Math 140
Introductory Statistics

Professor Silvia Fernández

Lecture 4
Based on the book *Statistics in Action*
by A. Watkins, R. Scheaffer, and G. Cobb.

Stemplots

- Also called **stem-and-leaf plots**.
- Numbers on the left are called **stems** (the first digits of the data value)
- Numbers on the right are the **leaves**. (the last digit of the data value)

Mammal speeds:
- 11,12,20,25,30,30,30,32,35, 39,40,40,40,42,45,48,50,70.

1	1 2
2	0 5
3	0 0 0 2 5 9
4	0 0 0 2 5 8
5	0
6	
7	0

3 | 9 represents 39 miles per hour.

Stemplots (split)

- Each original stem becomes two stems.
- The unit digits 0,1,2,3,4 are associated with the first stem and they are placed on the first line.
- The unit digits 5,6,7,8,9 are associated with the second stem and they are placed on the second line from that stem.

1	1 2
-	
2	0
-	5
3	0 0 0 2
-	5 9
4	0 0 0 2
-	5 8
5	0
-	
6	
-	
7	0

3 | 9 represents 39 miles per hour.

Stemplot vs split stemplot

Mammal speeds:
- 11,12,20,25,30,30,30,32,35,39,40,40,40,42,45,48,50,70.

1	1 2
2	0 5
3	0 0 0 2 5 9
4	0 0 0 2 5 8
5	0
6	
7	0

3 | 9 represents 39 miles per hour.

1	1 2
-	
2	0
-	5
3	0 0 0 2
-	5 9
4	0 0 0 2
-	5 8
5	0
-	
6	
-	
7	0

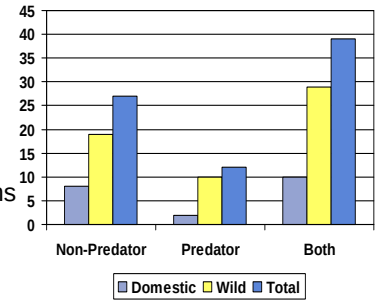
3 | 9 represents 39 miles per hour.

Stemplots

- Stemplots work best when
  - Plotting a single quantitative variable
  - Small number of values to plot
  - Want to keep track of individual values (at least approximately)
  - Have two or more groups that we want to compare

Bar Graphs

- One bar for each category.
- The height of the bar tells the frequency.
- Bar graphs have categories in the horizontal axis, as opposed to histograms which have measurements.



2.3 Measures of Center and Spread

- Before we used visual methods (estimations) to find out center (e.g. mean) and spread (e.g. SD). Now we will learn how to calculate them exactly.
- Measures of Center
  - Mean
  - Median
- Measures of Spread
  - Standard Deviation
  - Inter Quartile Range

Measures of Center

- Mean

The average of the data values denoted $\overline{x}$.

- Calculated as:

$$\overline{x} = \frac{\text{sum of values}}{\text{number of values}} = \frac{\sum x}{n}$$

- Example. Data Set: 5,12,34,18,37,11,9,21,30,6

$$\overline{x} = \frac{5+12+34+18+37+11+9+21+30+6}{10} = 18.3$$

Measures of Center

- **Median**

The value that divides the data into equal halves. Denoted $median$ or Q_2.

- Calculated as:
  - List all values in increasing order and find the middle one.
  - If there are n values then the middle one is $(n+1)/2$
  - If n is even use the fact that the mid-value between a and b is $(a+b)/2$

Measures of Center

- **Median (examples)**

Data set: 5,12,34,18,37,11,9,21,30,6.

Ordered data set:

5,6,9,11,12,18,21,30,34,37

$$median = \frac{12+18}{2} = 15$$

2. Data set: 6, 5 , 9, 12, 30, 18, 11, 34, 21.

Ordered data set:

5,6,9,11,12,18,21,30,34

Median = 12

Measure of spread around the Mean

- Most useful measure of spread when working with random samples.
- The deviation of a value is how far apart is it from the mean.

  $x - \bar{x}$
- Unfortunately it is easy to see that
- $\sum (x - \bar{x}) = 0$

- Standard Deviation
- There are two kinds σ_n and σ_{n-1}.
- The default is σ_{n-1}.
- They are calculated as:

$$\sigma_n = \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$$

$$\sigma_{n-1} = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

Measure of spread around the Mean

- Example. Data: 2,7,8,12,12,19
- $n = 6,\ \bar{x} = (2+7+8+12+12+19)/6 = 10$

x	$x-\bar{x}$	$(x-\bar{x})^2$
2	-8	64
7	-3	9
8	-2	4
12	2	4
12	2	4
19	9	81
Sum	60	0	166

$$\sigma_n = \sqrt{\frac{\sum (x-\bar{x})^2}{n}}$$

$$\sigma_{n-1} = \sqrt{\frac{\sum (x-\bar{x})^2}{n-1}}$$

$$\sigma_n = \sqrt{\frac{166}{6}} \approx 5.2599$$

$$\sigma_{n-1} = \sqrt{\frac{166}{5}} \approx 5.7619$$

Measure of spread around the Median

- Q_1 = First Quartile or Lower Quartile.
- Q_3 = Third Quartile or Upper Quartile.
- These are calculated as the medians of each of the two halves determined by the original median.
- In case n is odd then the original median is removed from each of the two halves.

- Inter Quartile Range

IQR = The distance between the Lower Quartile and the Upper Quartile.

$$IQR = Q_3 - Q_1$$

- About 50% of the values are between Q_1 and Q_3.

Five Number Summary

- min = Minimum (value)
- Q_1 = Lower or First Quartile
- Q_2 = Median
- Q_3 = Upper or Third Quartile
- max = Maximum (value)

In addition we also have
- Range = $max - min$
- $IQR = Q_3 - Q_1$

Example: Mammal speeds,
11,12,20,25,30,30,30,32,35,39,40,40,40,42,45,48,50,70.

- min = 11
- Q_1 = 30
- Median = (35+39)/2 = 37
- Q_3 = 42
- max = 70.

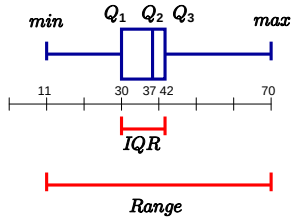
- Range = 70 – 11 = 59
- IQR = 42 – 30 = 12

Box Plots

- A **Box Plot** is a $graphical\ display$ of a five-point summary.

Example: Mammal speeds,
11,12,20,25,30,30,30,32,35,39,40,40,40,42,45,48,50,70.

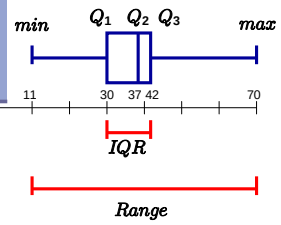
- min = 11
- Q_1 = 30
- Median = (35+39)/2 = 37
- Q_3 = 42
- max = 70.
- Range = 70 – 11 = 59
- IQR = 42 – 30 = 12



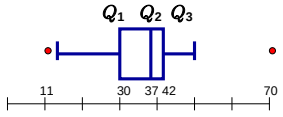
Modified Box Plots

- A **Modified Box Plot** also takes into account the **outliers**.
- An **outlier** is a value which is more than 1.5 times the IQR from the nearest quartile.

Example: Mammal speeds,
11,12,20,25,30,30,30,32,35,39,40,40,40,42,45,48,50,70.



- IQR = 42 – 30 = 12
- (1.5)IQR = (1.5)12 = 18
- 30 – 18 = 12 > 11, so 11 is an outlier.
- 42 + 18 = 60 <70, so 70 is an outlier.



Box Plots (Modified)

- Box Plots and Modified Box Plots are useful when plotting a single quantitative variable and:
  - We want to compare shape, center, and spread of two or more distributions.
  - The distribution has a large number of values
  - Individual values do not need to be identified.
  - (Modified) We want to identify outliers.