Example: A study was conducted in 1982 to compare the abilities of men and women to perform the strenuous tasks required of a shipboard firefighter. The study reports the pulling force (in Newtons) that a firefighter was able to exert in pulling the starter cord of a P-250 water pump. The study also giver the weight and the sex of the firefighter.

Who:

- *Cases:*

What:

When:

Where:

Why:

How:

Variable:

- *Type:*
- *Units:*

Variable:

- *Type:*
- *Units:*

Variable:

- *Type:*
- *Labels:*

The new Brew Pub manufactures and distributes three types of beers. a sample of 450 beer drinkers was selected. Individuals were asked to state their preference, defined as their first choice.

Who:

What:

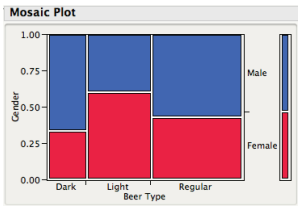
Compute marginal totals and percentages.

	Beer Type			
	Light	Regular	Dark	Total
Male	60	120	60	
Female	90	90	30	
Total				

Compute the conditional percentages for gender conditioning on beer type = "Light" (down the column).

	Beer Type			
	Light	Regular	Dark	Total
Male	60	120	60	240
Female	90	90	30	210
Total	150	210	90	450

Are these two variables independent?



Answer:

Chapter 4 Lecture Notes

Displaying and Summarizing Quantitative Data

Describing One Quantitative Variable

Histogram

1. Divide the *what* values into **equal-width** bins (piles).
2. Count the number of *whos* that belong to each bin.
3. Plot the bin values on the x-axis.
4. Plot the counts on the y-axis.

Example: Barry Bonds Home Runs

No. HR	16	25	24	19	33	25	34	46	37	33	42
Year	86	87	88	89	90	91	92	93	94	95	96
No. HR	40	37	34	49	73	46	45	45	5	26	
Year	97	98	99	00	01	02	03	04	05	06	

BBs HRs ordered: 5, 16, 19, 24, 25, 25, 26, 33, 33, 34, 34, 37, 37, 40, 42, 45, 45, 46, 46, 49, 73

Histogram-Barry Bonds

Bins	
0-10	
10-20	
20-30	
30-40	
40-50	
50-60	
60-70	
70-80	

Characteristics of a Stem and Leaf Display

Constructing a Stem-and-Leaf Display:

1. Order the values.
2. Separate each observation into a stem (all but the last digit) and a leaf (the last digit).
3. Write stems in a vertical column in increasing order from top to bottom (smallest at top, largest at bottom). Draw a vertical line to the right of this column.
4. Write each leaf in the row to the right of its stem, **in increasing order** out from the stem.
5. Provide key at bottom to decode plot.

Example of Stem and Leaf Display

BBs HRs ordered: 5, 16, 19, 24, 25, 25, 26, 33, 33, 34, 34, 37, 37, 40, 42, 45, 45, 46, 46, 49, 73

HAs HRs ordered: 10, 12, 13, 20, 24, 26, 27, 29, 30, 32, 34, 34, 38, 39, 39, 40, 40, 44, 44, 44, 44, 45, 47

"Back-2-Back" Stem and Leaf Display:

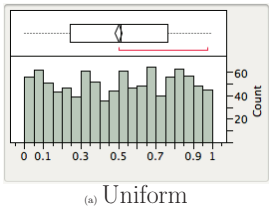
Barry Bonds		Hank Aaron
5	0	
9 6	1	0 2 3
6 5 5 4	2	0 4 6 7 9
7 7 4 4 3 3	3	0 2 4 4 8 9 9
9 6 6 5 5 2 0	4	0 0 4 4 4 4 5 7
	5	
	6	
3	7	

Interpreting Histograms and Stem and Leaf Displays

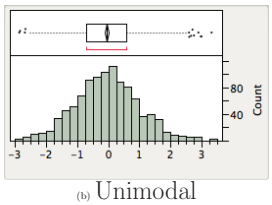
- we are interested in three properties
  1.
  2.
  3.

How many Modes?

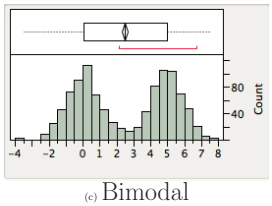
- None:
- One:
- Two:
- Three or more:



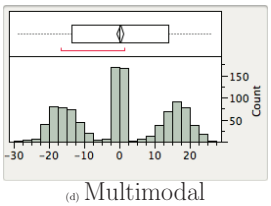
(a) Uniform



(b) Unimodal



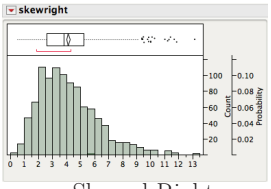
(c) Bimodal



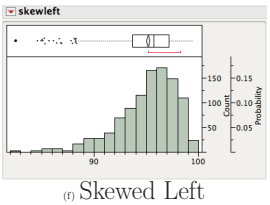
(d) Multimodal

Is the distribution symmetric?

- Symmetric:
- Skewed:



(e) Skewed Right

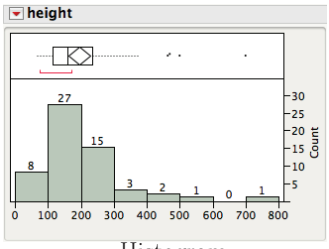


(f) Skewed Left

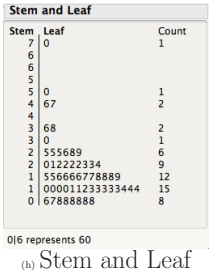
Are there any outliers?

Example identifying modes, skewness and outliers

Heights of Bungy Jumps in the USA



(g) Histogram



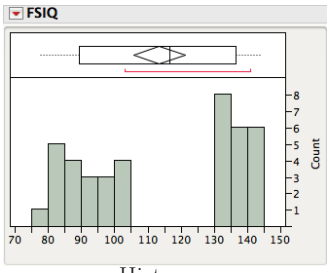
Stem and Leaf

Stem	Leaf	Count
7	0	1
6		
6		
5		
5	0	1
4	67	2
4		
3	68	2
3	0	1
2	555689	6
2	012222334	9
1	556666778889	12
1	000011233333444	15
0	67888888	8

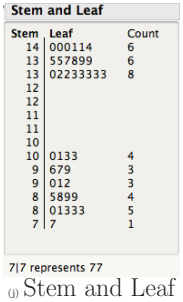
0|6 represents 60

(h) Stem and Leaf

IQ scores of men and women from a brain size study



(i) Histogram



Stem and Leaf

Stem	Leaf	Count
14	000114	6
13	557899	6
13	02233333	8
12		
12		
11		
11		
10		
10	0133	4
9	679	3
9	012	3
8	5899	4
8	01333	5
7	7	1

7|7 represents 77

(j) Stem and Leaf

Describing Distributions Numerically

- Measuring the Center

- Measuring the Spread

Median

Finding the Median:

1. Order the data from smallest to largest
2. Locate the middle number on the list
   - If the total number of observations (n) is odd: the median is the middle observation in the ordered list, i.e. the $\left(\frac{n+1}{2}\right)^{th}$ observation.
   - Ex: $n = 11$,
   - If the total number of observations (n) is even: the median is the average of the middle two observations in the ordered list, i.e. the average of the $\left(\frac{n}{2}\right)^{th}$ and $\left(\frac{n}{2}+1\right)^{th}$ observations.
   - Ex: $n = 36$,

Example: Barry Bonds HRs - (ordered):

5, 16, 19, 24, 25, 25, 26, 33, 33, 34, 34,
37, 37, 40, 42, 45, 45, 46, 46, 49, 73

n (sample size) = 21

median =

Example: New York Marathon Times for Men Between 20 and 29 in minutes

182, 201, 221, 234, 237, 251, 261, 266, 267, 273, 286,
291, 292, 296, 296, 296, 326, 352, 359, 365

n (sample size) = 20

median =

The Range

The problem with the range:

Quartiles: 3 numbers that divide the ordered data into 4 equally sized groups (i.e. each group contains 25% of data)

- Q_1:
- Q_2:
- Q_3:

Interquartile Range (IQR):

- $IQR =$

The **5-number summary** consists of

Min. Q_1 Median Q_3 Max.

Five-Number Summary for Barry Bonds Home-Runs

BBs HRs ordered:

5, 16, 19, 24, 25, 25, 26, 33, 33, 34, 34,
37, 37, 40, 42, 45, 45, 46, 46, 49, 73

Mean

Formula:

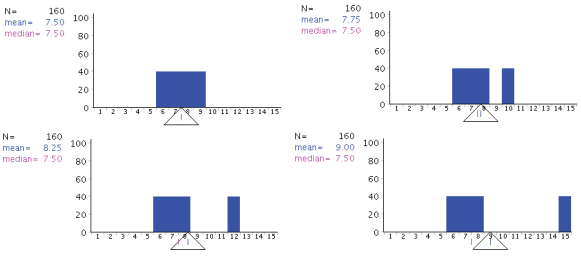
$$\bar{x} = \frac{x_1 + x_2 + x_3 + \cdots + x_n}{n} = \frac{\text{Total}}{n} = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$$

Example: Highway mileage for Honda Civics

x_i	x_1	x_2	x_3	x_4	x_5	x_6	x_7	x_8	x_9
	33	32	32	29	32	34	31	36	29

Mean:

Properties of the Mean



Standard Deviation

Formula:

$$s = \sqrt{\frac{(x_1 - \bar{x})^2 + (x_2 - \bar{x})^2 + \cdots + (x_n - \bar{x})^2}{n-1}}$$

$$= \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}}$$

Example: Calculating the standard deviation of Highway Mileage for Honda Civics

$$s = \sqrt{\frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1}} \qquad \bar{x} = 32$$

	H-mpg	$(x_i - \bar{x})$	$(x_i - \bar{x})^2$
x_1	33		
x_2	32		
x_3	32		
x_4	29		
x_5	32		
x_6	34		
x_7	31		
x_8	36		
x_9	29		
sum			

Influence of outlier on the mean and median

Small Example: Income in a small town of 6 people

$25,000 $27,000 $29,000
$35,000 $37,000 $38,000

- Median income is
- Mean income is

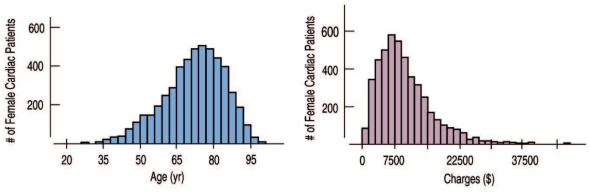
Tiger Woods moves to town.

$25,000 $27,000 $29,000
$35,000 $37,000 $38,000 $100,000,000

- The median income is
- The mean income is

Influence of Skewness on the mean and median

The observations in the "tail" of a distribution influence the mean but not the median.



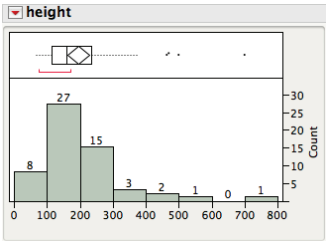
Report Range and IQR when you report Median Value.
Report Standard Deviation when you report Mean Value.
Always question when means are reported for skewed data

Which summaries are best?

- Five Number Summary
- Mean and Standard Deviation
- ALWAYS GET A PICTURE OF YOUR DATA.

Example choosing descriptive statistics

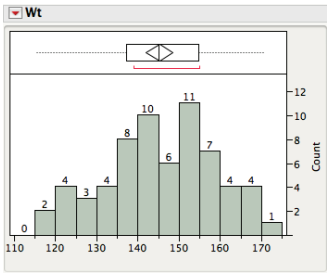
Heights of Bungy Jumps in the USA



- Which statistics would best describe the *center* and *variation* of the height (ft) of US bungy jumps as displayed in the histogram?

- How do we expect the mean and median to compare?

Weights of Sumo Wrestlers in Japan



- Which statistics would best describe tthe *center* and *variation* of the weights (kg) of sumo wrestlers as displayed in the histogram?

- How do we expect the mean and median to compare?