Math 140
Introductory Statistics

Professor B. Ábrego
Lecture 3
Sections 2.1, 2.2

People added to the class.

- Sara Nejad Hashemi
- Nazir Atayee
- Expo Aggie
- Mirna Chamorro
- Sean-Michael Schumacher
- Ruth Zepeda
- Kent Allison

Wait till the END of the class to ask me for a permission number.

- Next on the list
  - Ziyao Zhu

Quantitative vs. Categorical Data

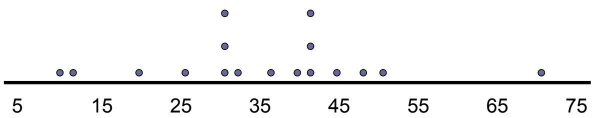
- **Quantitative**: Data about the cases in the form of numbers that can be compared and that can take a large number of values.
- **Categorical**: Data where a case either belongs to a category or not.

Different ways to visualize data

- Quantitative Variables
  - Dot Plots
  - Histograms
  - Stemplots
- Categorical Variables
  - Bar Graphs

Dot Plots

- Each dot represents the value associated to a case.
  - Dots may have different symbols.
  - Dots may represent more than one case.

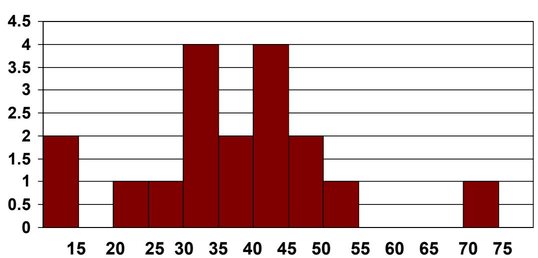


Dot Plots

- Dot Plots work best when
  - Relatively small number of values to plot
  - Want to keep track of individuals
  - Want to see the shape of the distribution
  - Have one group or a small number of groups that we want to compare

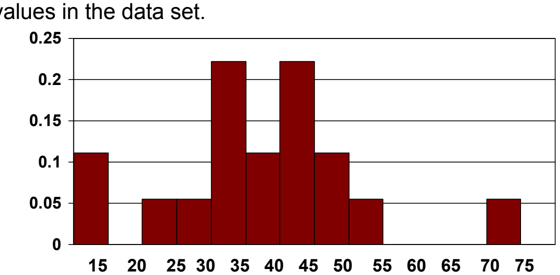
Histograms

- Groups of cases represented as rectangles or bars
- The vertical axis gives the number of cases (called **frequency** or **count**) for a given group of values.
- By convention borderline values go to the bar on the right.
- There is no prescribed number for the width of the bars.



Relative Frequency Histograms

- The height of each bar is the proportion of values in that range. (always a number between 0 and 1)
- The sum of the heights of all the bars equals 1.
- To change a regular histogram to a relative frequency histogram just divide the frequency of each bar by the total number of values in the data set.



Histograms (Relative Frequency)

- Histograms work best when
  - Large number of values to plot
  - Don't need to see individual values
  - Want to see the general shape of the distribution
  - Have one or a small number of distributions we want to compare
  - We can use a calculator or computer to draw the plots

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

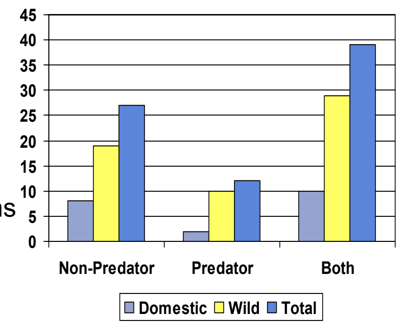
3 | 9 represents 39 miles per hour.

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.



2.2 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 $\bar{x}$.

- Calculated as:

Measures of Center

- Mean

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

- Calculated as:

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

Measures of Center

- Mean

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

- Calculated as:

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

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

Measures of Center

- Mean

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

- Calculated as:

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

- 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
- 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 $(a+b)/2$
- Example. Ordered data set:

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

Measures of Center

- **Median**
- 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 $(a+b)/2$
- Example. Ordered data set:

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

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

Measure of spread around the Median

- First Quartile or Lower Quartile. Denoted Q_1.
- Third Quartile or Upper Quartile. Denoted Q_3.
- 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

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

Measure of spread around the Median

- First Quartile or Lower Quartile. Denoted Q_1.
- Third Quartile or Upper Quartile. Denoted Q_3.
- 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

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

- $\boxed{IQR = Q_3 - Q_1}$
- About 50% of the values are between Q_1 and Q_3.

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.

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.

  $x-\overline{x}$
- Unfortunately it is easy to see that

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.

  $x-\overline{x}$
- Unfortunately it is easy to see that
- $\sum(x-\overline{x})=0$
- Standard Deviation
- There are two kinds σ_n and σ_{n-1}.
- The default is σ_{n-1}.
- They are calculated as:

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.

  $x-\overline{x}$
- Unfortunately it is easy to see that
- $\sum(x-\overline{x})=0$
- Standard Deviation
- There are two kinds σ_n and σ_{n-1}.
- The default is σ_{n-1}.
- They are calculated as:

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

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

Measure of spread around the Mean

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

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

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

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

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

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

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

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

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

Five Number Summary

- Minimum = smallest value = min
- Lower or First Quartile = Q_1.
- Median = Q_2.
- Upper or Third Quartile = Q_3
- Maximum = largest value = max

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

Five Number Summary

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

- 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 = 37
  - Q_3 = 42
  - max = 70.

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

Box Plots

- 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 = 37
  - Q_3 = 42
  - max = 70.

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

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

Box Plots

- 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 = 37
  - Q_3 = 42
  - max = 70.

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

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