ECO239 Statistics I

Week 5

Describing Data: Numerically

- Central Tendency
  – Mean
  – Median
  – Mode
- Variation
  – Range
  – Interquartile Range
  – Variance
  – Standard Deviation

Mean

The *sample mean*, denoted as $\bar{x}$, can be calculated as

$$\bar{x} = \frac{x_1 + x_2 + \cdots + x_n}{n},$$

where x_1, x_2, ..., x_n represent the *n* observed values.

The *population mean* is also computed the same way but is denoted as μ. It is often not possible to calculate μ since population data are rarely available.

The sample mean is a *sample statistic*, and serves as a *point estimate* of the population mean. This estimate may not be perfect, but if the sample is good (representative of the population),
it is usually a pretty good estimate.

Calculating Mean

e.g. 1 {1, 2, 3, 4, 5}

e.g. 2 {1, 2, 3, 4, 20}

- Check raw data to detect outliers.
- Mean is affected by outliers/extreme values.

Weighted Mean

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

- GPA (Grade Point Average) Calculation

A1 = 4
A2 = 3.5
B1 = 3
B2= 2.5
C1 = 2
C2 = 1.5
D1 = 1
D2 = 0.5
F = 0

	Grade	Score (X)	Credit (W)	X*W
ECO239	A1		3	
ECO336	A2		3	
ECO448	B1		3	
ING250	C2		2	
ING350	A2		2	
TOTAL			13	

Median

The *median* is the value that splits the data in half when ordered in ascending order.

$$0, 1, 2, 3, 4$$

If there are an even number of observations, then the median is the average of the two values in the middle.

$$0, 1, \underline{2, 3}, 4, 5 \rightarrow \frac{2+3}{2} = 2.5$$

Since the median is the midpoint of the data, 50% of the values are below it. Hence, it is also the 50th percentile.

Median

Finding Median

Step 1: Order the data in ascending order
Step 2: Find Median Position = (n+1)/2
Step 3: Find the Median at the Median Position

If n is **odd**, Median is the middle number.
e.g. n=5 => Median position = (5+1)/2 = 3.

If n is even, Median is the average of two middle numbers.
e.g.n=12, Median position = (12+1)/2=6.5. Median is average of 6^{th} and 7^{th} values.

Median

- Data{ 8, 4, 3, 5, 9, 7, 8}

Find Median.

Median

- Data { 4, 3, 5, 7, 8, 8, 20}

*Median is not affected by an extreme value.

Median

Data {4, 3, 5, 7, 8, 8, 9, 20}

Mode

- Value that occurs most often.

- There may not be any mode.
- There may be multiple mode.

e.g.
1, 3, 4, 5, 5, 7, 9, 9, 9, 10, 12, 12, 13, 14

Mode = 9

Mode

e.g. 3, 5, 7, 4, 8, 8, 9
e.g. 3, 5, 7, 4, 8, 8, 30

=> Not affected by extreme values.

Mode

e.g. 0, 1, 2, 3, 5, 6
Mode = No Mode

e.g. 0, 1, 1, 1, 2, 3, 3, 3 ,4, 5, 6
Mode = 1 and 3

Practice

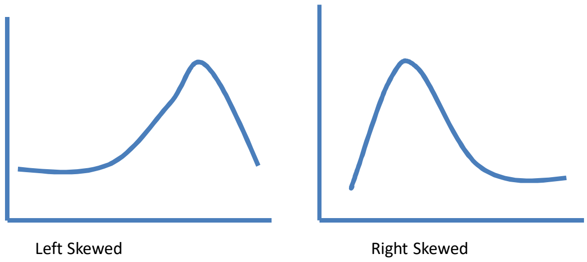
Housing Prices
1. $ 2,000,000
2. $ 500,000
3. $ 300,000
4. $ 100,000
5. $ 100,000

Q: Find mean, median and mode.

When do we use Mean and when do we better use Median???

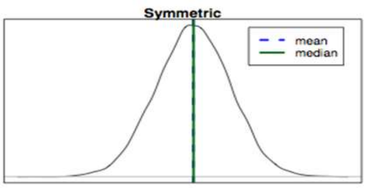
Shape of Distribution

- We can judge skewness using Mean and Median
- Consider the relative size of Mean and Median in the following cases.



Mean vs. Median

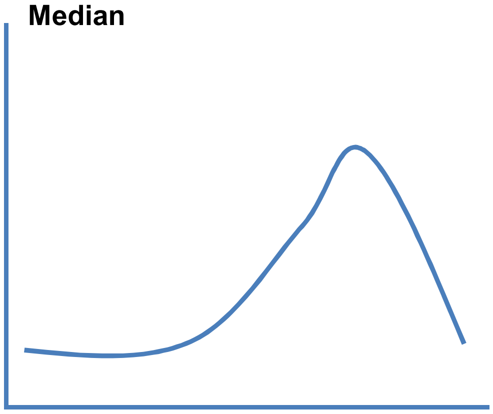
If the distribution is symmetric, center is often defined as the mean:
mean ~ median



If the distribution is skewed or has extreme outliers, center is often defined as the median

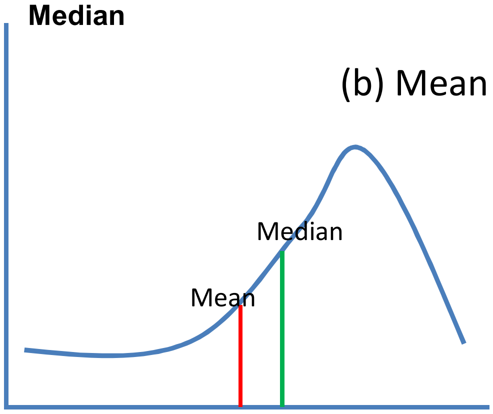
- Right-skewed: mean > median
- Left-skewed: mean < median

Question: For Left Skewed Distribution, the relative location of Mean and Median are…? (a) Mean>Median, (b) Mean<Median, (c) Mean = Median



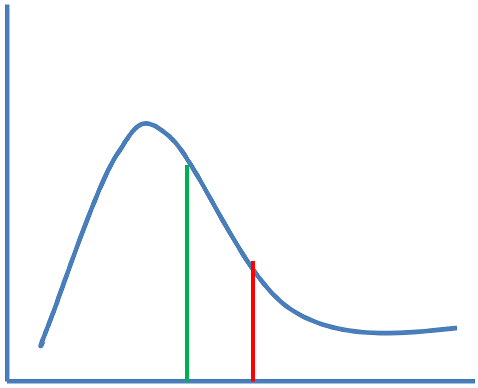
Question: For Left Skewed Distribution, the relative location of Mean and Median are…? (a) Mean>Median, (b) Mean<Median, (c) Mean = Median

(b) Mean < Median



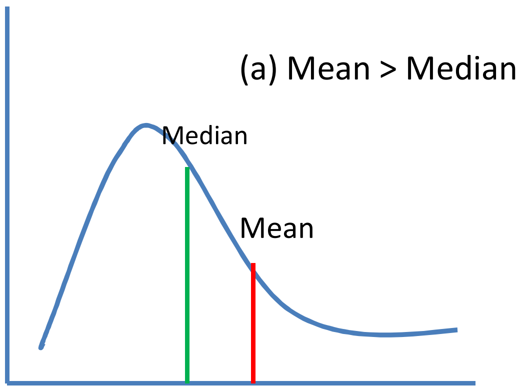
Question: For Right Skewed Distribution,

the relative location of Mean and Median are…? (a) Mean>Median, (b) Mean<Median, (c) Mean = Median



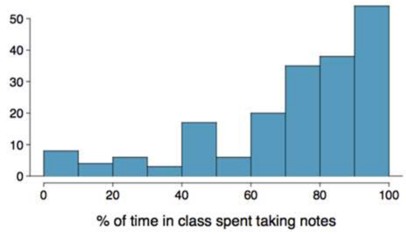
Question: For Right Skewed Distribution,

the relative location of Mean and Median are…? (a) Mean>Median, (b) Mean<Median, (c) Mean = Median



Practice

Which is most likely true for the distribution of percentage of time actually spent taking notes in class versus on Facebook, Twitter, etc.?



(a) mean > median
(b) mean ~ median
(c) mean < median
(d) impossible to tell

Practice

Data { 40, 45, 50, 51, 55, 60, 80, 99}
n= 8

Comment on skewness.

Measures of Variation

- Range
- Interquartile Range
- Variance
- Standard Deviation

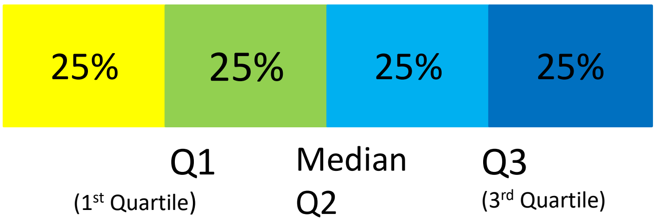
Range

- Range = X_largest – X_smallest

- E.g. {7, 8, 9, 11, 12}
- Range = 12-7 = 5

Interquartile Range
(used in Box plot & more)

- Quartiles: split the ranked data into 4 segments with an equal number of values per segment.



Interquartile Range = Q3 – Q1

Q1 locates in ¼ (n+1) position
(25% below, 75% above)

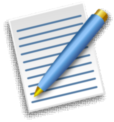
Q2 locates in ½(n+1) position
(50% below, 50% above)

Q3 locates in ¾(n+1) position
(75% below, 25%above)

IQR: Practice 1

- Data {11, 7, 14, 16, 21, 0, 3, 2, 17, 9, 13}
- n = 11

- Derive Q1, Q2 and Q3 values.



IQR: Practice 2

Data {11, 7, 14, 22,16, 21, 0, 3, 2, 17, 23, 9, 13}
n= 13

Derive IQR.

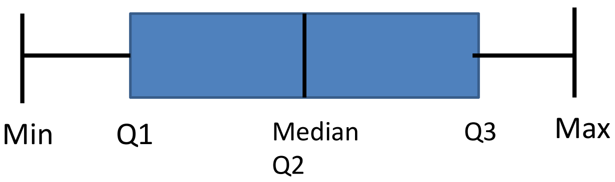
IQR: Practice 3

Data { 0, 2, 3, 7, 9, 11, 13, 14, 16, 17, 21, 22}
n= 12

Derive IQR. 

Five Number Summary

Min < Q1 < Median (Q2) < Q3 < Max



IQR

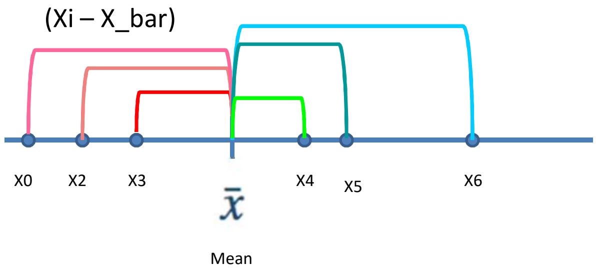
Compare the following two cases.

Data{ 0, 2, 3, 7, 9, 11, 13, 14, 16, 17, 21}
IQR = ?

Data{ 0, 2, 3, 3, 4, 4, 14, 15, 16, 16, 21}
IQR= ?

Variance

Measuring the average of the total distance between each observation and the mean.



Variance: Calculation

Step 1: Compute the distance between each data point and mean.

Step 2: Square the each distance

Step 3: Sum all the squared distances and divide by observation size (for population) OR by observation size – 1 (for sample data)

Population Variance

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

μ = Population Mean

X_i = each observation, i = 1,… N.

N = population size

Sample Variance

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

$\bar{x}$:Sample Mean

n : Sample Size

Why do we use the squared deviation in the calculation of variance?

- To get rid of negatives so that observations equally distant from the mean are weighed equally.
- To weigh larger deviations more heavily.

WHY???

- Do we divide by (n-1), instead of n for sample variance???

A: Sample variance is an unbiased estimator of the population variance. It's a better estimator of the population variance if divided by n-1.

<u>This will be discussed in detail in ECO240.</u>

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

Sample Data { 50, 60, 65, 70, 88}
Calculate sample variance.

NOTE: Variance

- If you forget to square the distance, the calculated value =

Standard Deviation

: Average spread around the mean

: Square root of the variance.

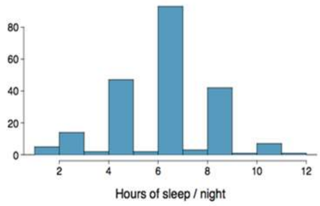
: Has the same unit as the original data

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

Standard Deviation

The *standard deviation* is the square root of the variance, and has the same units as the data.

- The standard deviation of amount of sleep students get per night can be calculated as: $s = \sqrt{s^2}$



$$s = \sqrt{4.11} = 2.03\ hours$$

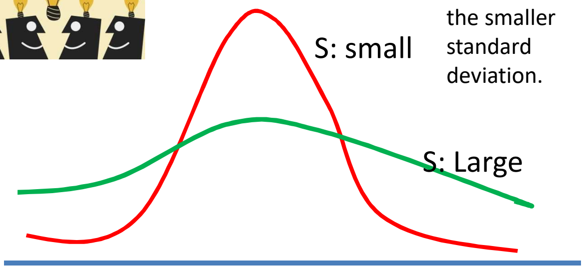
- We can see that all of the data are within 3 standard deviations of the mean.

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

Data { 50, 60, 65, 70, 88}
Calculate standard deviation.

*IF the data is exam score, for example, the unit of standard deviation is also “exam points”.

Which s is smaller / larger???



Red one has the smaller standard deviation.

Compare standard deviations

Data1 { 11, 12, 13, 16, 16, 17, 18, 21}
Data2 { 14, 15, 15, 15, 16, 16, 16, 17}
Data3 { 11, 11, 11, 12, 19, 20, 20, 20}

1. Guess which one has the smallest/largest standard deviation.
2. Calculate Mean and Standard Deviation.

Compare.

Summary Statistics in R

min, Q1, median, mean, Q3, max for all variables included in the file.

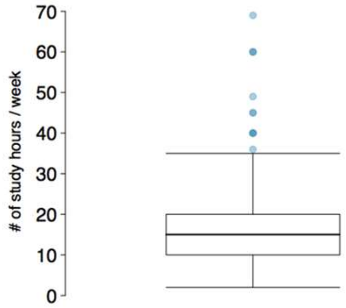
summary(gpa_sec2)

same summary for just one of the variables in gpa_sec2.csv file.

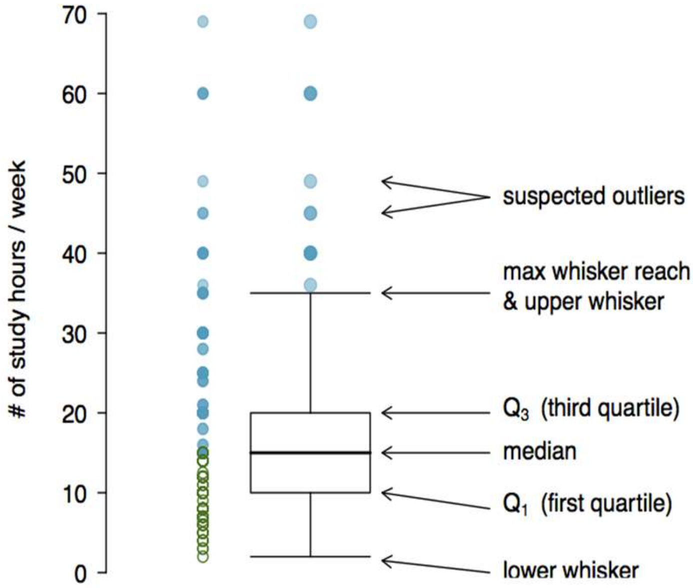
summary(gpa_sec2$gpa)

Box Plot

The box in a *box plot* represents the middle 50% of the data, and the thick line in the box is the median.



Anatomy of a Box Plot



Whiskers and Outliers

Whiskers of a box plot can extend up to 1.5 x IQR away from the quartiles.

max upper whisker reach = Q3 + 1.5 x IQR
max lower whisker reach = Q1 - 1.5 x IQR

Whiskers and Outliers

Whiskers of a box plot can extend up to 1.5 x IQR away from the quartiles.

max upper whisker reach = Q3 + 1.5 x IQR

max lower whisker reach = Q1 - 1.5 x IQR

IQR: 20 - 10 = 10

max upper whisker reach = 20 + 1.5 x 10 = 35

max lower whisker reach = 10 - 1.5 x 10 = -5

Whiskers and Outliers

Whiskers of a box plot can extend up to 1.5 x IQR away from the quartiles.

max upper whisker reach = Q3 + 1.5 x IQR

max lower whisker reach = Q1 - 1.5 x IQR

IQR: 20 - 10 = 10

max upper whisker reach = 20 + 1.5 x 10 = 35

max lower whisker reach = 10 - 1.5 x 10 = -5

A potential *outlier* is defined as an observation beyond the maximum reach of the whiskers. It is an observation that appears extreme relative to the rest of the data.

Outliers (cont.)

Why is it important to look for outliers?

Outliers (cont.)

Why is it important to look for outliers?

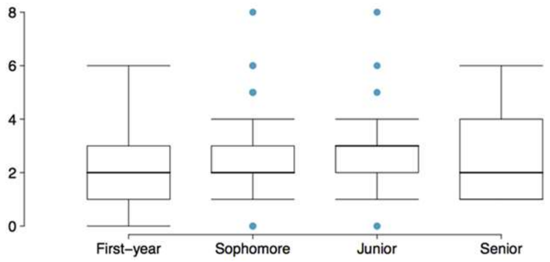
- Identify extreme skew in the distribution.
- Identify data collection and entry errors.
- Provide insight into interesting features of the data.

Box Plot

```
attach(gpa_sec2)
boxplot(gpa~gender,data=gpa_sec2,names=c("
male","female"), main="GPA by gender",
xlab="Gender", ylab="GPA")
```

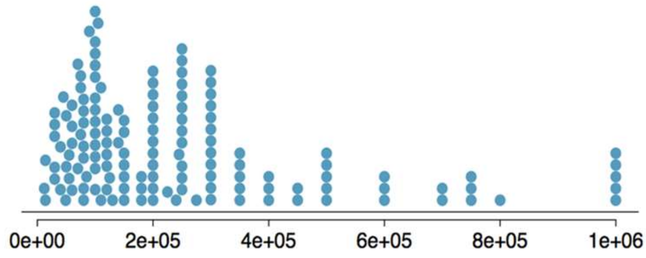
Comparing Numerical Data Across Groups

Does there appear to be a relationship between class year and number of clubs students are in?

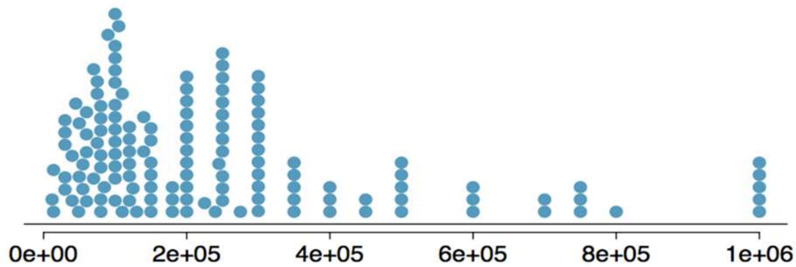


Extreme Observations

How would sample statistics such as mean, median, SD, and IQR of household income be affected if the largest value was replaced with $10 million?



Robust Statistics



	robust		not robust	
scenario	median	IQR	$\bar{x}$	s
original data	190K	200K	245K	226K
move largest to $10 million	190K	200K	309K	853K

Robust Statistics

Median and IQR are more robust to skewness and outliers than mean and SD. Therefore,

- for skewed distributions it is often more helpful to use median and IQR to describe the center and spread
- for symmetric distributions it is often more helpful to use the mean and SD to describe the center and spread

If you would like to estimate the typical household income for a student, would you be more interested in the mean or median income?

Extremely Skewed Data

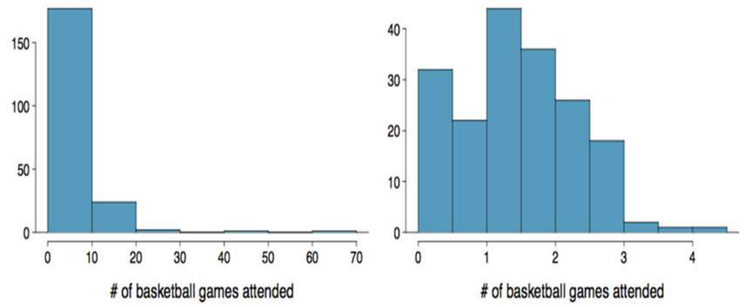
When data are extremely skewed, transforming them might make modeling easier.

A common transformation is the *log transformation*.

Extremely Skewed Data

When data are extremely skewed, transforming them might make modeling easier. A common transformation is the log transformation.

The histograms on the left shows the distribution of number of basketball games attended by students. The histogram on the right shows the distribution of log of number of games attended.



Pros and Cons of Transformations

- Skewed data are easier to model with when they are transformed because outliers tend to become far less prominent after an appropriate transformation.

  # of games 70 50 25 ...

  # of games 4.25 3.91 3.22 ...

- However, results of an analysis might be difficult to interpret because the log of a measured variable is usually meaningless.

  What other variables would you expect to be extremely skewed?

  Salary, housing prices, etc.

Measures of Relationship between Variables

- Covariance
- Correlation Coefficient

Covariance

- A measure of the linear relationship between two variables
- Only concerned with the direction of the relationship.

Population Covariance

$$\mathrm{Cov}(X,Y) = \sigma_{xy} = \frac{\sum_{i=1}^{N}(X_i - \mu_x)(Y_i - \mu_Y)}{N}$$

Sample Covariance

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

Covariance: Meaning

Cov(x,y) > 0 => X and Y tend to move in the same direction.

Cov(x,y) < 0 => X and Y tend to move in the opposite direction.

Cov(x,y)= 0 => X and Y are independent

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

Practice: Calculate COV(X,Y)

X (# Workers)	Y (# Cell phones)			
12	20			
30	60			
15	27			
24	50			
14	23			

Correlation Coefficient

- Measures the relative strength and direction of the linear relationship between two variables.

Population Correlation Coefficient

$$\rho = \frac{\mathrm{Cov}(X,Y)}{\sigma_X \sigma_Y}$$

Population Covariance

Standard Deviation of X, and Y.

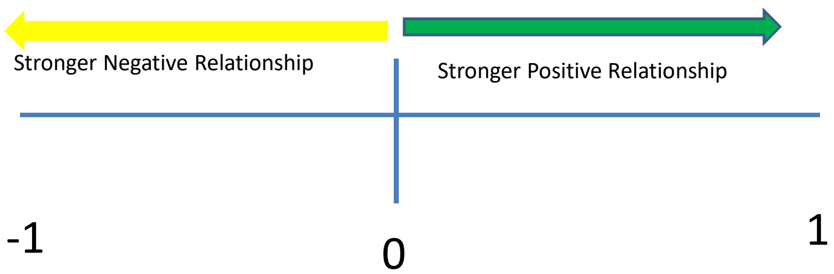
Sample Correlation Coefficient

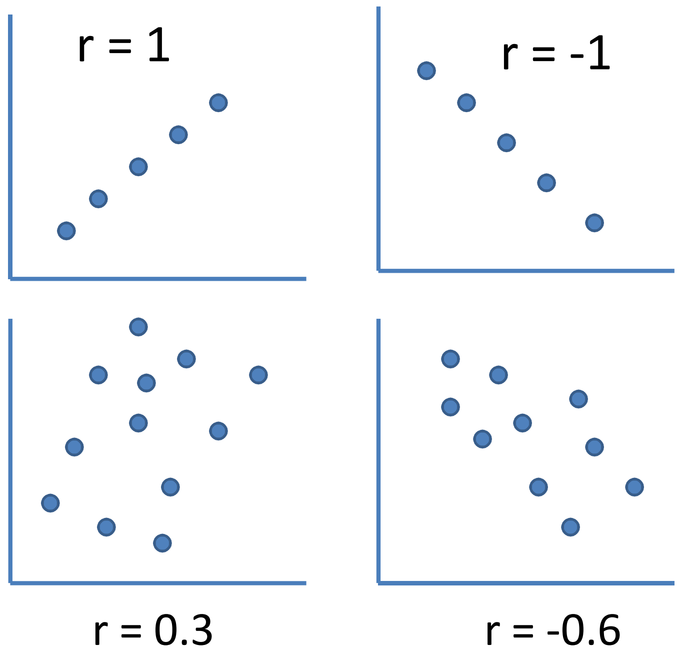
$$r = \frac{\mathrm{Cov}(x,y)}{s_X s_Y}$$

Sample Covariance

Standard Deviation of x and y.

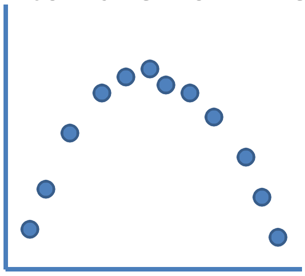
Correlation Coefficient: Meaning

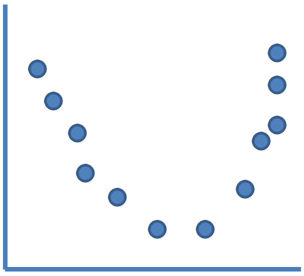




Correlation Coefficient: Caution!

- This coefficient measures LINEAR relationship, not Non-Linear.
- Even if r = 0, it is possible for x and y variables to have non-linear relationship.





Calculate Correlation Coefficient

X (# Workers)	Y (# Cell phones)
12	20
30	60
15	27
24	50
14	23
X_bar=19	Y_bar = 36

- COV(x,y) = 136.75.

$$r = \frac{\mathrm{Cov}(x, y)}{s_X s_Y}$$

- Variance

var(gpa)

- Standard Deviation

sd(gpa)

- Covariance

cov(x,y)

- Correlation Coefficient

cor(x,y)