

# **STA 291**

# **Summer 2008**

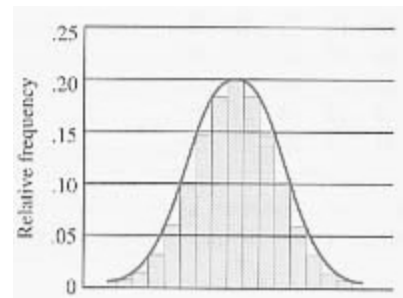
## **Lecture 3**

# Review: Graphical/Tabular Descriptive Statistics

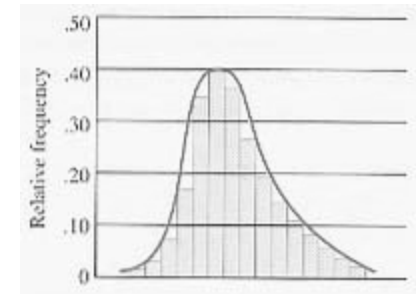
- Summarize data
- Condense the information from the dataset
- Always useful: Frequency distribution
- Interval data: Histogram
- Nominal/Ordinal data: Bar chart, Pie chart

# Review: Shapes of Distributions

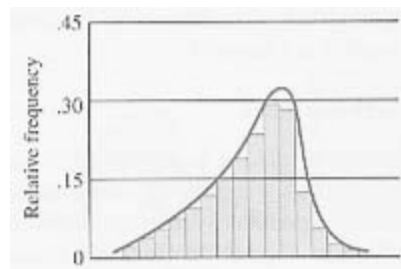
Symmetric Distribution



Skewed to the right

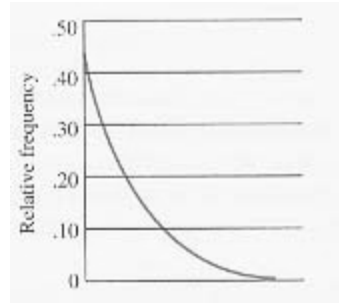


Skewed to the left

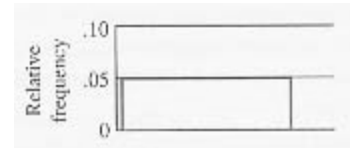


# Examples

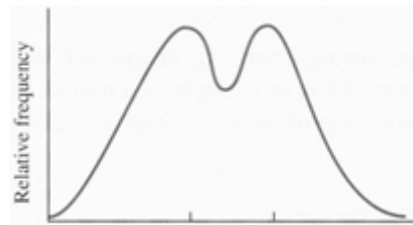
Exponential Distribution



Uniform Distribution



Bimodal Distribution



# Review: Good vs. Bad Graphics

- Good Graphics
  - ...present a lot of information clearly and concisely,
  - ...are not misleading
- Bad Graphics
  - ...are not clearly labeled or scaled
  - ...distort by stretching/shrinking the vertical or horizontal axis
  - ...use bar charts with bars of unequal width

# Summarizing Data Numerically

- Center of the data
  - Mean
  - Median
  - Mode
- Dispersion of the data
  - Variance, Standard deviation
  - Interquartile range
  - Range

# Population/Sample

- **Parameter**
  - numerical characteristic of the **p**opulation
  - calculated using the whole population
- **Statistic**
  - numerical characteristic of the **s**ample
  - calculated using the sample

# Measuring Central Tendency

- “What is a typical measurement in the sample/population?”
- Mean: Arithmetic average
- Median: Midpoint of the observations when they are arranged in increasing order
- Mode: Most frequent value

# Mean (Average)

- Mean (or Average): Sum of measurements divided by the number of subjects
- Example: Observations 3,8,19,12  
Mean =

# Mathematical Notation: Sample Mean

- Sample size  $n$
- Observations  $x_1, x_2, \dots, x_n$
- Sample Mean "x-bar"

$$\begin{aligned}\bar{x} &= (x_1 + x_2 + \dots + x_n) / n \\ &= \frac{1}{n} \sum_{i=1}^n x_i\end{aligned}$$

$\Sigma$  = SUM

# Mathematical Notation: Population Mean for a finite population of size $N$

- Population size  $N$
- Observations  $x_1, x_2, \dots, x_N$
- Population Mean “mu”

$$\mathbf{m} = (x_1 + x_2 + \dots + x_N) / N$$

$$= \frac{1}{N} \sum_{i=1}^N x_i$$

$\Sigma$  = SUM

# Mean (Average)

- The mean requires numerical values. Only appropriate for quantitative data.
- It does not make sense to compute the mean for nominal variables.
- Example “Nationality” (nominal):
  - Germany = 1, China = 2,
  - U.S. = 3, Norway = 4
- Mean nationality = 2.4???

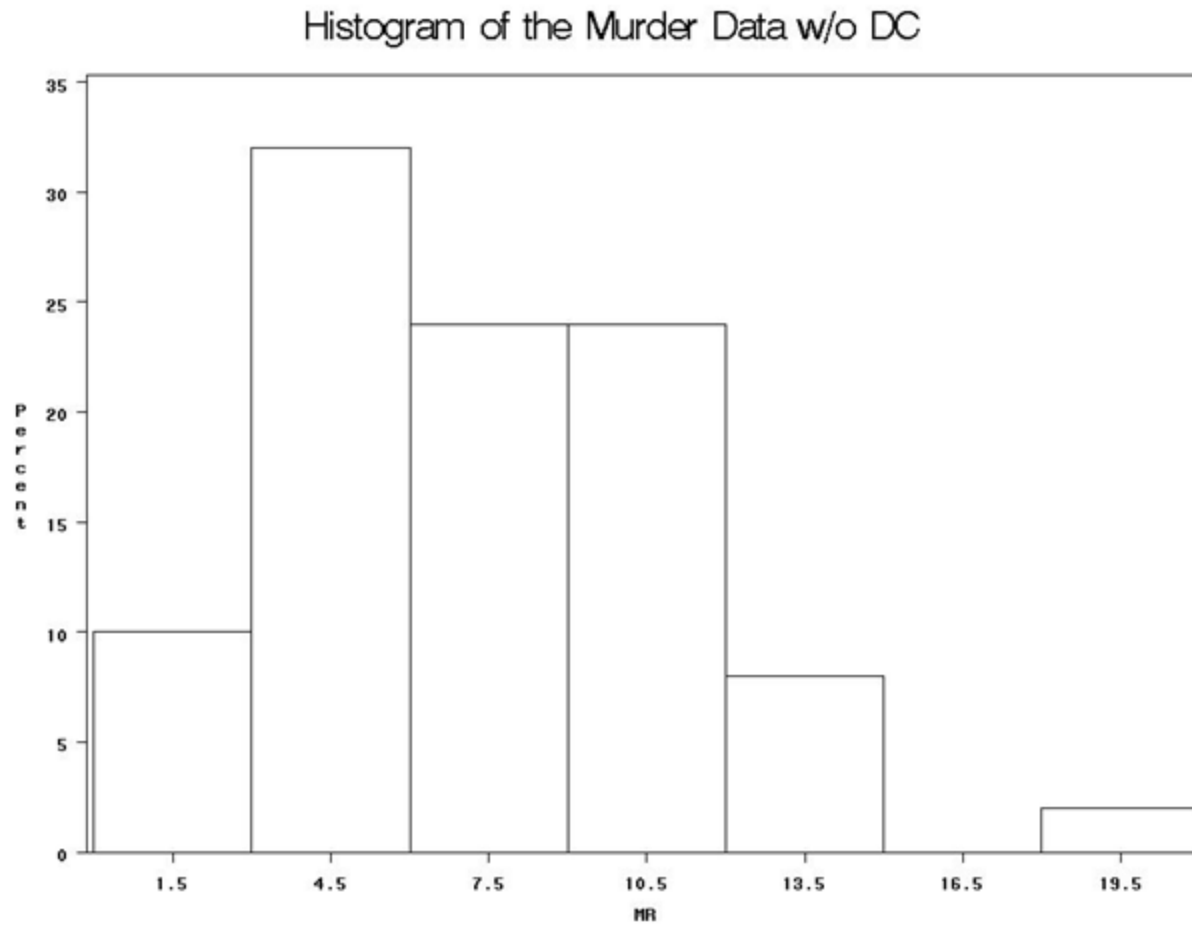
# Mean

- Sometimes, the mean is calculated for ordinal variables, but this does not always make sense.
- Example “Weather” (on an ordinal scale):  
Sun=1, Partly Cloudy=2, Cloudy=3,  
Rain=4, Thunderstorm=5
- Mean (average) weather=2.8
- Another example: “GPA = 3.8” is also a mean of observations measured on an ordinal scale

# Mean

- The mean is highly influenced by outliers. That is, data points that are far from the rest of the data.
- Example: Murder Rate Data  
Mean incl. DC: 8.73  
Mean w/o DC: 7.33
- Right skewed distribution:  
The mean is pulled to the right.

# Mean



# Mean

- If the distribution is highly skewed, then the mean is not representative of a typical observation

- Example:

Monthly income for five persons

1,000   2,000   3,000   4,000   100,000

- Average monthly income:
- Not representative of a typical observation.

# Geometric Interpretation of the Mean

- Assume that each measurement has the same “weight”
- Then, the mean is the center of gravity for the set of observations
- This is because the sum of the distances to the mean is the same for the observations above the mean as for the observations below the mean

# Median

- The median is the measurement that falls in the middle of the ordered sample
- When the sample size  $n$  is odd, there is a middle value
- It has the ordered index  $(n+1)/2$
- Example: 1.1, 2.3, 4.6, 7.9, 8.1  
 $n=5$ ,  $(n+1)/2=6/2=3$ , Index = 3,  
Median = 3<sup>rd</sup> smallest observation = 4.6

# Median

- When the sample size  $n$  is even, average the two middle values
- Example: 3, 7, 8, 9,  $n=4$ ,  
 $(n+1)/2=5/2=2.5$ , Index = 2.5

Median = midpoint between 2<sup>nd</sup> and 3<sup>rd</sup>  
smallest observation =  $(7+8)/2 = 7.5$

# Mean and Median

- For skewed distributions, the median is often a more appropriate measure of central tendency than the mean
- The median usually better describes a “typical value” when the sample distribution is highly skewed

- Example:

Monthly income for five persons

1,000   2,000   3,000   4,000   100,000

- Median monthly income: 3000

# Mean and Median

- Example: Murder Rate Data

- Mean incl. DC: 8.73

Mean w/o DC: 7.33

- Median incl. DC: 6.8

Median w/o DC: 6.7

# Summary: Measures of Location

**Mean-** Arithmetic Average

{ Mean of a Sample -  $\bar{x}$   
{ Mean of a Population -  $\mu$

**Median** – Midpoint of the observations when they are arranged in increasing order

Notation: Subscripted variables  
 $n$  = # of units in the sample  
 $N$  = # of units in the population  
 $x$  = Variable to be measured  
 $x_i$  = Measurement of the *ith* unit

**Mode-** Most frequent value.

# Mean and Median

- Is there a compromise between the median and the mean?
- Yes!
- Trimmed mean:
  1. Order the data from smallest to largest
  2. Delete a selected number of values from each end of the ordered list
  3. Find the mean of the remaining values
- The trimming percentage is the percentage of values that have been deleted from each end of the ordered list.

# Median for Grouped or Ordinal Data

- Example: Highest Degree Completed (Frequency Table)

Highest Degree	Frequency	Percentage
Not a high school graduate	38,012	21.4
High school only	65,291	36.8
Some college, no degree	33,191	18.7
Associate, Bachelor, Master, Doctorate, Professional	41,124	23.2
Total	177,618	100

# Calculate the Median

- $n=177,618$
- $(n+1)/2=88,809.5$
- Median = midpoint between the 88809<sup>th</sup> smallest and 88810<sup>th</sup> smallest observation
- Both are in the category “High school only”
- Mean would not make sense here, since the variable is only ordinal

# Median

- The median can be used for interval data and for ordinal data
- The median can not be used for nominal data because the observations can not be ordered on a scale

# Mean versus Median

- Mean: Interval data with an approximately symmetric distribution
- Median: Interval or ordinal data
- The mean is sensitive to outliers, the median is not

# Mean vs. Median

Observations	Median	Mean
1, 2, 3, 4, 5	3	3
1, 2, 3, 4, 100		
3, 3, 3, 3, 3		
1, 2, 3, 100, 100		

# Mean vs. Median

- If the distribution is symmetric, then Mean=Median
- If the distribution is skewed, then the mean lies more toward the direction of skew
- Mean and Median Online Applet

# Median

- Disadvantage: Insensitive to changes **within** the lower or upper half of the data
- Example: 1, 2, 3, 4, 5, 6, 7      vs.  
                  1, 2, 3, 4, 100, 100, 100
- *Sometimes*, the mean is more informative even when the distribution is skewed

# Example

- “Number of people you have known personally who have committed suicide in the last 12 months”

Response	Frequency	Percentage
0	1344	88.8
1	133	8.8
2	25	1.7
3	11	0.7
4	1	0.1

- $N=1514$ , index = 757.5
- Median = 0
- Mean = 0.145 (interval data)

# Example (contd.)

- Change data set (shift 586 responses from category 0 to 4)

Response	Frequency (old)	Frequency (new)
0	1344	758
1	133	133
2	25	25
3	11	11
4	1	587

- $N=1514$ , index = 757.5
- Median = 0
- Mean = 1.694

# Mode

- The mode is the value that occurs most frequently
- The mode need not be near the center of the distribution. Therefore, it is not really a measure of central tendency
- Can be used for **all** types of data (nominal, ordinal, interval)
- Bimodal distribution: Two peaks (only for ordinal or interval scale)
- Unimodal, symmetric distribution:  
Mean=Median=Mode

# Mean vs. Median vs. Mode

- Mean: Interval data with an approximately symmetric distribution
- Median: Interval or ordinal data
- Mode: All types of data

# Mean vs. Median vs. Mode

- The mean is sensitive to outliers, median and mode are not
- In general, the median is more appropriate for skewed data than the mean
- In some situations, the median may be too insensitive to changes in the data
- The mode may not be unique

# Mean and Median

- Example: For towns with population size 2500 to 4599 in the U.S. Northeast in 1994, the mean salary of chiefs of police was \$37,527, and the median was \$30,500.
- Does this suggest that the distribution of salary was skewed to the left, symmetric, or skewed to the right?

# Another Example

## (Mean, Median, Mode)

- “How often do you read the newspaper?”

Response	Frequency	Relative Frequency
every day	969	
a few times a week	452	
once a week	261	
less than once a week	196	
Never	76	
TOTAL		

- Identify the mode
- Identify the median response
- Calculate the mean, if possible

# Percentiles

- The  $p$ th percentile ( $L_p$ ) is a number such that  $p\%$  of the observations take values below it, and  $(100-p)\%$  take values above it
- 50<sup>th</sup> percentile = median
- 25<sup>th</sup> percentile = lower quartile
- 75<sup>th</sup> percentile = upper quartile

## *Two Step Procedure*

1. The **index** for  $L_p$  can be calculated as  
 **$(n+1)p/100$**
2.  $L_p$  is the  $(n+1)p/100$  th smallest observation

# Quartiles

- 25<sup>th</sup> percentile  
= lower quartile  
= approximately median of the observations below the median
- 75<sup>th</sup> percentile  
= upper quartile  
= approximately median of the observations above the median

- Median and Quartiles can be found from a stem and leaf plot
- Example: Murder Rate Data (w/o DC)

Stem	Leaf	#
20	3	1
19		
18		
17		
16		
15		
14		
13	135	3
12	7	1
11	334469	6
10	2234	4
9	08	2
8	03469	5
7	5	1
6	034689	6
5	0238	4
4	46	2
3	0144468999	10
2	039	3
1	67	2

A quarter of the states  
has murder rate above...

The median murder rate is...

A quarter of the states  
has murder rate below...

-----+-----+-----+-----+

# Five-Number Summary

- Maximum, Upper Quartile, Median, Lower Quartile, Minimum
- Statistical Software SAS output  
(Murder Rate Data)

<b>Quantile</b>	<b>Estimate</b>
<b>100% Max</b>	<b>20.30</b>
<b>75% Q3</b>	<b>10.30</b>
<b>50% Median</b>	<b>6.70</b>
<b>25% Q1</b>	<b>3.90</b>
<b>0% Min</b>	<b>1.60</b>

# Five-Number Summary

- Maximum, Upper Quartile, Median, Lower Quartile, Minimum
- Example: The five-number summary for a data set is minimum=4, Q1=256, median=530, Q3=1105, maximum=320,000.
- What does this suggest about the shape of the distribution?

# Measures of Variation

- Mean and Median only describe a typical value, but not the spread of the data
- Two distributions may have the same mean, but different variability
- Statistics that describe variability are called measures of variation

# Sample Measures of Variation

- Sample Range:

Difference between maximum and minimum sample value

- Sample Variance:  $s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$

- Sample Standard Deviation:

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

- Sample Interquartile Range:

Difference between upper and lower quartile of the sample

# Population Measures of Variation

- Population Range:

Difference between maximum and minimum population value

- Population Variance:  $s^2 = \frac{\sum (x_i - \mathbf{m})^2}{N}$

- Population Standard Deviation:  $s = \sqrt{\frac{\sum (x_i - \mathbf{m})^2}{N}}$

- Population Interquartile Range:

Difference between upper and lower quartile of the population

# Range

- Range: Difference between the largest and smallest observation
- Very much affected by outliers (one misrecorded observation may lead to an outlier, and affect the range)
- The range does not always reveal different variation about the mean

# Range: Example

- Murder Rate Data with DC:  
Smallest Observation: 1.6  
Largest Observation: 78.5  
Range:
- Murder Rate Data without DC:  
Smallest Observation: 1.6  
Largest Observation: 20.3  
Range:

# Deviations

- The deviation of the  $i$ th observation  $x_i$  from the sample mean  $\bar{x}$  is  $(x_i - \bar{x})$ . The difference between them
- The sum of all deviations is zero because the sample mean is the center of gravity of the data
- Therefore, people use either the sum of the absolute deviations or the sum of the squared deviations as a measure of variation

# Deviations: Example

- Data: 1, 7, 4, 3, 10
- Mean:  $25/5=5$

Observation	Deviation
1	
3	
4	
7	
10	

# Sample Variance

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

The variance of  $n$  observations is the sum of the squared deviations, divided by  $n-1$ .

# Variance: Example

Observation	Mean	Deviation	Squared Deviation
1			
3			
4			
7			
10			
Sum of the Squared Deviations			
$n-1$			
Sum of the Squared Deviations / $(n-1)$			

# Variance: Interpretation

- The variance is about the average of the squared deviations
- “average squared distance from the mean”
- Unit: square of the unit for the original data
- Difficult to interpret
- Solution: Take the square root of the variance, and the unit is the same as for the original data

# Standard Deviation

- The standard deviation  $s$  is the positive square root of the variance

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

# Standard Deviation: Properties

- $s \geq 0$  always
- $s=0$  only when all observations are the same
- If data is collected for the whole population instead of a sample, then  $n-1$  is replaced by  $n$
- $s$  is sensitive to outliers

# Standard Deviation

## Interpretation: Empirical Rule

- If the histogram of the data is approximately symmetric and bell-shaped, then
  - About **68%** of the data are within **one** standard deviation from the mean
  - About **95%** of the data are within **two** standard deviations from the mean
  - About **99.7%** of the data are within **three** standard deviations from the mean

# Another (Better) Example

- Distribution of SAT score is scaled to be approximately bell-shaped with mean 500 and standard deviation 100
- About 68% of the scores are between \_\_\_\_\_
- About 95% are between \_\_\_\_\_
- If you have a score above 700, you are in the top \_\_\_\_\_%

# Yet Another Example

- “Number of people you have known personally who have committed suicide in the last 12 months”

Response	Frequency	Percentage
0	1344	88.8
1	133	8.8
2	25	1.7
3	11	0.7
4	1	0.1

- Mean = 0.15
- Standard Deviation = 0.46
- Are 68% of the observations between  $-0.31$  and  $0.61$ ?

# Interquartile Range

- The Interquartile Range (IQR) is the difference between upper and lower quartile
- $IQR = Q3 - Q1$
- IQR= Range of values that contains the middle 50% of the data
- IQR increases as variability increases
- Example: Murder Rate Data  
Q1= 3.9, Q3= 10.3, IQR=\_\_\_\_\_

# Summary

- Measures of Location / Central Tendency
  - Where is the data located?
  - Where is the “middle” of the data?
  - Mean, Median, Mode
- Measures of Variation
  - How variable are the data?
  - How spread out about the “middle” are the data?
  - Range, Variance, Standard Deviation, Interquartile Range

# Sample Statistics and Population Parameters

- Population mean and population standard deviation are denoted by the Greek letters  $\mu$  (mu) and  $\sigma$  (sigma)
- They are unknown constants that we would like to estimate
- Sample mean and sample standard deviation are denoted by  $\bar{x}$  and  $s$
- They are random variables, because their values vary according to the random sample that has been selected

# Sample Measures of Variation

- Sample Range:

Difference between maximum and minimum sample value

- Sample Variance:  $s^2 = \frac{\sum (x_i - \bar{x})^2}{n - 1}$

- Sample Standard Deviation:

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

- Sample Interquartile Range:

Difference between upper and lower quartile of the sample

# Population Measures of Variation

- Population Range:

Difference between maximum and minimum population value

- Population Variance:  $s^2 = \frac{\sum (x_i - \mathbf{m})^2}{N}$

- Population Standard Deviation:  $s = \sqrt{\frac{\sum (x_i - \mathbf{m})^2}{N}}$

- Population Interquartile Range:

Difference between upper and lower quartile of the population

# Review: Range

- Range: Difference between the largest and smallest observation
- Very much affected by outliers (one misrecorded observation may lead to an outlier, and affect the range)
- The range does not always reveal different variation about the mean

# Sample Variance and Standard Deviation

The sample variance of  $n$  observations is the sum of the squared deviations, divided by  $n-1$ .

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

The sample standard deviation  $s$  is the positive square root of the sample variance.

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

# Sample Variance: Example

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

- The variance of  $n$  observations is the sum of the squared deviations, divided by  $n-1$ .
- Calculate the variance and standard deviation for the following data set:  
281, 289, 289, 290, 291

# Variance and Standard Deviation: Example

Observation	Mean	Deviation	Squared Deviation
Sum of the Squared Deviations			
$n-1$			
Variance = Sum of the Squared Deviations / $(n-1)$			
Standard Deviation = Square Root of the Variance			

# Variance: Interpretation

- The variance is about the average of the squared deviations
- “average squared distance from the mean”
- Unit: square of the unit for the original data
- Difficult to interpret
- Solution: Take the square root of the variance, and the unit is the same as for the original data

# Standard Deviation: Properties

- $s \geq 0$  always
- $s=0$  only when all observations are the same
- If data is collected for the whole population instead of a sample, then  $n-1$  is replaced by  $n$
- $s$  is sensitive to outliers
- All these properties also apply to the variance

# Standard Deviation

## Interpretation: Empirical Rule

- If the histogram of the data is approximately symmetric and bell-shaped, then
  - About **68%** of the data are within **one standard deviation** from the mean
  - About **95%** of the data are within **two** standard deviations from the mean
  - About **99.7%** of the data are within **three** standard deviations from the mean

# Another (Better) Example

- Distribution of SAT score is scaled to be approximately bell-shaped with mean 500 and standard deviation 100
- About 68% of the scores are between \_\_\_\_\_
- About 95% are between \_\_\_\_\_
- If you have a score above 700, you are in the top \_\_\_\_\_%

# Yet Another Example

- “Number of people you have known personally who have committed suicide in the last 12 months”

Response	Frequency	Percentage
0	1344	88.8
1	133	8.8
2	25	1.7
3	11	0.7
4	1	0.1

- Mean = 0.15
- Standard Deviation = 0.46
- Are 68% of the observations between  $-0.31$  and  $0.61$ ?

# Interquartile Range

- The Interquartile Range (IQR) is the difference between upper and lower quartile
- $IQR = Q3 - Q1$
- IQR= Range of values that contains the middle 50% of the data
- IQR increases as variability increases
- IQR is robust (not much affected by outliers)
- Example: Murder Rate Data  
Q1= 3.9, Q3= 10.3, IQR=\_\_\_\_\_

# Summary

- Measures of Location / Central Tendency
  - Where is the data located?
  - Where is the “middle” of the data?
  - Mean, Median, Mode
- Measures of Variation
  - How variable are the data?
  - How spread out about the “middle” are the data?
  - Range, Variance, Standard Deviation, Interquartile Range

# Sample Statistics and Population Parameters

- Population mean and population standard deviation are denoted by the Greek letters  $\mu$  (mu) and  $\sigma$  (sigma)
- They are unknown constants that we would like to estimate
- Sample mean and sample standard deviation are denoted by  $\bar{x}$  and  $s$
- They are random variables, because their values vary according to the random sample that has been selected

# Problem 1

- According to the National Association of Home Builders, the U.S. nationwide median selling price of homes sold in 1995 was \$118,000
- Would you expect the mean to be larger, smaller, or equal to \$118,000?
- Which of the following is the most plausible value for the standard deviation:  
(a)  $-15,000$ , (b)  $1,000$ , (c)  $45,000$ , (d)  $1,000,000$ ?

# Problem 2

- For the following multiple-choice item, select the correct response(s).
  - In Canada in 1981, for the categories Catholic, Protestant, Eastern Orthodox, Jewish, None, Other for religious affiliation, the relative frequencies were 47.3%, 41.2%, 1.5%, 1.2%, 7.3%, 1.5%
- (a) The median religion is Protestant
  - (b) The distribution is bimodal
  - (c) Only 2.7% of the subjects fall within one standard deviation of the mean
  - (d) The mode is Catholic
  - (e) The "Other" response is an outlier

# Box Plots

- A box plot is basically a graphical version of the five number summary (unless there are outliers)
- It consists of a **box** that contains the central 50% of the distribution (from lower quartile to upper quartile),
- A **line** within the box that marks the median,
- And **whiskers** that extend to the maximum and minimum values, unless there are outliers

# Outlier

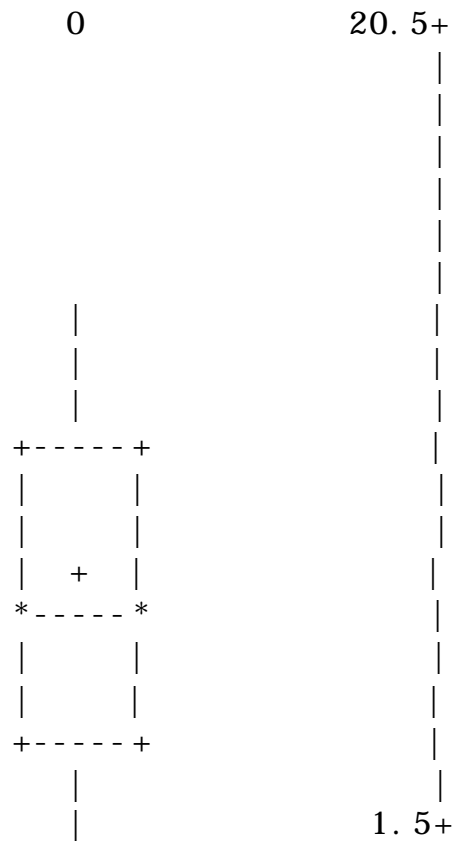
- An observation is an outlier if it falls
  - more than 1.5 IQR above the upper quartile or
  - more than 1.5 IQR below the lower quartile
- Example: Murder Rate Data w/o DC
  - upper quartile  $Q3 = 10.3$
  - IQR = 6.4
  - $Q3 + 1.5 \text{ IQR} = \underline{\hspace{2cm}}$
  - Any outliers?

# Box Plots (contd.)

- The ***whiskers*** only extend to the most extreme observations within 1.5 IQR beyond the quartiles
- If an observation is an ***outlier***, it is marked by an ***O*** (in SAS)
- If an observation is an ***extreme outlier*** (more than 3.0 IQR beyond the quartiles), it is marked by an asterisk ***\****
- In the box plot, the mean is represented by a ***+***

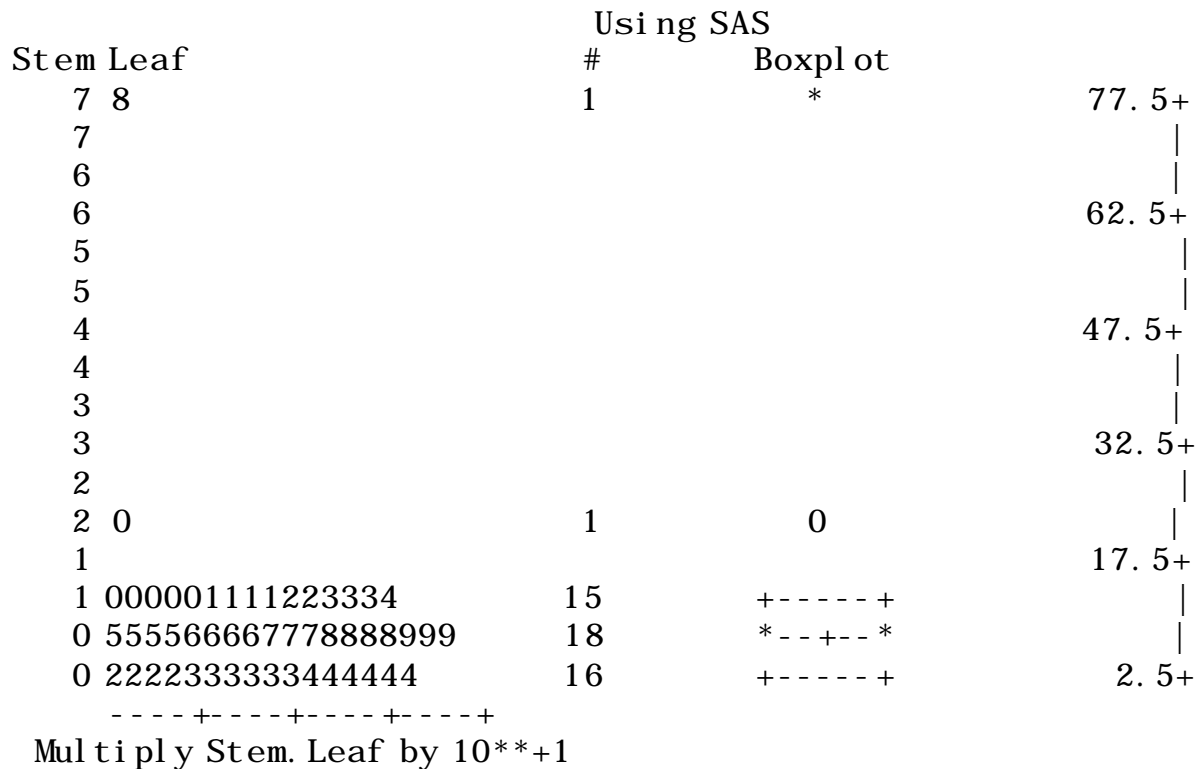
# Box Plot: Example

Boxplot for the Murder Rate Data w/o DC  
Using SAS



# Stem and Leaf Plot and Box Plot: Example

Stem and Leaf Plot and Boxplot  
for the Murder Rate Data with DC



# Box Plot: Another Example

- Observations:  
148, 154, 158, 160, 161,  
162, 166, 170, 182, 195, 236
- Create a box plot.

# Example Data Sets

- One Variable Statistical Calculator
- Modify the data sets and see how mean and median, as well as standard deviation and interquartile range change
- Look at the histograms and stem-and-leaf plots – does the empirical rule apply?
- Make yourself familiar with the standard deviation
- Interpreting the standard deviation takes experience

# Analyzing Linear Relationships Between Two Quantitative Variables

- Is there an association between the two variables?
- Positive or negative?
- How strong is the association?
- Notation
  - Response variable:  $Y$
  - Explanatory variable:  $X$

# Sample Measures of Linear Relationship

- Sample Covariance:

$$s_{xy} = \frac{\sum (x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{1}{n-1} \left( \sum x_i y_i - \frac{1}{n} \sum x_i \sum y_i \right)$$

- Sample Correlation Coefficient:

$$r = \frac{s_{xy}}{s_x s_y}$$

- Population measures: Divide by N instead of n-1

# Properties of the Correlation I

- The value of  $r$  does not depend on the units (e.g., changing from inches to centimeters), whereas the covariance does
- $r$  is standardized
- $r$  is always between  $-1$  and  $1$ , whereas the covariance can take *any* number
- $r$  measures the ***strength and direction of the linear association*** between  $X$  and  $Y$
- $r > 0$  positive linear association
- $r < 0$  negative linear association

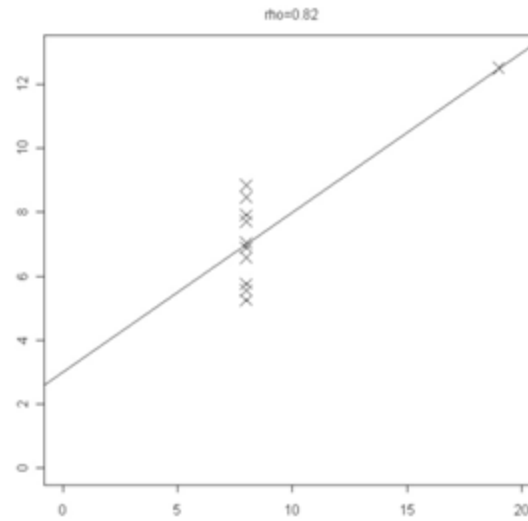
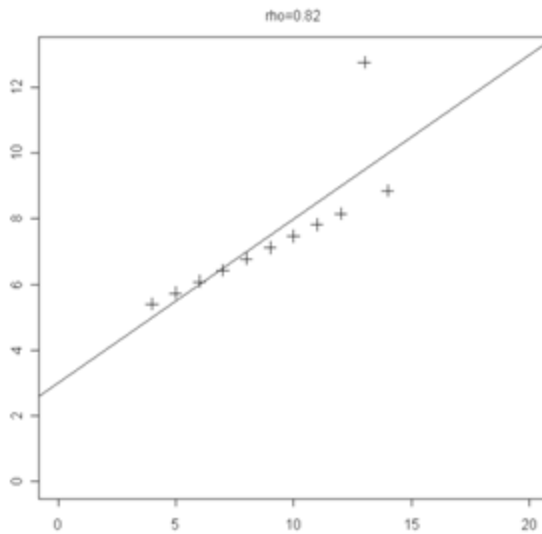
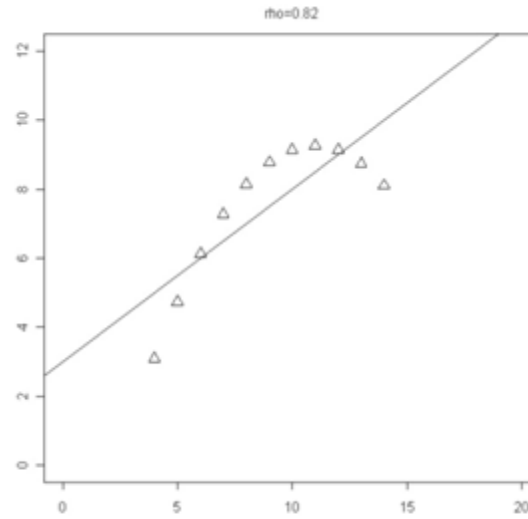
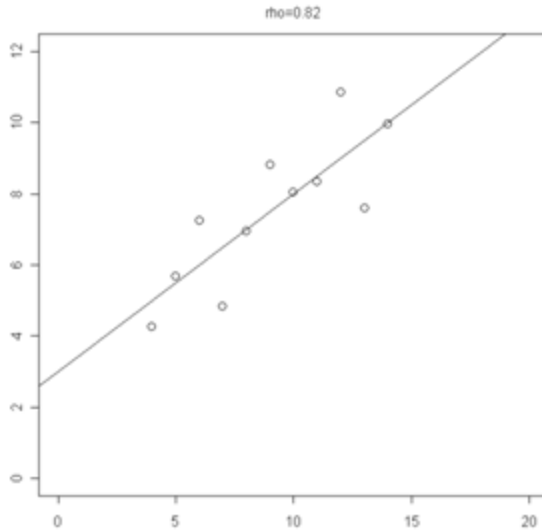
# Properties of the Correlation II

- $r = 1$  when all sample points fall exactly on a line with positive slope (*perfect positive association*)
- $r = -1$  when all sample points fall exactly on a line with negative slope (*perfect negative association*)
- The larger the absolute value of  $r$ , the stronger is the degree of linear association

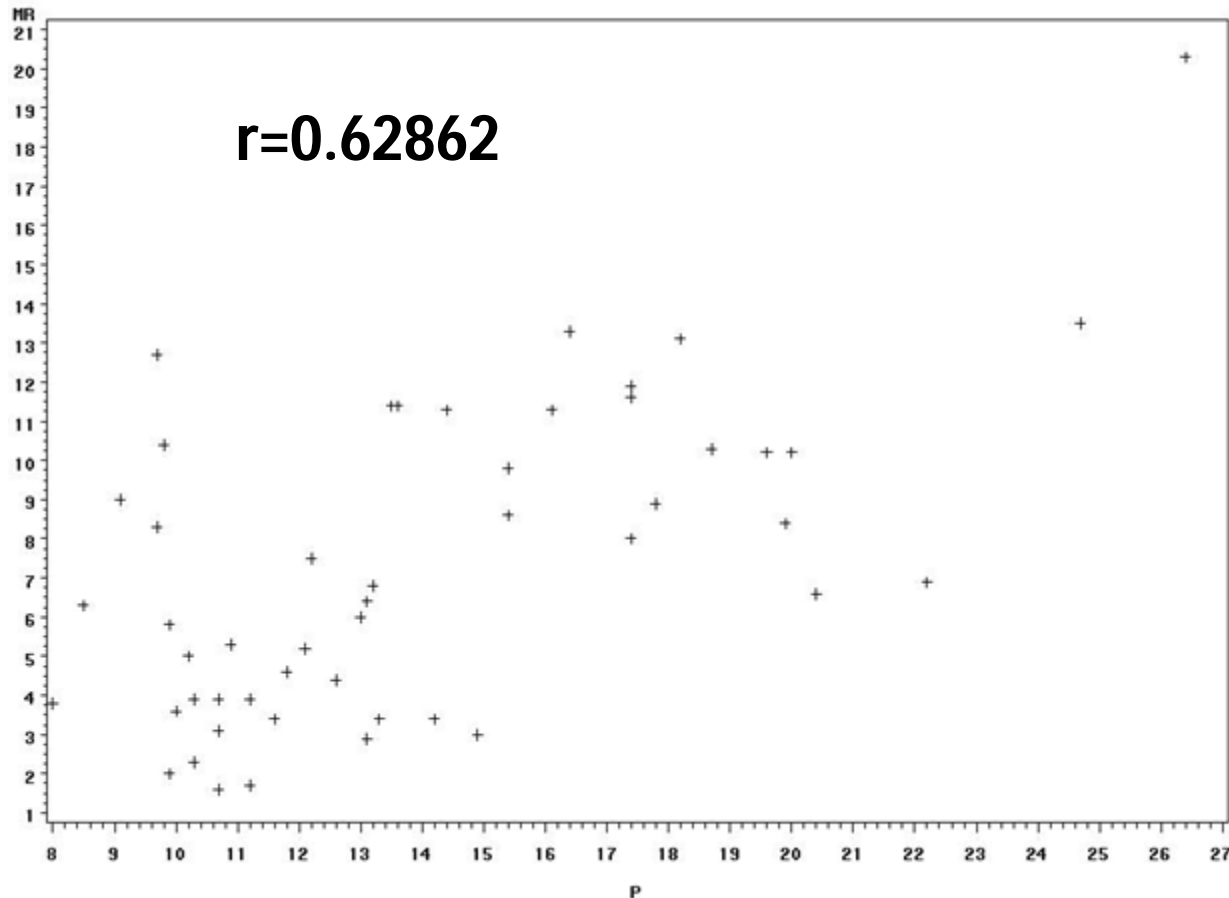
# Properties of the Correlation III

- If  $r$  is close to 0, this does not necessarily mean that the variables are not associated
- It only means that they are not *linearly associated*
- The correlation treats  $X$  and  $Y$  symmetrically
- That is, it does not matter which variable is explanatory ( $X$ ) and which one is response ( $Y$ ), the correlation remains the same

# Example: Correlation = 0.82



# Scatter Diagram of Murder Rate (Y) and Poverty Rate (X) for the 50 States



Correlation and Scatterplot Applet

Correlation by Eye Applet

Simple Regression Analysis Tool

# Model Assumptions and Violations

- **Factors Influencing the Correlation**
- The sample correlation depends on the range of  $X$ -value sampled
- When a sample has a much narrower range of variation in  $X$  than the population, the sample correlation tends to underestimate the population correlation
- The sample  $(X,Y)$  values should be a random sample of the population
- It should be representative of the  $X$  population values as well as the  $Y$  values

# Correlation: Example

- For a sample of 100 people, the correlation coefficient between  $X =$  years of education and  $Y =$  annual income (in dollars) equals 0.50
  - a) Suppose instead  $Y$  refers to annual income in thousands of dollars. State the correlation.
  - b) Suppose that  $Y$  is treated as the explanatory variable and  $X$  is treated as the response variable. Will the correlation coefficient change in value?