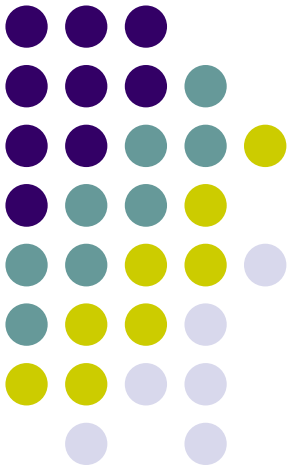
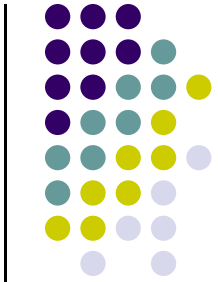
Econ 3790: Business and Economics Statistics

Instructor: Yogesh Uppal
Email: yuppal@ysu.edu

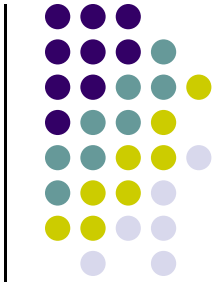


Today's Lecture



- Numerical methods for summarizing data:
  - Location
  - Variability
  - Distribution measures

Measures of Location

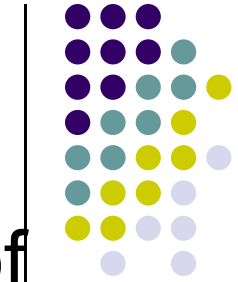


- Mean
- Median
- Mode
- Percentiles
- Quartiles

If the measures are computed for data from a sample, they are called sample statistics.

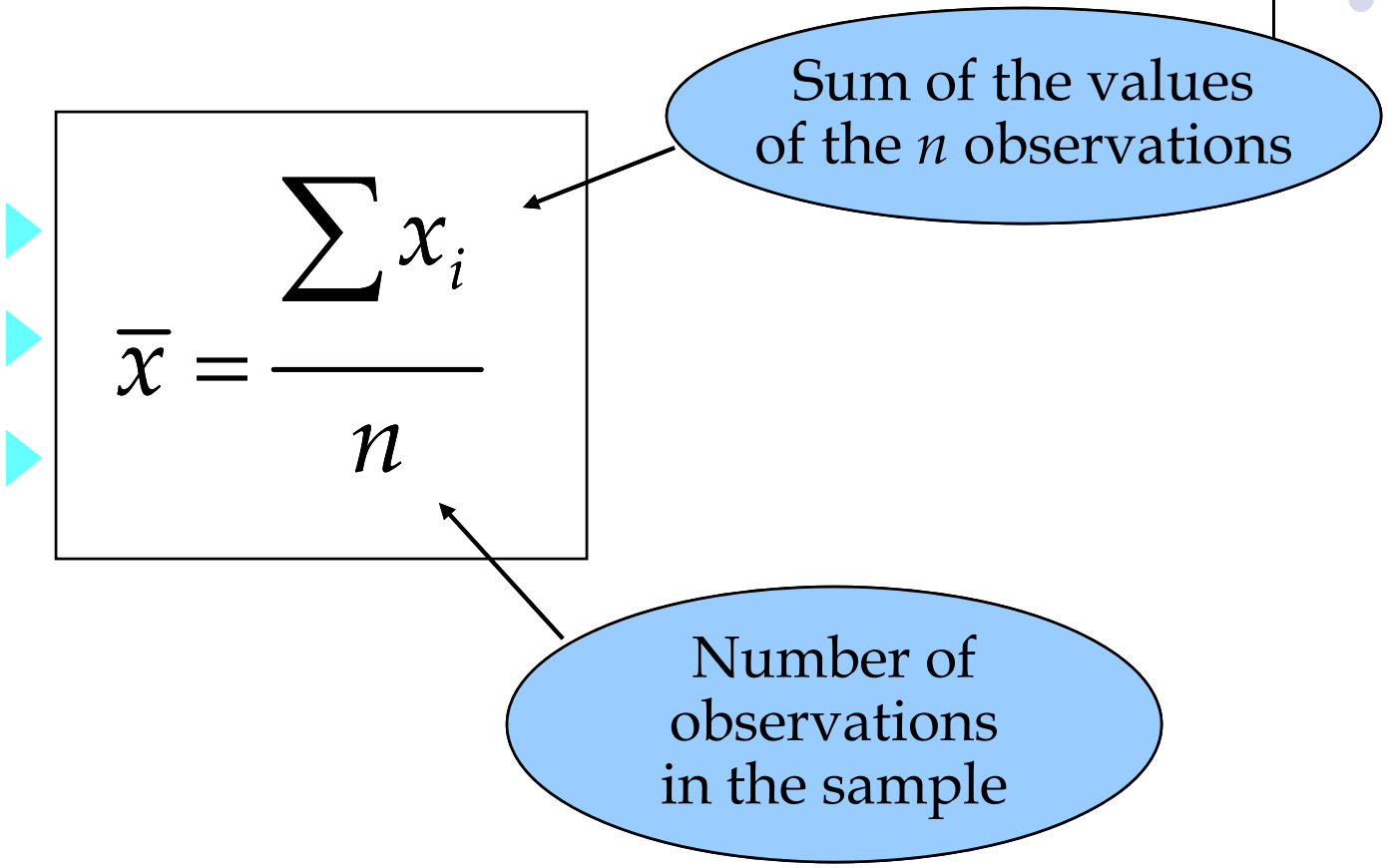
If the measures are computed for data from a population, they are called population parameters.

Mean

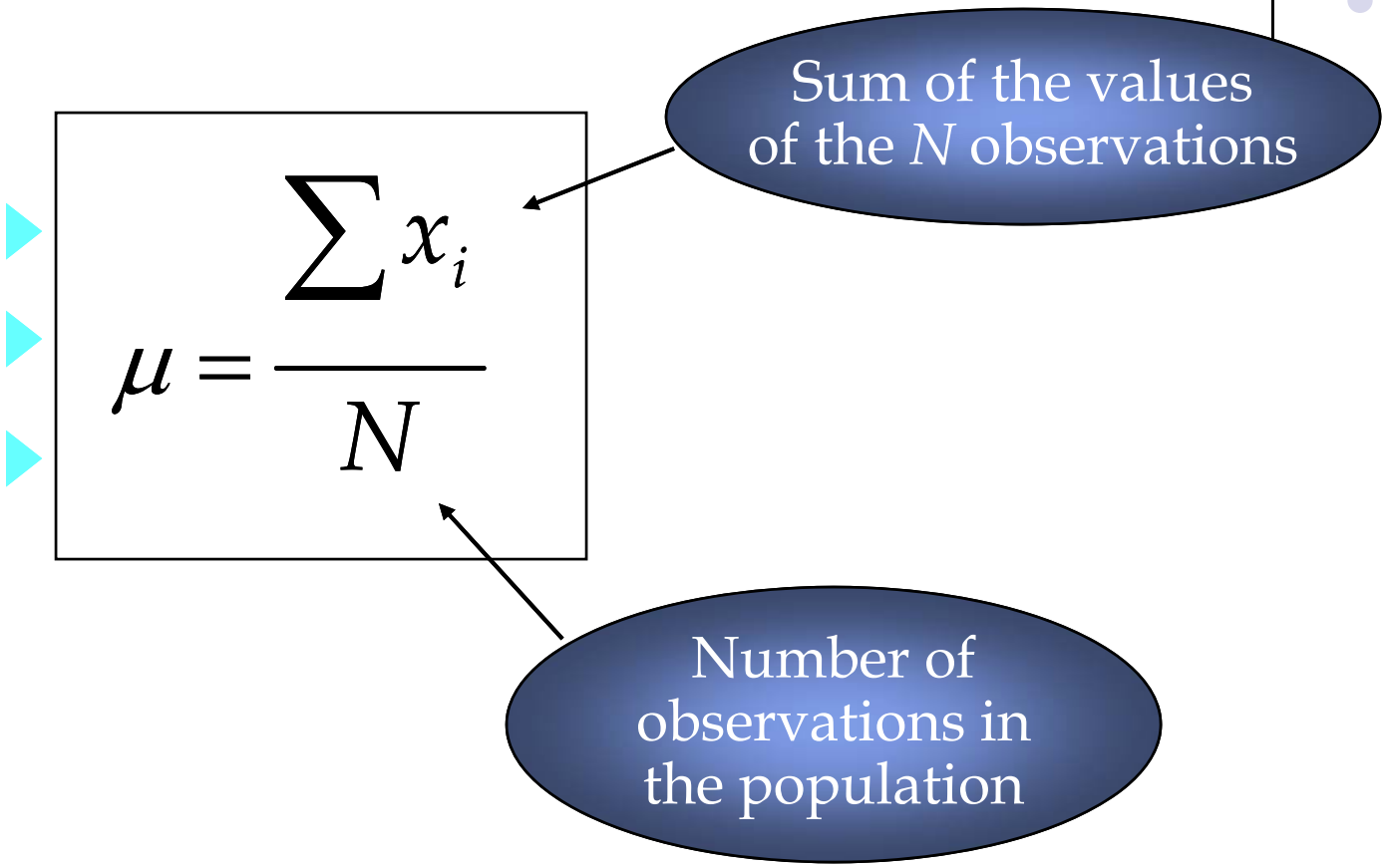


- The mean of a data set is the average of all the data values.
- The sample mean $\bar{x}$ is the point estimator of the population mean μ.

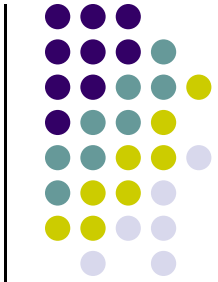
Sample Mean ($\overline{x}$)



Population Mean μ

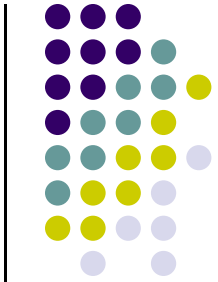


Go Penguins, Again!!!



Month	Opponents	Rushing TDs
Sep	SLIPPERY ROCK	4
Sep	NORTHEASTERN	4
Sep	at Liberty	1
Sep	at Pittsburgh	0
Oct	ILLINOIS STATE	1
Oct	at Indiana State	4
Oct	WESTERN ILLINOIS	2
Oct	MISSOURI STATE	4
Oct	at Northern Iowa	0
Nov	at Southern Illinois	0
Nov	WESTERN KENTUCKY	3

Sample Mean of TDs



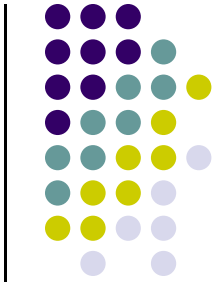
$$\bar{x} = \frac{\sum x_i}{n} = \frac{23}{11} = 2.09$$

Go Penguins……

Rushing TDs	Frequency	f_i.*x_i
0	3	0*3=0
1	2	1*2=2
2	1	2*1=2
3	1	3*1=3
4	4	4*4=16
	Total=11	Total=23

$$\bar{x} = \frac{\sum f_i x_i}{n} = \frac{23}{11} = 2.09$$

Sample Mean for Grouped Data



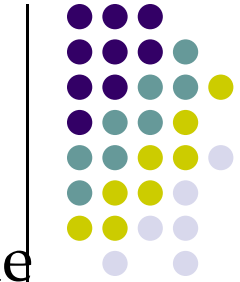
$$\bar{x} = \frac{\sum f_i M_i}{n}$$

Where M_i = the mid-point for class i

f_i = the frequency for class i

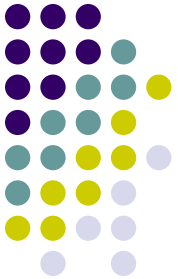
n = the sample size

Median



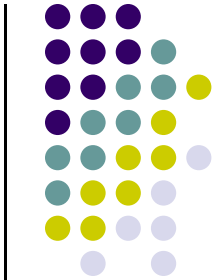
- The median of a data set is the value in the middle when the data items are arranged in ascending order.
- Whenever a data set has extreme values, the median is the preferred measure of central location.
- The median is the measure of location most often reported for annual income and property value data.
- A few extremely large incomes or property values can inflate the mean.

Median with odd number of Obs.



Month	Opponents	Rushing TDs
Sep	at Pittsburgh	0
Oct	at Northern Iowa	0
Nov	at Southern Illinois	0
Sep	at Liberty	1
Oct	Illinois State	1
Oct	Western Illinois	2
Nov	Western Kentucky	3
Sep	Slippery Rock	4
Sep	Northeastern	4
Oct	at Indiana State	4
Oct	Missouri State	4

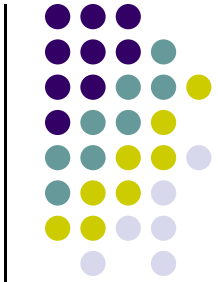
Median



- So the middle value is the game against Western Illinois in October.

- The median number of TDs is 2.

Median with even number of Obs.



- For an even number of observations:

26	18	27	12	14	27	30	19

8 observations

12	14	18	19	26	27	27	30

in ascending order

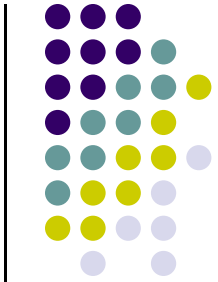
the median is the average of the middle two values.

Median = (19 + 26)/2 = 22.5

Mode

- The mode of a data set is the value that occurs with greatest frequency.
- The greatest frequency can occur at two or more different values.
- If the data have exactly two modes, the data are bimodal.
- If the data have more than two modes, the data are multimodal.

Mode

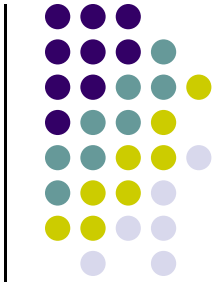


- Modal value for our Go Penguins example is 4 TDs.

Percentiles

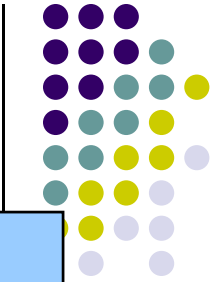
- A percentile provides information about how the data are spread over the interval from the smallest value to the largest value.
- Admission test scores for colleges and universities are frequently reported in terms of percentiles.

Percentiles



- The *p*th percentile of a data set is a value such that at least *p* percent of the items take on this value or less and at least (100 - *p*) percent of the items take on this value or more.

Percentiles

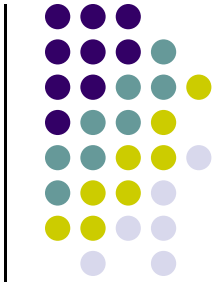


- Arrange the data in ascending order.
- Compute index i, the position of the pth percentile.

$$i = (p/100)n$$

- If i is not an integer, round up to the next integer. The pth percentile is the value in the ith position.
- If i is an integer, the pth percentile is the average of the values in positions i and $i+1$.

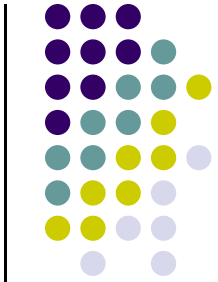
Example: Rental Market in Youngstown Again



- Sample of 28 rental listings from craigslist:

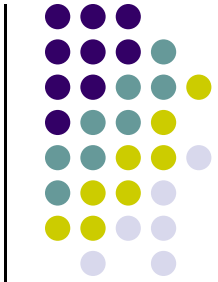
330	470	540	595
350	470	550	599
390	480	570	610
410	490	580	650
410	500	585	660
430	520	590	700
440	520	595	740

90^{th} Percentile



- $i = (p/100)*n = (90/100)*28 = 25.2$
- Rounding it to the next integer, which is the 26^{th} position
- 90th Percentile = 660

50^{th} Percentile

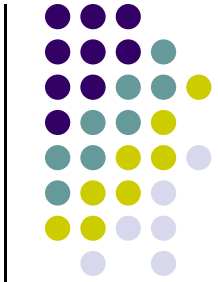


- $i = (p/100)n = (50/100)28 = 14$

- Averaging the 14th and 15th data values:

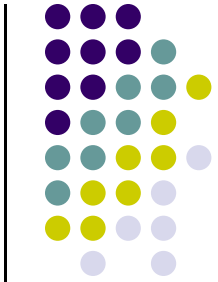
- 50th Percentile = (520 + 540)/2 = 530

Percentile Rank



- The percentile rank of a data value of a variable is the percentage of all elements with values less than or equal to that data value.
- Of course it is related to p^{th} percentile we found earlier.
- If the p^{th} percentile of a dataset is some value (Let us say 400), then the percentile rank of 400 is p%.

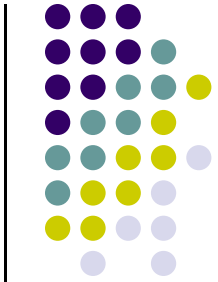
Percentile Rank (Cont'd)



- Percentile Rank can be calculated as follows:

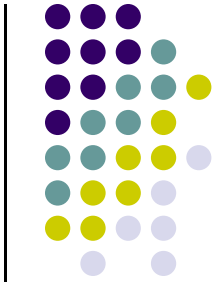
PR of a score =
(Cumulative Fre. of that score / Total Fre.)*100
Or
Cumulative Relative frequency * 100

Example 1: Go Penguins (Cont'd)



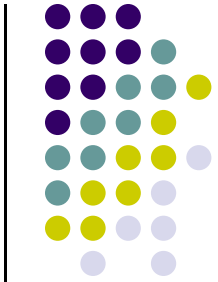
Rushing TDs	Cumulative Fre.	Cumulative Relative Fre.	PR
0	3	0.27	27%
1	5	0.45	45%
2	6	0.54	54%
3	7	0.63	63%
4	11	1	100%
Total			

Quartiles



- Quartiles are specific percentiles.
- First Quartile = 25th Percentile
- Second Quartile = 50th Percentile = Median
- Third Quartile = 75th Percentile

First Quartile



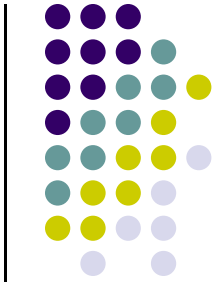
- First quartile = 25th percentile

$i = (p/100)n = (25/100)28 = 7$

- Averaging 7th and 8th data values

First quartile = (440+470)/2 = 455

Third Quartile



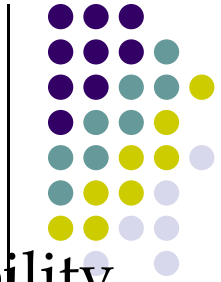
- Third quartile = 75th percentile

$i = (p/100)n = (75/100)28 = 21$

- Averaging 21st and 22nd data values

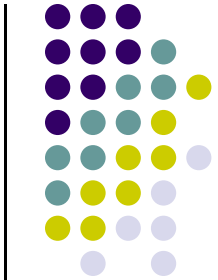
Third quartile = (595+595)/2 = 595

Measures of Variability



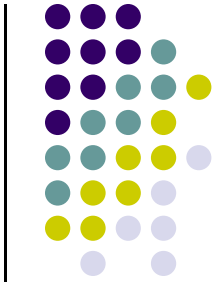
- It is often desirable to consider measures of variability (dispersion), as well as measures of location.
- For example, in choosing supplier A or supplier B we might consider not only the average delivery time for each, but also the variability in delivery time for each.

Measures of Variability



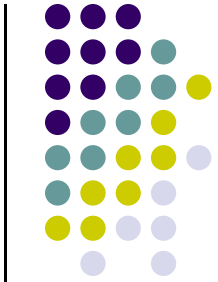
- Range
- Interquartile Range
- Variance
- Standard Deviation
- Coefficient of Variation

Range



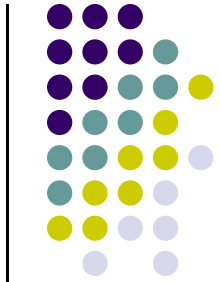
- The range of a data set is the difference between the largest and smallest data values.
- It is the simplest measure of variability.
- It is very sensitive to the smallest and largest data values.

Consider our penguins TDs data



Month	Opponents	Rushing TDs
Sep	at Pittsburgh	0
Oct	at Northern Iowa	0
Nov	at Southern Illinois	0
Sep	at Liberty	1
Oct	Illinois State	1
Oct	Western Illinois	2
Nov	Western Kentucky	3
Sep	Slippery Rock	4
Sep	Northeastern	4
Oct	at Indiana State	4
Oct	Missouri State	4

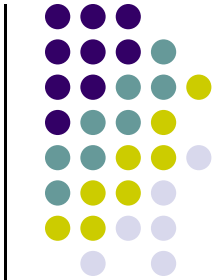
Range



- Range = largest value - smallest value

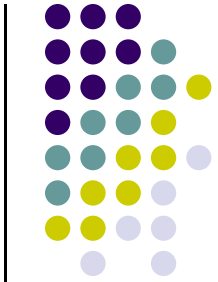
Range = 4 - 0 = 4

Interquartile Range



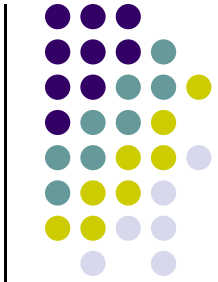
- The interquartile range of a data set is the difference between the third quartile and the first quartile.
- It is the range for the middle 50% of the data.
- It overcomes the sensitivity to extreme data values.

Interquartile Range



- 3rd Quartile ($Q3$) = (75/100)*11=8.25
  3rd Quartile is 4
- 1st Quartile ($Q1$) = 0
- Interquartile Range = $Q3$ - $Q1$ = 4 - 0 = 4

Variance

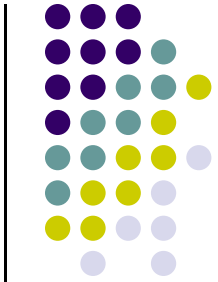


- The variance is a measure of variability that utilizes all the data.

- It is based on the difference between the value of each observation (x_i) and the mean ($\overline{x}$ for a sample, μ for a population).

- Basically we are talking about how the data is SPREAD around the mean.

Variance



- The variance is the average of the squared differences between each data value and the mean.
- The variance is computed as follows:

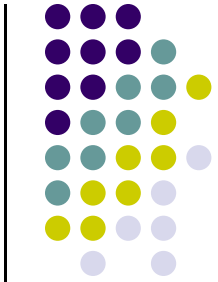
$$s^2 = \frac{\sum (x_i - \bar{x})^2}{n-1}$$

for a sample

$$\sigma^2 = \frac{\sum (x_i - \mu)^2}{N}$$

for a population

When data is presented as a frequency Table

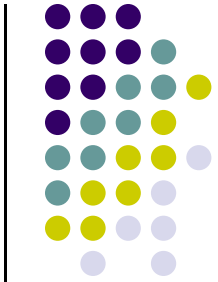


- Then, the variance is computed as follows:

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

$$\sigma^2 = \frac{\sum f_i (x_i - \mu)^2}{N}$$

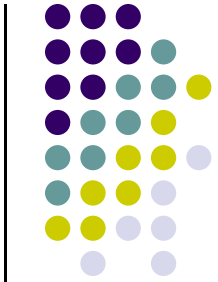
Standard Deviation



- The standard deviation of a data set is the positive square root of the variance.

- It is measured in the same units as the data, making it more easily interpreted than the variance

Standard Deviation



- The standard deviation is computed as follows:

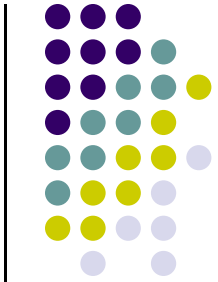
$$s = \sqrt{s^2}$$

for a sample

$$\sigma = \sqrt{\sigma^2}$$

for a population

Coefficient of Variation



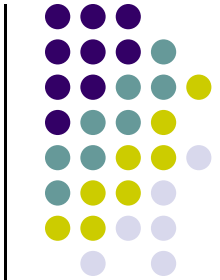
- The coefficient of variation indicates how large the standard deviation is in relation to the mean.

- The coefficient of variation is computed as follows:

$$\left(\frac{s}{\bar{x}} \times 100\right)\% \quad \leftarrow \text{for a sample}$$

$$\left(\frac{\sigma}{\mu} \times 100\right)\% \quad \leftarrow \text{for a population}$$

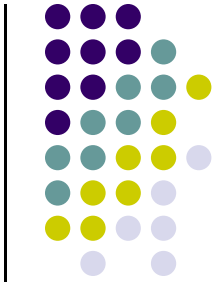
Coefficient of Variation (CV)



- CV is used in comparing variability of distributions with different means.

- A value of CV > 100% implies a data with high variance. A value of CV < 100% implies a data with low variance.

Covariance

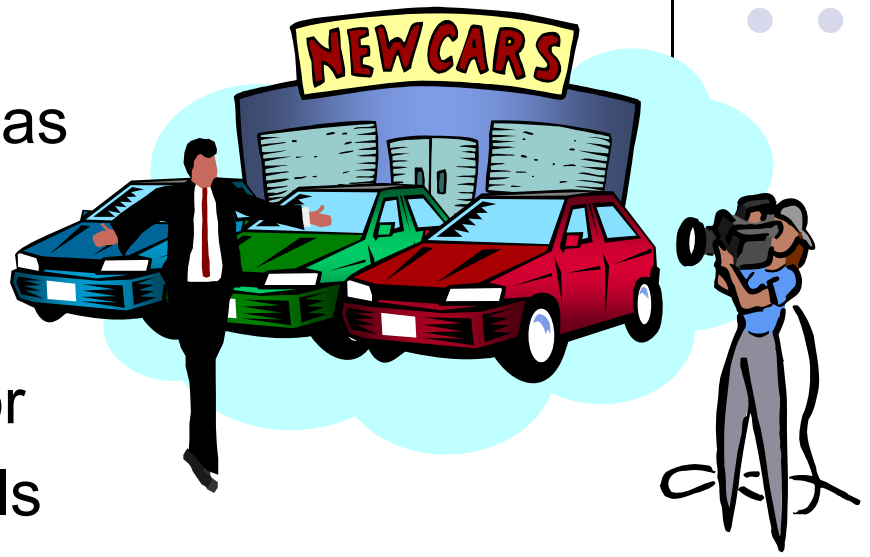


- Covariance between x and y is a measure of relationship between x and y.

$$\operatorname{cov}(x, y) = \frac{SS_{xy}}{n-1} = \frac{\sum (y-\bar{y})(x-\bar{x})}{n-1}$$

Covariance

- Example: Reed Auto Sales

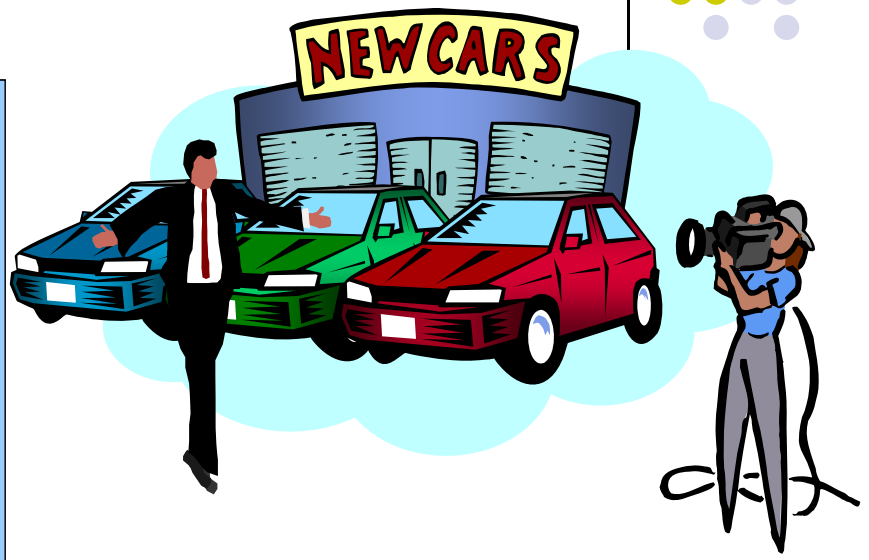


Reed Auto periodically has a special week-long sale. As part of the advertising campaign Reed runs one or more television commercials during the weekend preceding the sale. Data from a sample of 5 previous sales are shown on the next slide.

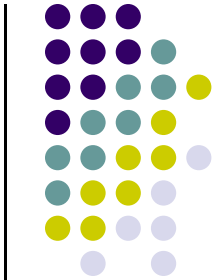
Covariance

- Example: Reed Auto Sales

Number of TV Ads	Number of Cars Sold
1	14
3	24
2	18
1	17
3	27



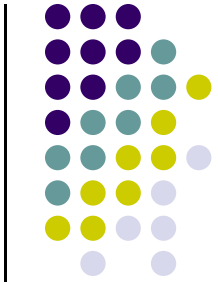
Covariance



x	y	$x-\bar{x}$	$y-\bar{y}$	$(x-\bar{x})(y-\bar{y})$
1	14	-1	-5.6	5.6
3	24	1	4.4	4.4
2	18	0	-1.6	0
1	17	-1	-2.6	2.6
3	25	1	5.4	5.4
Total=10	Total = 98			SS_{xy}=18

$$\mathrm{cov}(x,y)=\frac{SS_{xy}}{n-1}=\frac{18}{4}=4.5$$

Simple Correlation Coefficient



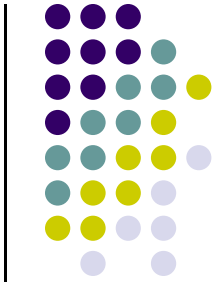
- **Simple Population Correlation Coefficient**

$$\rho = \frac{\text{cov}(x, y)}{\sigma_x \sigma_y}$$

$$-1 \leq \rho \leq +1$$

- **If ρ < 0, a negative relationship between x and y.**
- **If ρ > 0, a positive relationship between x and y.**

Simple Correlation Coefficient

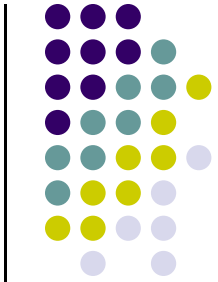


- Since population standard deviations of x and y are not known, we use their sample estimates to compute an estimate of ρ.

$$r = \frac{\text{cov}(x, y)}{s_x s_y}$$

$$-1 \leq r \leq +1$$

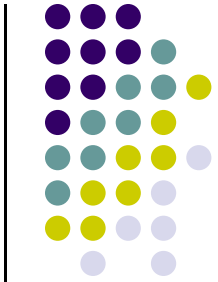
Simple Correlation Coefficient



- Example: Reed Auto Sales

x	y	$x-\bar{x}$	$y-\bar{y}$	SS_x	SS_y
1	14	-1	-5.6	1	31.36
3	24	1	4.4	1	19.36
2	18	0	-1.6	0	2.56
1	17	-1	-2.6	1	6.76
3	25	1	5.4	1	29.16
Total=10	Total=98			Total=4	Total= 89.2

Simple Correlation Coefficient

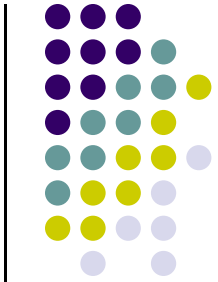


$$s_x = \sqrt{\frac{\sum (x - \bar{x})^2}{n-1}} = \sqrt{\frac{4}{4}} = 1$$

$$s_y = \sqrt{\frac{\sum (y - \bar{y})^2}{n-1}} = \sqrt{\frac{89.2}{4}} = 4.72$$

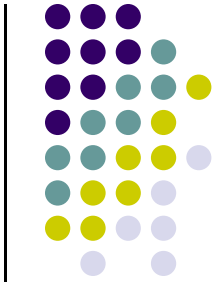
$$r = \frac{\mathrm{cov}(x, y)}{s_x s_y} = \frac{4.5}{1 * 4.72} = 0.95$$

Measures of Distribution Shape, Relative Location, and Detecting Outliers



- Distribution Shape
- z-Scores
- Detecting Outliers

Distribution Shape: Skewness

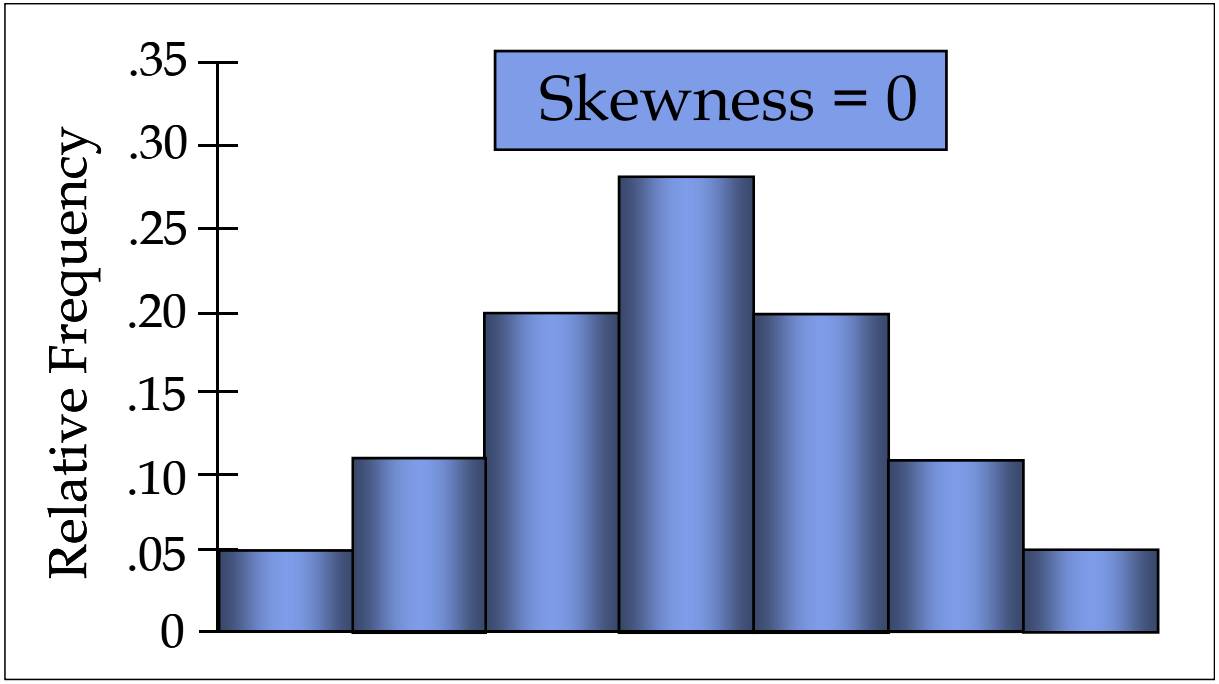


- An important measure of the shape of a distribution is called skewness.

- The formula for computing skewness for a data set is somewhat complex.

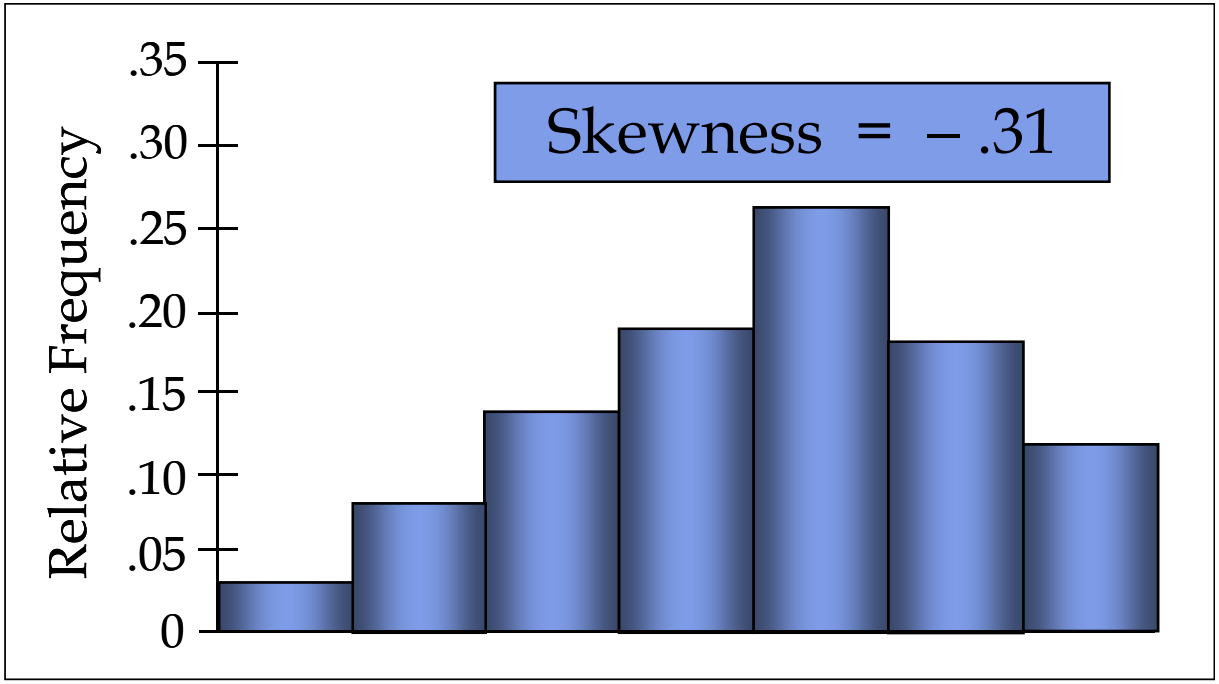
Distribution Shape: Skewness

- Symmetric (not skewed)
  - Skewness is zero.
  - Mean and median are equal.



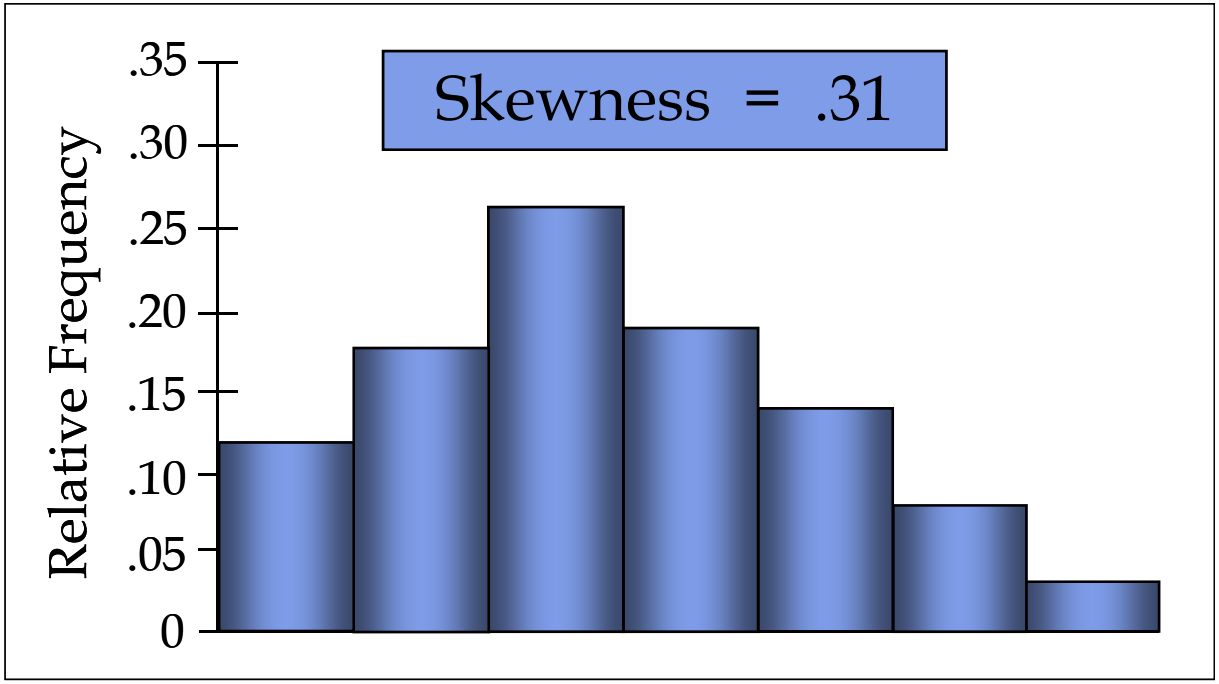
Distribution Shape: Skewness

- Moderately Skewed Left
  - Skewness is negative.
  - Mean will usually be less than the median.



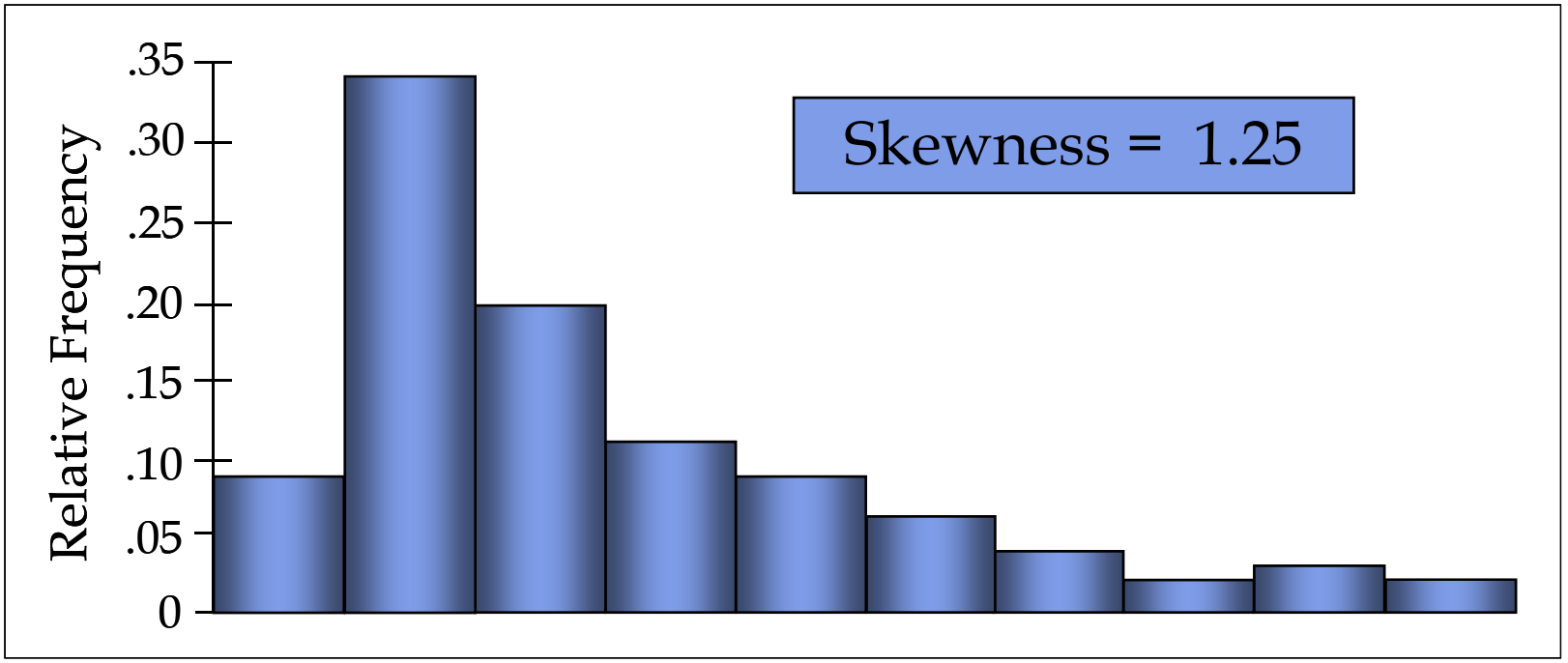
Distribution Shape: Skewness

- Moderately Skewed Right
  - Skewness is positive.
  - Mean will usually be more than the median.

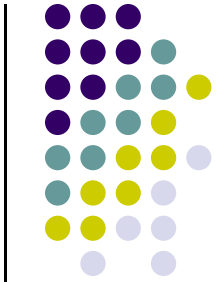


Distribution Shape: Skewness

- Highly Skewed Right
  - Skewness is positive.
  - Mean will usually be more than the median.



Z-scores



- Z-score is often called standardized scores.

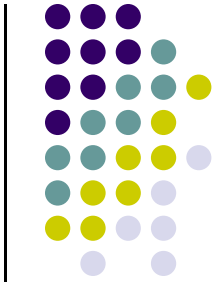
- It denotes the number of standard deviations a data value is from the mean.

$$z_i = \frac{x_i - \overline{x}}{s}$$

z-Scores

- An observation's z-score is a measure of the relative location of the observation in a data set.
- A data value less than the sample mean will have a z-score less than zero.
- A data value greater than the sample mean will have a z-score greater than zero.
- A data value equal to the sample mean will have a z-score of zero.

Detecting Outliers



- An outlier is an unusually small or unusually large value in a data set.

- A data value with a z-score less than -3 or greater than +3 might be considered an outlier.