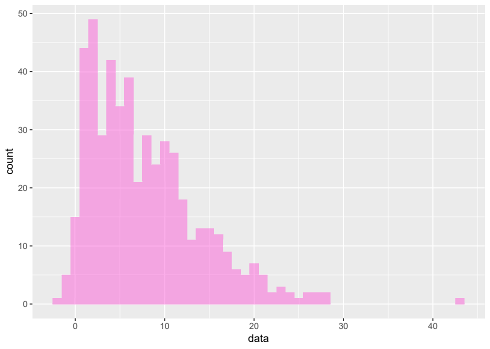
- Bell shape
- Left half is a mirror image of its right half
- **Mode = Median = Mean**

Shape of data (Left Skewed)



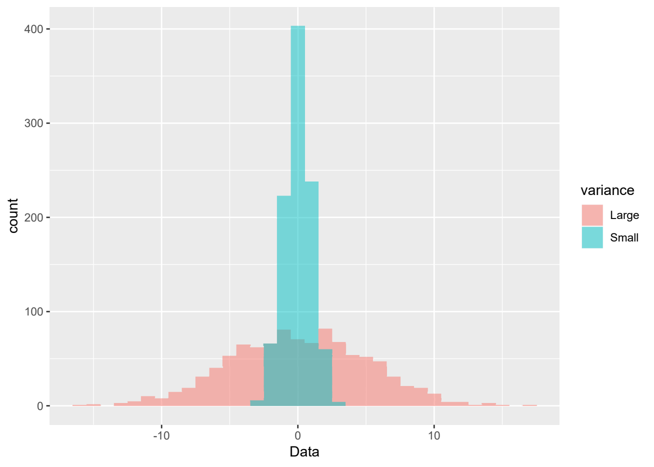
- Values are more spread out on the left side
- Values are concetrated on the right side (large value)
- **Mode > Median > Mean**

Shape of data (Right Skewed)

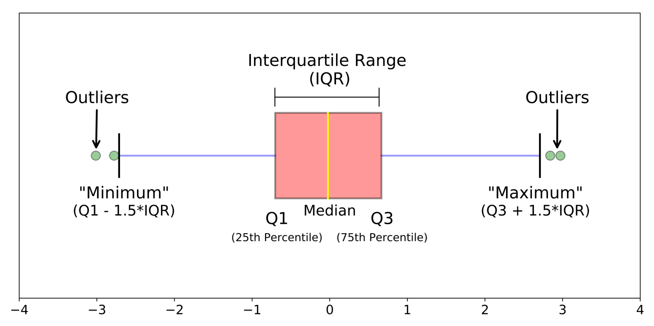


- Values are more spread out on the right side
- Values are concetrated on the left side (small value)
- **Mode < Median < Mean**

Shape of data (Variance)



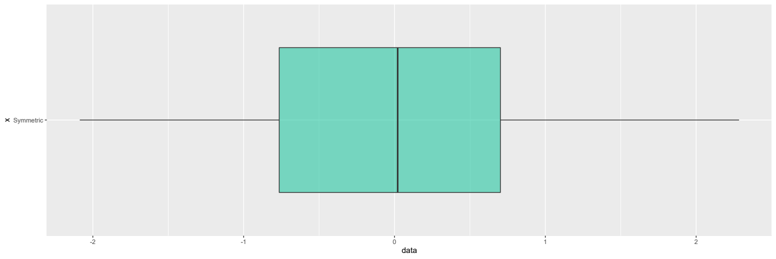
Box plot



- First find the five-number summary
- Find the lower, upper fence (Detect outliers)
    - No outliers → whiskers : min and max of the data
    - Outliers exist → whiskers : min and max values inside the lower, and upper fence

Shape of Box plot (Symmetric)

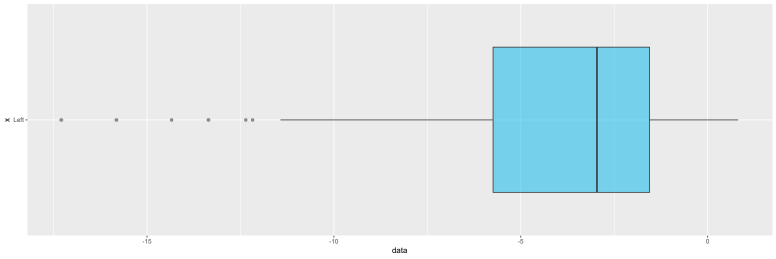
Box plot of Symmetric data



- Median located on the center of the box
- the left and right tails are equally balanced

Shape of Box plot (Left skewed)

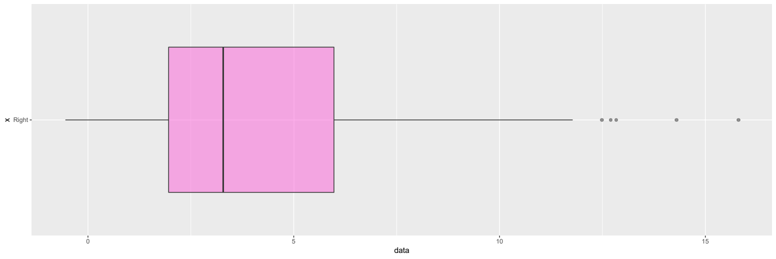
Box plot of the Left Skewed data



- Median closer to the upper quartile (Q_3)
- There are Low outliers (left side)
- Left whisker is longer than the right whisker

Shape of Box plot (Right skewed)

Box plot of the Right Skewed data

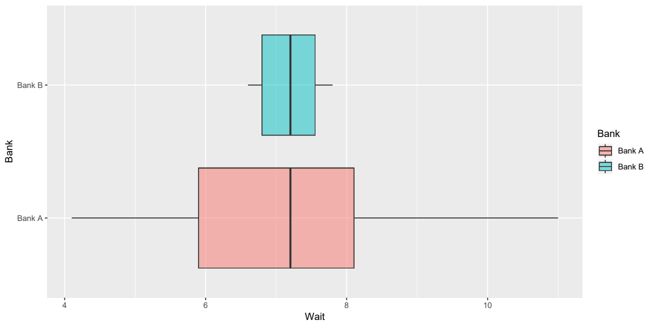


- Median closer to the lower quartile (Q_1)
- There are High outliers (right side)
- Right whisker is longer than the left whisker

Shape of Box plot (Variance)

Bank A	4.1	5.2	5.6	6.2	6.7	7.2	7.7	7.7	8.5	9.3	11.0
Bank B	6.6	6.7	6.7	6.9	7.1	7.2	7.3	7.4	7.7	7.8	7.8

Box-plot of the bank data



Normal distribution

The Normal Distribution

If we overlay the histogram with a smooth curve, the shape of this smooth distribution has **three** important characteristics:

1. Single peaked (Unimodal)
2. Symmetric around its single peak
3. “Bell-shaped” distribution

The smooth distribution, with these three characteristics, is called a **Normal distribution**.

Empirical Rule (Approximation)

The 68-95-99.7 rule for a Normal Distribution

- About 68% of the data values fall within 1 standard deviation of the mean.
- About 95% of the data values fall within 2 standard deviations of the mean.
- About 99.7% of the data values fall within 3 standard deviations of the mean.

Normal Probability (Exact)

- Find the probability from the given value :

  normalcdf(Lower value, Upper value, μ, σ)

- Find the value from the given probability :

  invnorm(Left Tail probability, μ, σ)