Descriptive Statistics

Carlos Hurtado
churtado@richmond.edu

Robins School of Business
University of Richmond

Aug 28, 2019

On the Agenda

1. Measures of Location
2. Measures of Variability
3. Measures of Distribution Shape
4. Measures of Association Between Two Variables

On the Agenda

Numerical Measures

- Sample statistics: Measures computed for data from a sample
- Population parameters: Measures computed for data from a population
- Point estimator: Sample statistic of the population parameter

Measures of Location

1. Mean
2. Median
3. Mode
4. Weighted Mean
5. Minimum and Maximum
6. Percentile

Measures of Location: Mean

- Perhaps the most important measure of location is the mean

- The mean provides a measure of central location

- 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

Measures of Location: Sample Mean $\bar{x}$

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n}$$

- $\sum_{i=1}^{n}$ represents the sum of the x_i values of the n observations
- n is the number of observations in the sample

Measures of Location: Population Mean μ

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

- $\sum_{i=1}^{n}$ represents the sum of the x_i values of the N observations
- N is the number of observations in the population

Example: Monthly Starting Salary

- A placement office wants to know the average starting salary of business graduates.

Graduate	Monthly Starting Salary ($)	Graduate	Monthly Starting Salary ($)
1	3,850	7	3,890
2	3,950	8	4,130
3	4,050	9	3,940
4	3,880	10	4,325
5	3,755	11	3,920
6	3,710	12	3,880

Example: Monthly Starting Salary

$$\bar{x} = \frac{\sum_{i=1}^{n} x_i}{n} = \frac{\sum_{i=1}^{12} x_i}{12} = \frac{47,280}{12} = 3,940$$

- $\sum_{i=1}^{n}$ represents the sum of the x_i values of the $n = 12$ observations

Measures of Location: Median

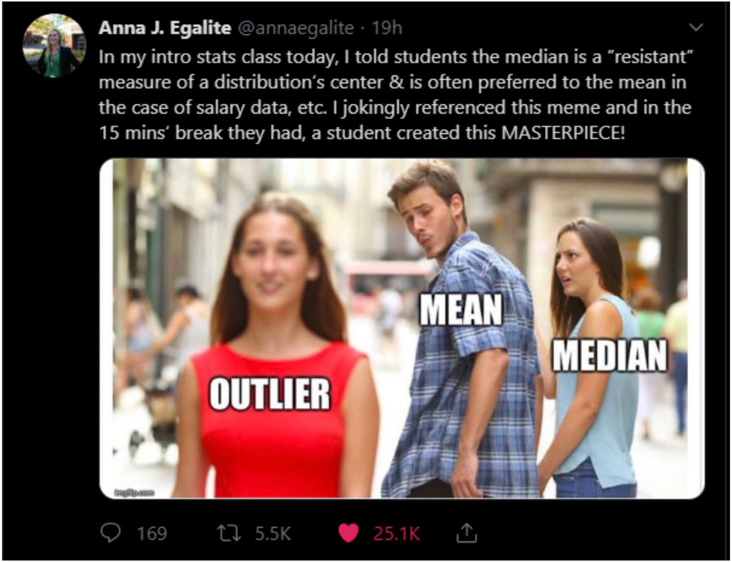
- Median: The middle observation or the average of the middle pair when the observations are ordered
- Whenever a data set has extreme values, median is the preferred measure of central location
- The median is the measure of location most often reported for annual income and property value data
- A few extremely large incomes or property values can inflate the mean

Measures of Location: Median

- Resistant Measure: is a measurement that doesn't change (or changes a tiny amount) when outliers are present

- The median is a resistant measure of a distribution's center

Measures of Location: Median



Example: Median for an odd number of observations

- 7 observations

26	18	27	12	14	27	19

Example: Median for an odd number of observations

- 7 observations In ascending order

12	14	18	19	26	27	27

Median is the middle value. Median $=19$

Example: Median for an even number of observations

- Monthly Starting Salary:

Monthly Starting Salary ($)	Monthly Starting Salary ($)
3,710	3,755
3,850	3,880
3,880	3,890
3,920	3,940
3,950	4,050
4,130	4,325

Averaging the 6th and 7th data values:
Median $= (3{,}890 + 3{,}920)/2 = 3{,}905$

Note: Data in ascending order

Measures of Location: Mode

- Mode: 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*

Example: Mode of Monthly Starting Salary

- Monthly Starting Salary:

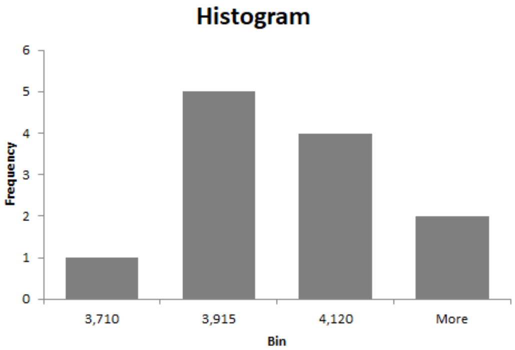
Monthly Starting Salary ($)	Monthly Starting Salary ($)
3,710	3,755
3,850	3,880
3,880	3,890
3,920	3,940
3,950	4,050
4,130	4,325

The only monthly starting salary that occurs more than once is 3,880

$$\text{Mode} = 3{,}880$$

Note: Data in ascending order

Example: Monthly Starting Salary



Measures of Location: Weighted Mean

- In some instances the mean is computed by giving each observation a weight that reflects its relative importance
- The choice of weights depends on the application
- The weights might be the number of credit hours earned for each grade, as in GPA
- In other weighted mean computations, quantities such as pounds, dollars, or volume are frequently used

Measures of Location: Weighted Mean

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i}$$

- $\sum_{i=1}^{n}$ represents the sum
- x_i is the value of observation i
- w_i is the weight for observation i
- Numerator: sum of the weighted data values
- Denominator: sum of the weights
- If data is from a population, μ replaces $\bar{x}$

Example: Purchase of Raw Material

- Consider the following sample of five purchases of a raw material over a period of three months

Purchase	Cost per Pound ($)	Number of Pounds
1	3.00	1200
2	3.4	500
3	2.8	2750
4	2.9	1000
5	3.25	800

$$\bar{x} = \frac{\sum_{i=1}^{n} w_i x_i}{\sum_{i=1}^{n} w_i} = \frac{18,500}{6,250} = 2.96$$

FYI, equally-weighted (simple) mean $= \$3.07$

Measures of Relative Standing: Minimum and Maximum

- Minimum: Smallest value in the observations

  0% of the observations are smaller than the minimum

- Maximum: Largest value in the observations

  100% of the observations are smaller than the maximum

Measures of Relative Standing: Percentiles

- p^{th} percentile: Is a value such that at least p percent of the items take on this value or less and at most $(100 - p)$ percent of the items take on this value or more
- 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

Measures of Relative Standing: Location of Percentile

$$L_p = \frac{p}{100} \times (n + 1)$$

- Arrange the data in ascending order
- Compute L_p, the "location" of the pth percentile
- This location formula excludes the extreme values (Why?)
- By convention: $L_0 = 1$ and $L_{100} = n$

Example: 80^{th} Percentile

- Monthly Starting Salary:
- Compute the location of the pth percentile

$$L_p = (p/100)(n+1) = (80/100)(12+1) = 10.4$$

- The 80^{th} percentile is the 10^{th} value plus 0.4 times the difference between the 11^{th} and 10^{th} values
- 80^{th} percentile = 4,050 + 0.4 (4,130 - 4,050)= 4,082

Monthly Starting Salary ($)	Monthly Starting Salary ($)
3,710	3,755
3,850	3,880
3,880	3,890
3,920	3,940
3,950	4,050
4,130	4,325

Example: 80^{th} Percentile

- Monthly Starting Salary:
- Compute the location of the pth percentile

$$L_p = (p/100)(n+1) = (80/100)(12+1) = 10.4$$

- The 80^{th} percentile is the 10^{th} value plus 0.4 times the difference between the 11^{th} and 10^{th} values
- 80^{th} percentile = 4,050 + 0.4 (4,130 - 4,050)= 4,082

Monthly Starting Salary ($)	Monthly Starting Salary ($)
3,710	3,755
3,850	3,880
3,880	3,890
3,920	3,940
3,950	4,050
4,130	4,325

Example: 80^{th} Percentile

- Monthly Starting Salary:
- At least 80% of the observations take on a value of 4,082 or less
- Why?
- At most 20% of the observations take on a value of 4,082 or more
- Why?

Monthly Starting Salary ($)	Monthly Starting Salary ($)
3,710	3,755
3,850	3,880
3,880	3,890
3,920	3,940
3,950	4,050
4,130	4,325

Example: 80^{th} Percentile

- Monthly Starting Salary:
- At least 80% of the observations take on a value of 4,082 or less
- Why?
- At most 20% of the observations take on a value of 4,082 or more
- Why?

Monthly Starting Salary ($)	Monthly Starting Salary ($)
3,710	3,755
3,850	3,880
3,880	3,890
3,920	3,940
3,950	4,050
4,130	4,325

Measures of Relative Standing: Quartiles

- Quartiles are specific percentiles
  - 1^{st} Quartile $=$ Q1 $=$ 25^{th} Percentile
  - 2^{nd} Quartile $=$ Q2 $=$ 50^{th} Percentile $=$ Median (Why?)
  - 3^{rd} Quartile $=$ Q3 $=$ 75^{th} Percentile

Example: Third Quartile

- Monthly Starting Salary:
- Compute the location of the 75^{th} percentile

$$L_p = (p/100)(n+1) = (75/100)(12+1) = 9.75$$

- The 75^{th} percentile is the 9^{th} value plus 0.75 times the difference between the 10^{th} and 9^{th} values
- Q3 = 3,950 + 0.75 (4,050 - 3,950)= 4,025

Monthly Starting Salary ($)	Monthly Starting Salary ($)
3,710	3,755
3,850	3,880
3,880	3,890
3,920	3,940
3,950	4,050
4,130	4,325

Example: Third Quartile

- Monthly Starting Salary:
- Compute the location of the 75^{th} percentile

$$L_p = (p/100)(n+1) = (75/100)(12+1) = 9.75$$

- The 75^{th} percentile is the 9^{th} value plus 0.75 times the difference between the 10^{th} and 9^{th} values
- Q3 $=$ 3,950 $+$ 0.75 (4,050 - 3,950)$=$ 4,025

Monthly Starting Salary ($)	Monthly Starting Salary ($)
3,710	3,755
3,850	3,880
3,880	3,890
3,920	3,940
3,950	4,050
4,130	4,325

On the Agenda

Measures of Variability

- It is often desirable to consider measures of variability (dispersion), as well as measures of location

- 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

1. Range
2. Interquartile Range
3. Variance
4. Standard Deviation
5. Coefficient of Variation

Measures of Variability: Range

- Range: Is the difference between the largest and smallest data values

  Range = Maximum - Minimum

- It is the simplest measure of variability

- It is very sensitive to the smallest and largest data values

Example: Range

- Monthly Starting Salary:

  Range = Maximum - Minimum

  Range = 4,325 - 3,710

Monthly Starting Salary ($)	Monthly Starting Salary ($)
3,710	3,755
3,850	3,880
3,880	3,890
3,920	3,940
3,950	4,050
4,130	4,325

Example: Range

- Monthly Starting Salary:

Range = Maximum - Minimum

Range = 4,325 - 3,710 = 615

Monthly Starting Salary ($)	Monthly Starting Salary ($)
3,710	3,755
3,850	3,880
3,880	3,890
3,920	3,940
3,950	4,050
4,130	4,325

Measures of Variability: Range

- Interquartile Range: Is the difference between the third quartile and the first quartile

  IQR = Q3 - Q1

- It is the range for the middle 50% of the data

- It overcomes the sensitivity to extreme data values

Example: Interquartile Range (IQR)

- Monthly Starting Salary:

  Q3 = 4,025

  Q1 = 3,858 (Why?)

  IQR = Q3 - Q1 = 4,025 - 3,858 = 167

Monthly Starting Salary ($)	Monthly Starting Salary ($)
3,710	3,755
3,850	3,880
3,880	3,890
3,920	3,940
3,950	4,050
4,130	4,325

Measures of Variability: Variance

- Variance: Is the average of the squared differences from the mean
- It is based on the difference between the value of
  - Each observation x_i
  - Sample mean $\bar{x}$

    OR
  - Population mean μ
- The variance is useful in comparing the variability of two or more variables

Measures of Variability: Sample Variance

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

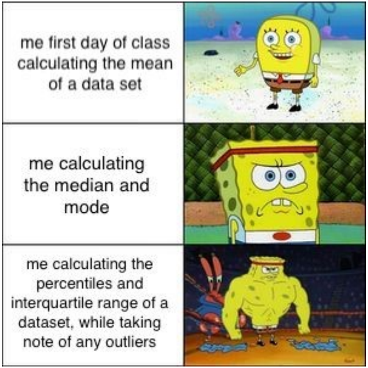
- $\sum_{i=1}^{n}$ represents the sum
- $(x_i - \bar{x})$ differences from the mean
- $(x_i - \bar{x})^2$ squared differences from the mean
- divide by $n - 1$ (Why?)

Measures of Variability: Population Variance

$$\sigma^2 = \frac{\sum_{i=1}^{N}(x_i - \mu)^2}{N}$$

- $\sum_{i=1}^{N}$ represents the sum
- $(x_i - \mu)$ differences from the mean
- $(x_i - \mu)^2$ squared differences from the mean
- divide by N

Measures of Variability



Example: Variance

- Monthly Starting Salary:

  $\bar{x} = 3,940$

Salary per Month: Xi	$(X_i - \overline{X})$	$(X_i - \overline{X})^2$
3,710		
3,755		
3,850		
3,880		
3,880		
3,890		
3,920		
3,940		
3,950		
4,050		
4,130		
4,325		

Total:

$s^2 =$

$\sigma^2 =$

Example: Variance

- Monthly Starting Salary:

  $\bar{x} = 3,940$

Salary per Month: Xi	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
3,710	-230	
3,755	-185	
3,850	-90	
3,880	-60	
3,880	-60	
3,890	-50	
3,920	-20	
3,940	0	
3,950	10	
4,050	110	
4,130	190	
4,325	385	
	Total:	
	$s^2 =$	
	$\sigma^2 =$	

Example: Variance

- Monthly Starting Salary:

$\bar{x} = 3,940$

Salary per Month: Xi	$(X_i - \bar{X})$	$(X_i - \bar{X})^2$
3,710	-230	52,900
3,755	-185	34,225
3,850	-90	8,100
3,880	-60	3,600
3,880	-60	3,600
3,890	-50	2,500
3,920	-20	400
3,940	0	0
3,950	10	100
4,050	110	12,100
4,130	190	36,100
4,325	385	148,225
	Total:	301,850
	$s^2 =$	27,440.91
	$\sigma^2 =$	25,154.17

Measures of Variability: Standard Deviation

- Standard Deviation: is the positive square root of the variance
- It is measured in the same units as the data
- It is more easily interpreted than the variance

Measures of Variability: Standard Deviation

- Sample

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

- Population

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

Measures of Variability: Coefficient of Variation

- Coefficient of Variation: It indicates how large the standard deviation is in relation to the mean

- Sample

$$CVS = \left(\frac{s}{\bar{x}}\right) \times 100\%$$

- Population

$$CVP = \left(\frac{\sigma}{\mu}\right) \times 100\%$$

Example: Sample Variance, Standard Deviation, SCV

- Monthly Starting Salary:
- Sample Variance

$$s^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} = 27,440.91$$

- Sample Standard Deviation

$$s = \sqrt{s^2} = \sqrt{27,440.91} = 165.65$$

- Sample Coefficient of Variation

$$CVS = \left(\frac{s}{\bar{x}}\right) \times 100\% = \left(\frac{165.65}{3,\bar{9}40}\right) \times 100\% = 4.2\%$$

On the Agenda

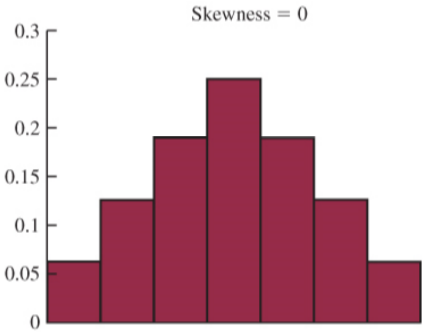
Measures of Distribution Shape

- Skewness is a measure of the shape of a distribution
- It is related to the asymmetry in a statistical distribution
- Skewness can be easily computed using statistical software
- The formula for the skewness of sample data is

$$b_1 = \frac{n}{(n-1)(n-2)} \sum_{i=1}^{n} \left[\frac{x_i - \bar{x}}{s} \right]^3$$

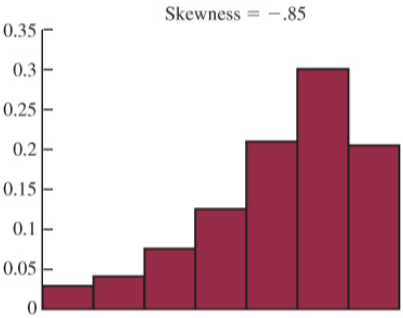
Measures of Distribution Shape: Skewness

- Symmetric: Not skewed
- Skewness is zero
- Mean and median are equal



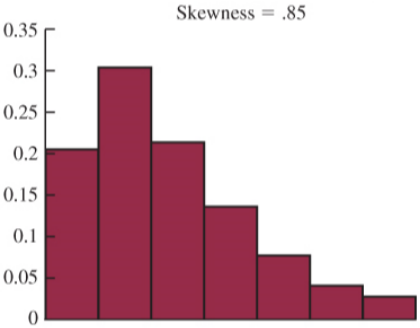
Measures of Distribution Shape: Skewness

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



Measures of Distribution Shape: Skewness

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



Measures of Relative Location: z-Scores

- The z-score denotes the number of standard deviations a data value x_i is from the mean

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

- The z-score is often called the standardized value

Measures of Relative Location: 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

Example: Class Size

- Class Size data:

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

Number of students In class	Deviation about the Mean	Z score $\left(\frac{x_i - \bar{x}}{s}\right)$
46		
54		
42		
46		
32		

Note: $\bar{x}=$? and $s=$? for the given data

Example: Class Size

- Class Size data:

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

Number of students In class	Deviation about the Mean	Z score $\left(\frac{x_i - \bar{x}}{s}\right)$
46	2	
54	10	
42	-2	
46	2	
32	-12	

Note: $\bar{x}$= 44 and s= 8 for the given data

Example: Class Size

- Class Size data:

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

Number of students In class	Deviation about the Mean	Z score $(\frac{x_i - \bar{x}}{s})$
46	2	2/8 = 0.25
54	10	10/8 = 1.25
42	-2	-2/8 = -0.25
46	2	2/8 = 0.25
32	-12	-12/8 = -1.5

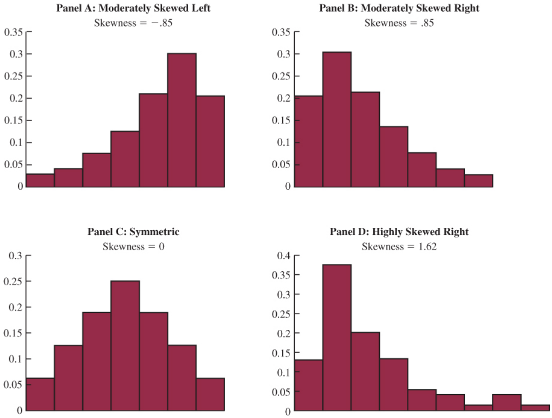
Note: $\bar{x}$= 44 and s= 8 for the given data

Chebyshev's Theorem

- At least $\left(1 - \frac{1}{k^2}\right)$ of the items in any data set will be within k standard deviations of the mean, where k is any value greater than 1
- Chebyshev's theorem requires $k > 1$, but k need not be an integer

Chebyshev's Theorem

- This is also called Chebyshev's inequality
- The theorem works for a wide class of distributions



Chebyshev's Theorem

- At least 75% of the data values must be within $k = 2$ standard deviations of the mean
- At least 89% of the data values must be within $k = 3$ standard deviations of the mean
- At least 94% of the data values must be within $k = 4$ standard deviations of the mean

Example: Marks of Students

- Marks of 100 students in a course had a mean of 70 and a standard deviation of 5
- We want to know the number of students having test scores between 60 and 80
- Notice that: 60 and 80 are 2 standard deviations below and above the mean respectively
- Hence, at least 75% of the data values must be within 60 and 80

Example: Marks of Students

- Marks of 100 students in a course had a mean of 70 and a standard deviation of 5
- We want to know the number of students having test scores between 58 and 72
- Notice that: 58 and 72 are 2.4 standard deviations below and above the mean respectively (Why?)
- Hence,

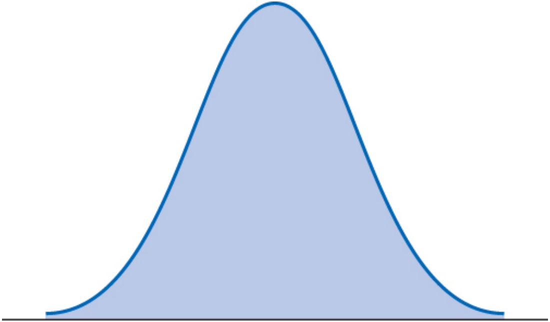
$$\left(1-\frac{1}{z^2}\right)=\left(1-\frac{1}{(2.4)^2}\right)=0.826=82.6\%$$

Empirical Rule

- When the data are believed to approximate a bell-shaped distribution:
  - The empirical rule can be used to determine the percentage of data values that must be within a specified number of standard deviations of the mean
  - The empirical rule is based on the normal distribution that we will cover later

Empirical Rule

- What is a bell-shaped distribution?



Empirical Rule

- For data having a bell-shaped distribution:
  - Approximately 68% of the data values will be within $\pm$ 1 standard deviation of its mean
  - Approximately 95% of the data values will be within $\pm$ 2 standard deviations of its mean
  - Approximately 100% of the data values will be within $\pm$ 3 standard deviations of its mean

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
- It might be:
  - an incorrectly recorded data value
  - a data value that was incorrectly included in the data set
  - a correctly recorded unusual data value that belongs in the data set

Example: Class Size

- Class Size data:

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

- -1.5 shows fifth class size is farthest from the mean
- No outliers are present as z value is within +/- 3 guideline for outliers

Number of students In class	Deviation about the Mean	Z score $(\frac{x_i - \bar{x}}{s})$
46	2	2/8 = 0.25
54	10	10/8 = 1.25
42	-2	-2/8 = -0.25
46	2	2/8 = 0.25
32	-12	-12/8 = -1.5

Note: $\bar{x}$= 44 and s= 8 for the given data

Five-Number Summary

- Minimum
- First Quartile
- Median
- Third Quartile
- Maximum

On the Agenda

1. Measures of Location
2. Measures of Variability
3. Measures of Distribution Shape
4. Measures of Association Between Two Variables

Measures of Association Between Two Variables

- We have examined numerical methods used to summarize the data for one variable at a time
- Often a manager or decision maker is interested in the relationship between two variables
- Two descriptive measures of the relationship between two variables are:
  - Covariance
  - Correlation Coefficient

Measures of Association Between Two Variables

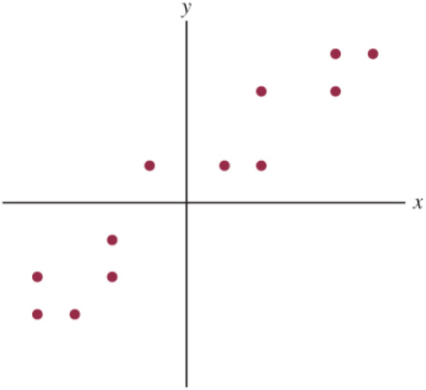
- Both terms measure the relationship and the linear dependency between two variables
- However, "covariance" indicates the direction of the <u>linear</u> relationship
- Whereas "correlation coefficient"measures both direction and strength of the <u>linear</u> relationship

Measures of Association Between Two Variables

- But what is a linear relationship anyway?

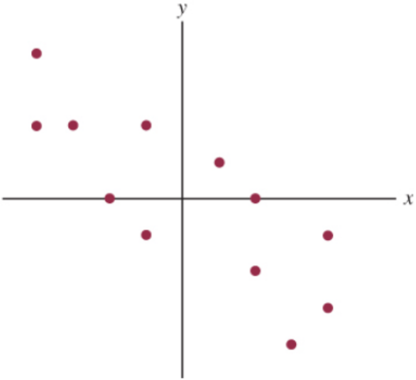
Measures of Association Between Two Variables

- But what is a linear relationship anyway?



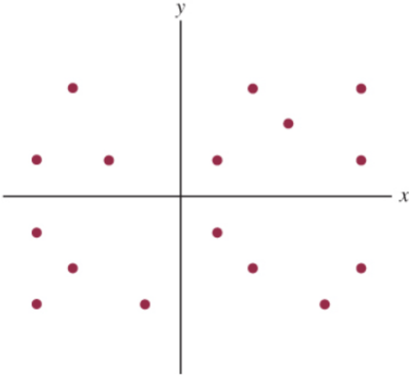
Measures of Association Between Two Variables

- But what is a <u>linear</u> relationship anyway?



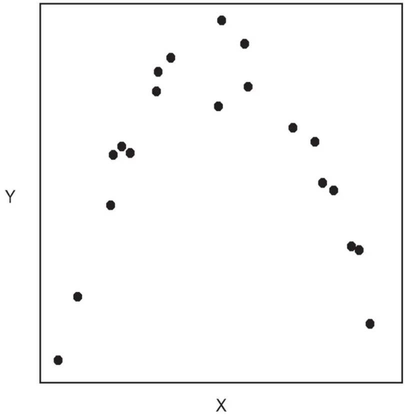
Measures of Association Between Two Variables

- But what is a <u>linear</u> relationship anyway?



Measures of Association Between Two Variables

- But what is a linear relationship anyway?



Covariance

- The covariance is a measure of the linear association between two variables
- Positive values indicate a positive relationship
- Negative values indicate a negative relationship

Covariance

- The covariance is computed as follows:
- For samples:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1}$$

- For population:

$$\sigma_{xy} = \frac{\sum_{i=1}^{N}(x_i - \mu_x)(y_i - \mu_y)}{N}$$

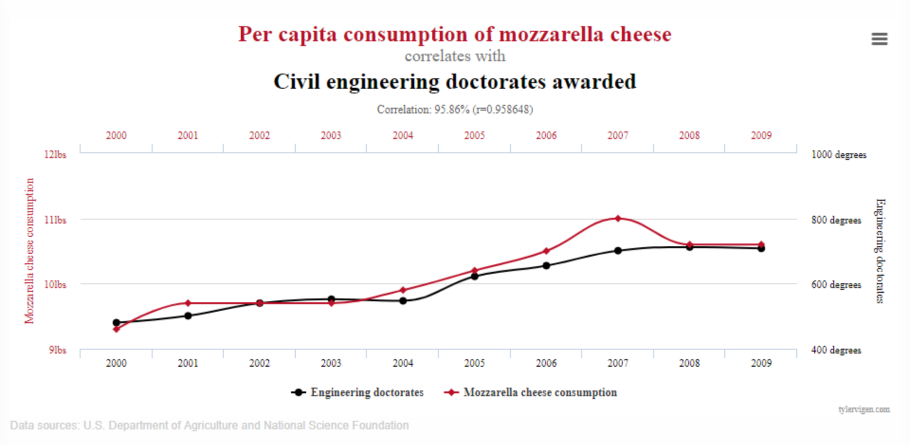
Correlation Coefficient

- Correlation is a measure of linear association and not necessarily causation

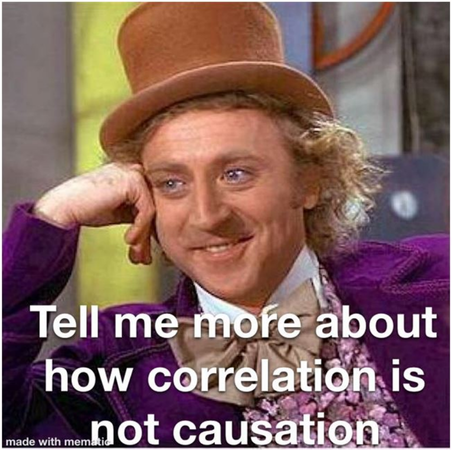
- Just because two variables are highly correlated, it does not mean that one variable is the cause of the other

Correlation Coefficient

- "correlation does not imply causation"



Correlation Coefficient



Correlation Coefficient

- The coefficient can take on values between -1 and $+1$
- Values near -1 indicate a strong negative linear relationship
- Values near 1 indicate a strong positive linear relationship
- The closer the correlation is to zero, the weaker the linear relationship

Correlation Coefficient

- The correlation coefficient is computed as follows:

- For samples:

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

- For population:

$$\rho_{xy} = \frac{\sigma_{xy}}{\sigma_x \sigma_y}$$

Example: Stereo and Sound Equipment Store

- The store's manager wants to determine the relationship (if any) between:
  - the number of weekend television commercials shown
  - the sales at the store during the following week

Week	Number of Commercials	Sales ($100s)
1	2	50
2	5	57
3	1	41
4	3	54
5	4	54
6	1	38
7	5	63
8	3	48
9	4	59
10	2	49

Example: Stereo and Sound Equipment Store

	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	$(x_i - \bar{x})(y_i - \bar{y})$
	2	50	−1	−1	1
	5	57	2	6	12
	1	41	−2	−10	20
	3	54	0	3	0
	4	54	1	3	3
	1	38	−2	−13	26
	5	63	2	12	24
	3	48	0	−3	0
	4	59	1	8	8
	2	46	−1	−5	5
Totals	30	510	0	0	99

$$s_{xy} = \frac{\sum(x_i - \bar{x})(y_i - \bar{y})}{n - 1} = \frac{99}{10 - 1} = 11$$

Example: Stereo and Sound Equipment Store

- Sample Covariance:

$$s_{xy} = \frac{\sum_{i=1}^{n}(x_i - \bar{x})(y_i - \bar{y})}{n-1} = \frac{99}{9} = 11$$

- Sample Correlation Coefficient

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{11}{1.49 \times 7.93} = 0.93$$