LECTURE 15
COVARIANCE & CORRELATION

PSY2002

Hye Won Suk

RELATIONSHIP BETWEEN TWO VARIABLES

- We have examined the relationship between a categorical variable and a continuous variable.

- For example, we can use an independent-samples *t*-test to compare males and females on their midterm scores

    - Gender: a categorical variable (male, female)
    - Midterm score: a continuous variable (score)

RELATIONSHIP BETWEEN TWO VARIABLES

- What if we want to investigate the relationship between two continuous (interval or ratio) variables, for example,
  - Age and physical capacity
  - SES and math achievement
  - BMI and abdominal circumference

- We can use
  - Scatterplot
  - Covariance/Correlation

SCATTER PLOT (산포도)

- BMI (체질량) and abdominal circumference (복부 둘레)
  - Each point (circle) represents each person.

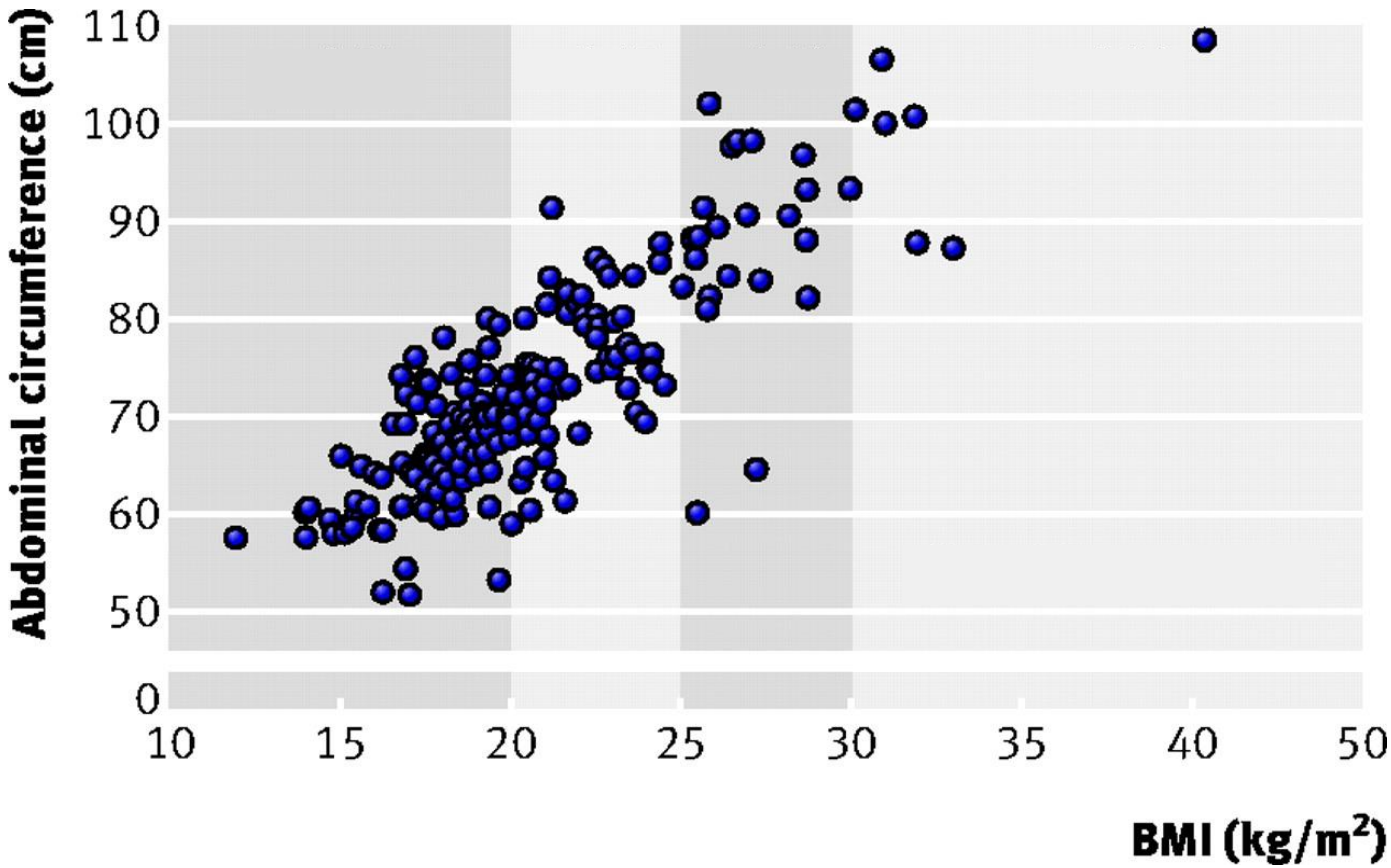
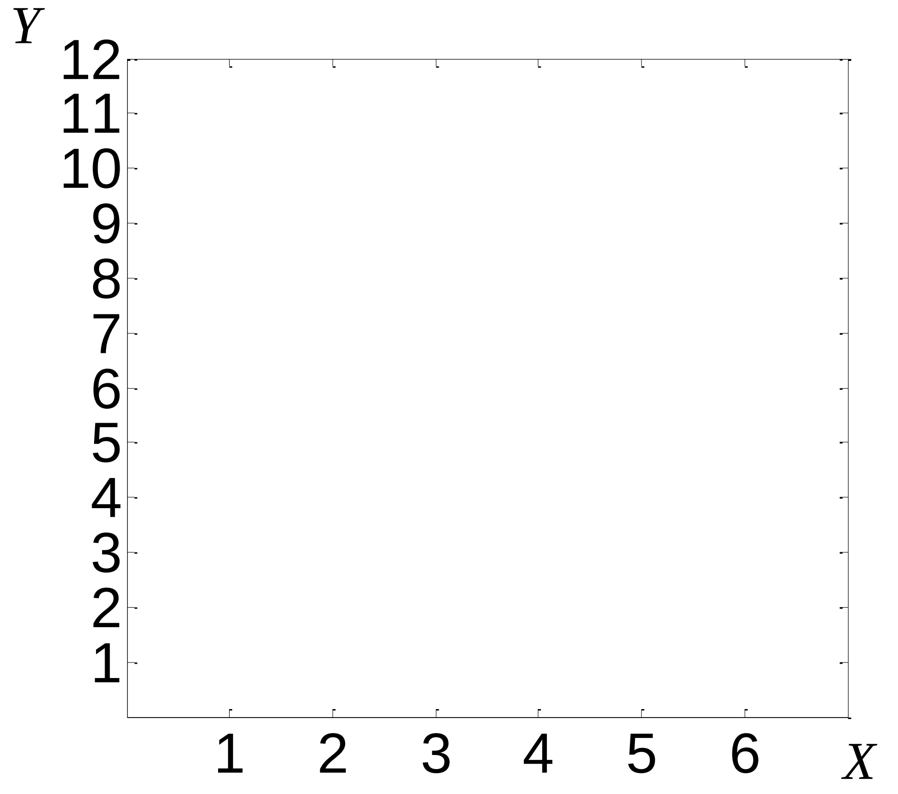


Image retrieved from: http://www.bmj.com/content/342/bmj.d556

SCATTER PLOT

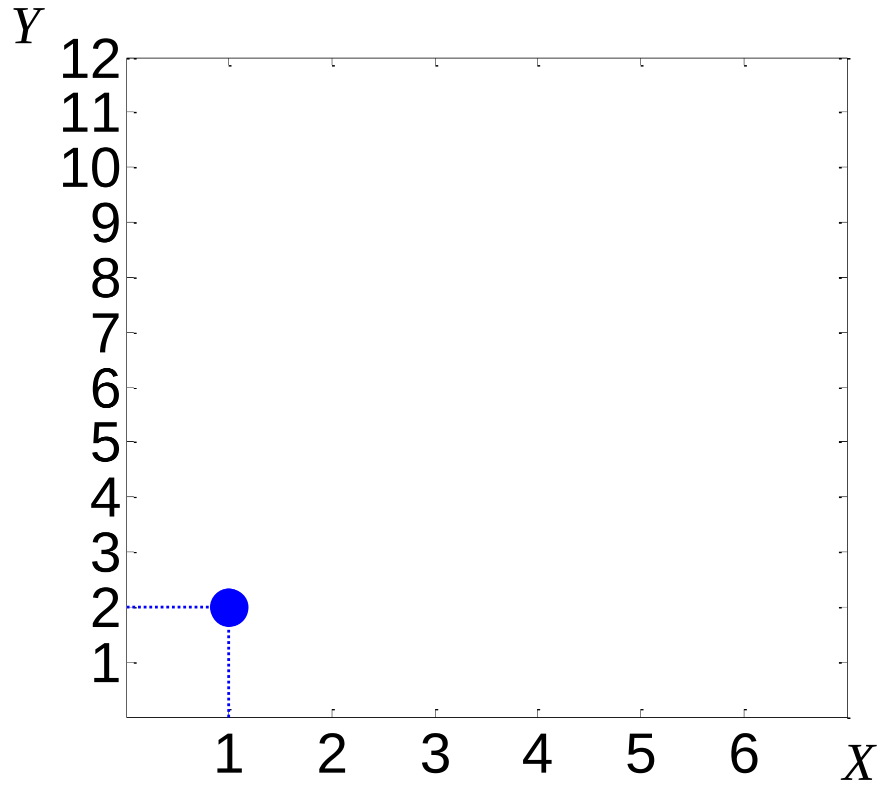
- How to make a scatter plot? Let's consider this toy example.

X	Y
1	2
2	5
3	6
4	10
5	11
6	11



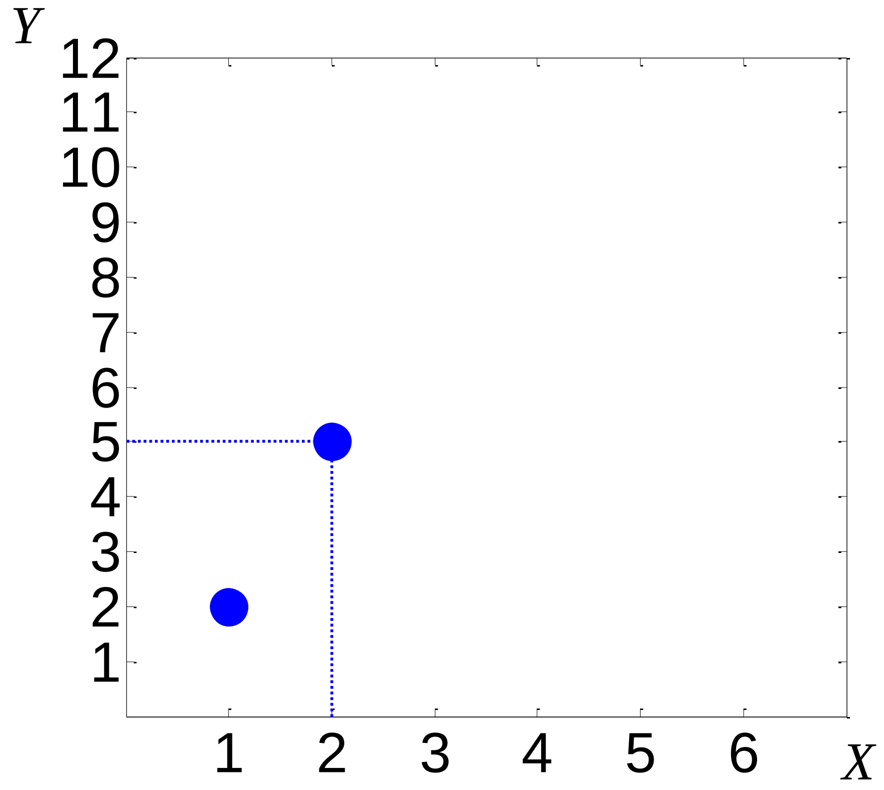
SCATTER PLOT

X	Y
1	2
2	5
3	6
4	10
5	11
6	11



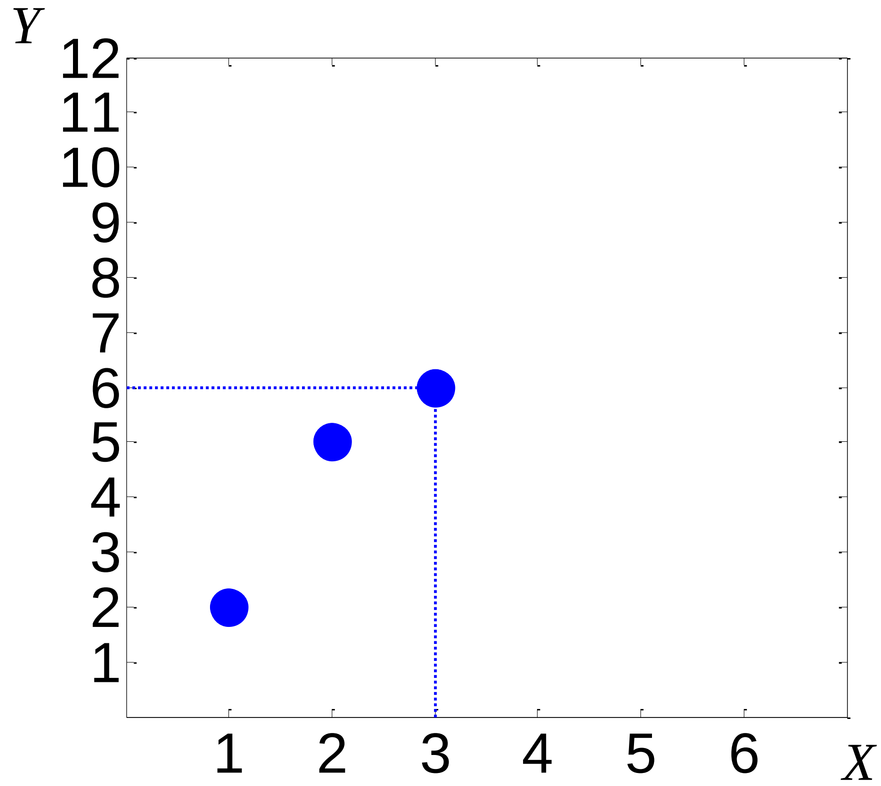
SCATTER PLOT

X	Y
1	2
2	**5**
3	6
4	10
5	11
6	11



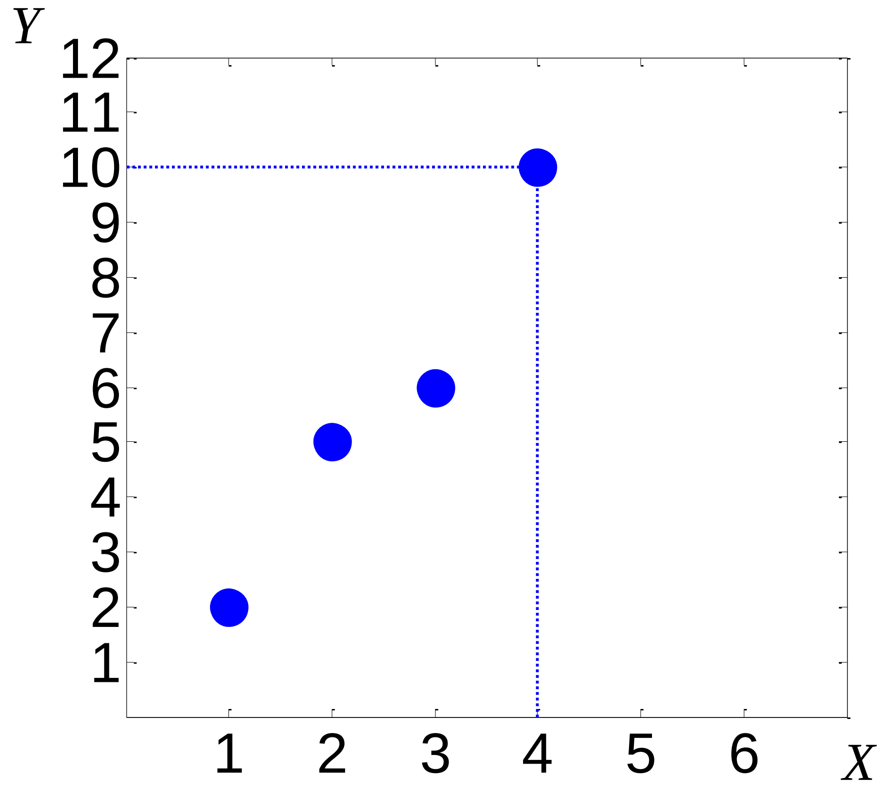
SCATTER PLOT

X	Y
1	2
2	5
3	**6**
4	10
5	11
6	11



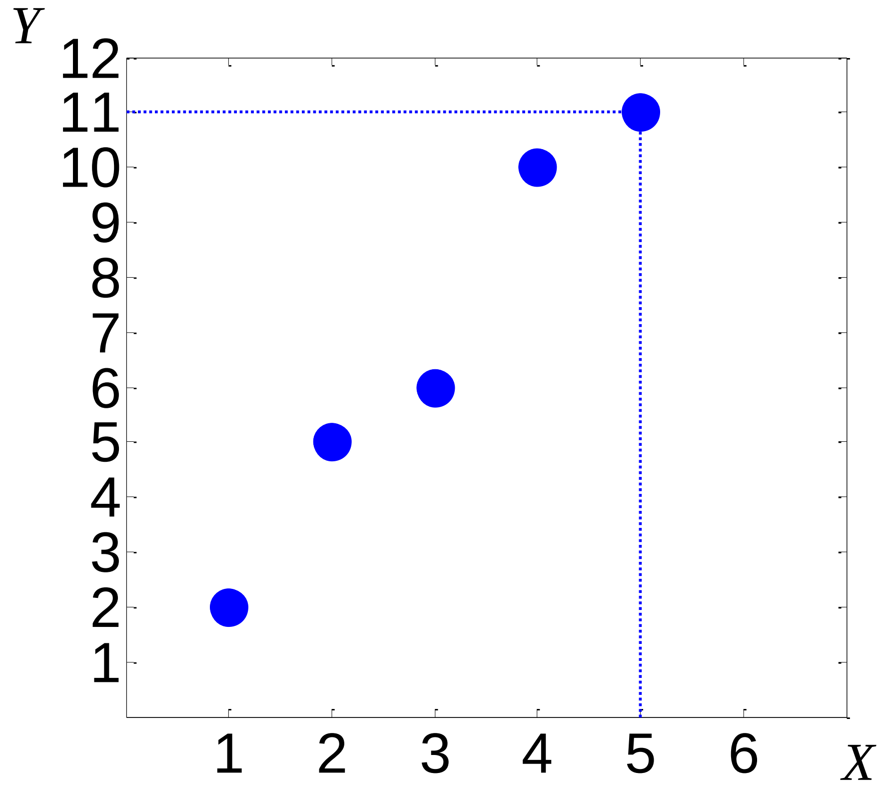
SCATTER PLOT

X	Y
1	2
2	5
3	6
4	**10**
5	11
6	11



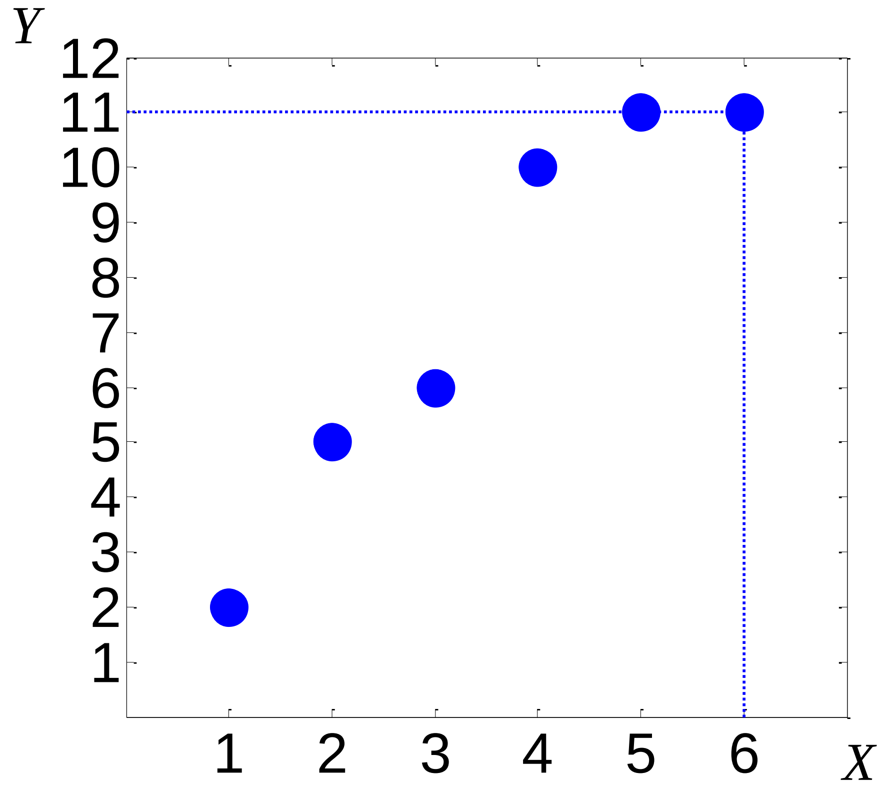
SCATTER PLOT

X	Y
1	2
2	5
3	6
4	10
5	**11**
6	11



SCATTER PLOT

X	Y
1	2
2	5
3	6
4	10
5	11
6	11



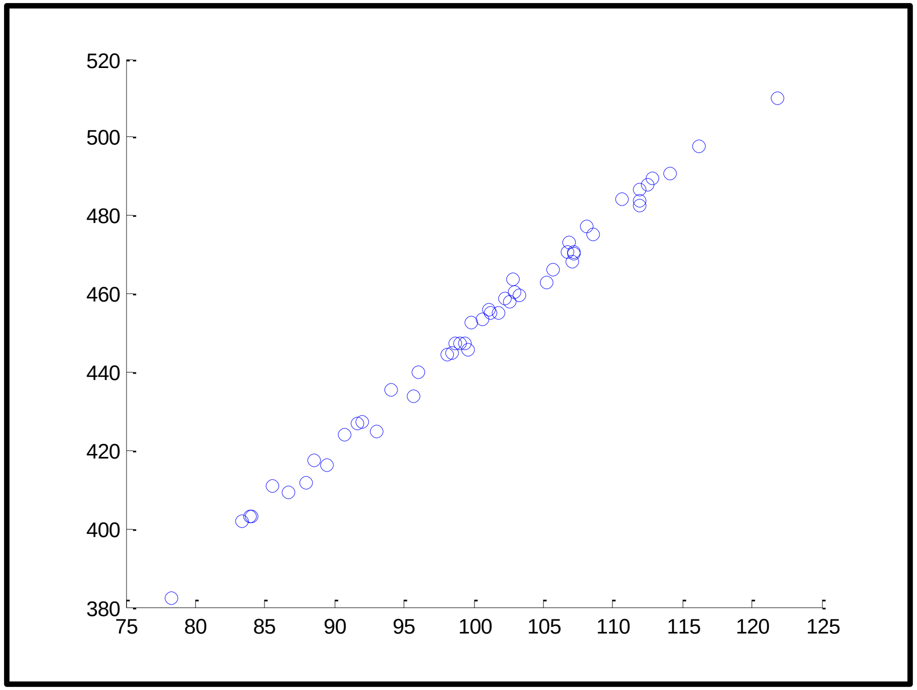
SCATTER PLOT

- You need to interpret three things for a scatterplot:

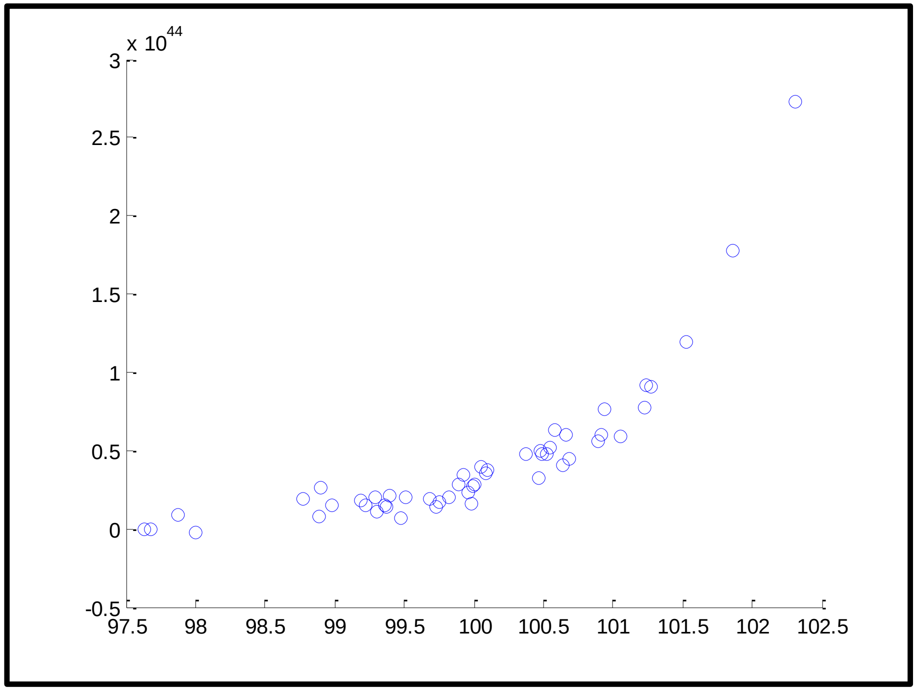
  - Form of relationship (linear or nonlinear)
  - Direction of relationship (positive or negative)
  - Strength of relationship (no, weak, or strong)

1. FORM OF RELATIONSHIP

Linear

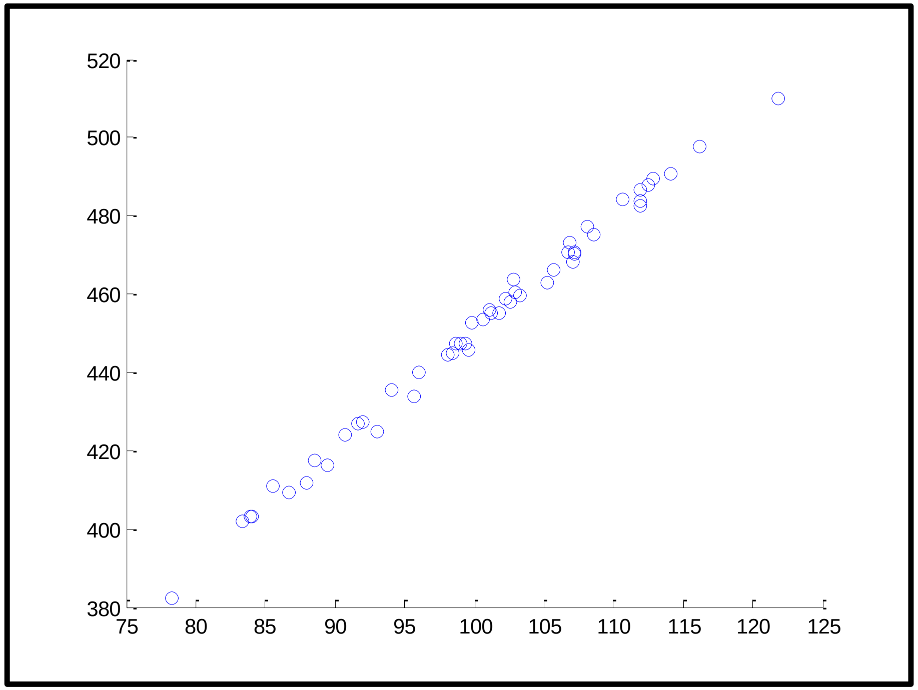


Nonlinear

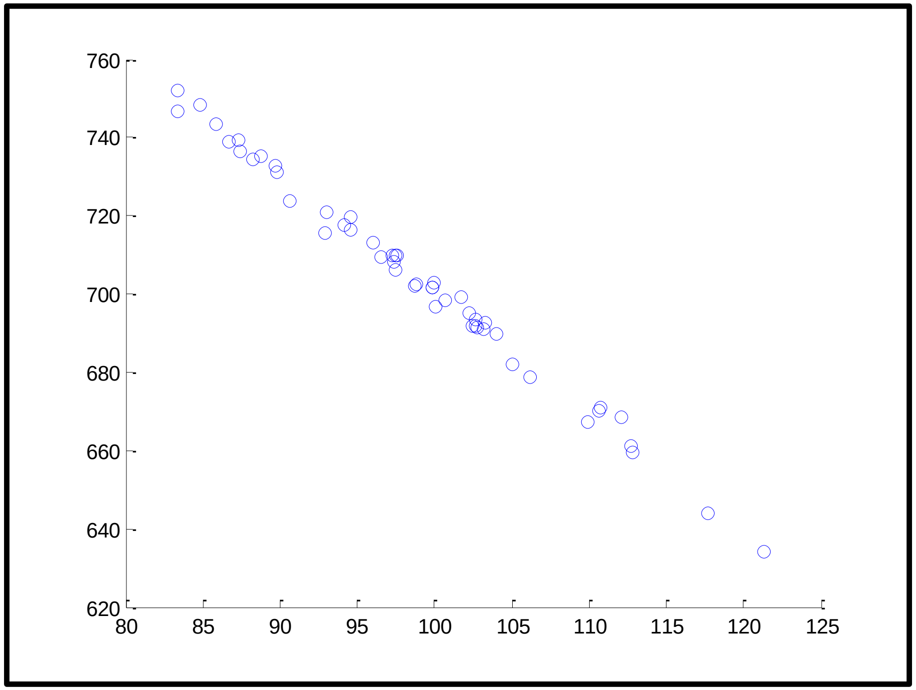


2. DIRECTION OF RELATIONSHIP (WHEN LINEAR)

Positive

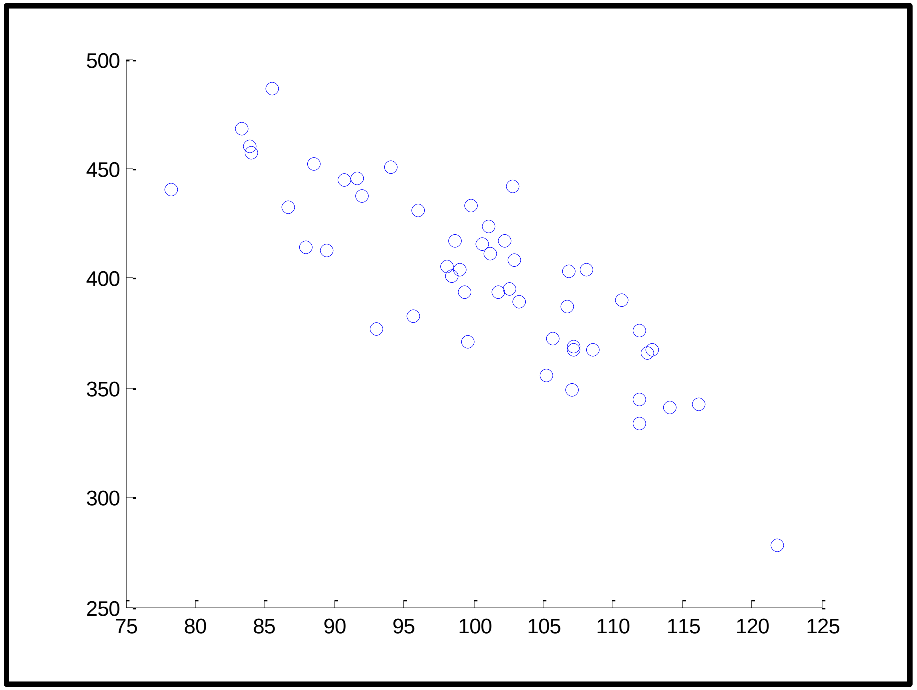


Negative

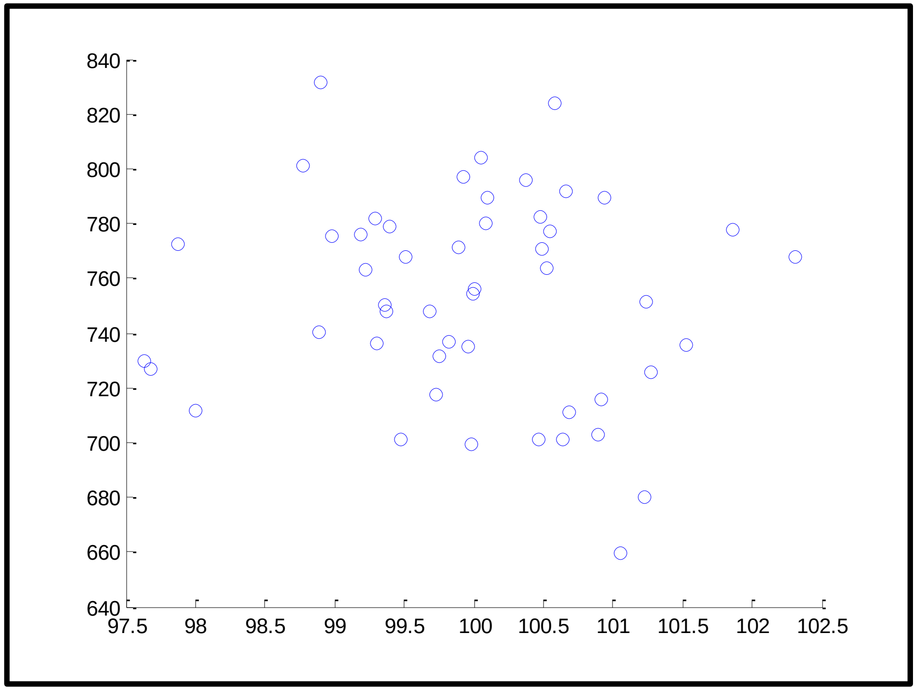


3. STRENGTH OF RELATIONSHIP

Strong



Weak



LIMITATIONS OF SCATTER PLOTS

- The interpretation of a scatterplot might be subjective because we should examine the relationship by our eyes.

→ We need an objective quantity that measures the relationship between two continuous variables.

  - COVARIANCE
  - CORRELATION

  - To calculate covariance or correlation, we need to calculate SP (sum of products of deviations). Let's take a look at how to calculate it.

SUM OF PRODUCTS OF DEVIATIONS

- Sum of products of deviations (SP)
  - Population SP

$$\boxed{SP = \sum_{i=1}^{N} (X_i - \mu_X)(Y_i - \mu_Y)}$$

- μ_X: population mean of X
- μ_Y: population mean of Y

  - Sample SP

$$\boxed{SP = \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y})}$$

PROPERTIES OF SP

- If *X* and *Y* are positively related, SP will be positive.

- If *X* and *Y* are negatively related, SP will be negative.

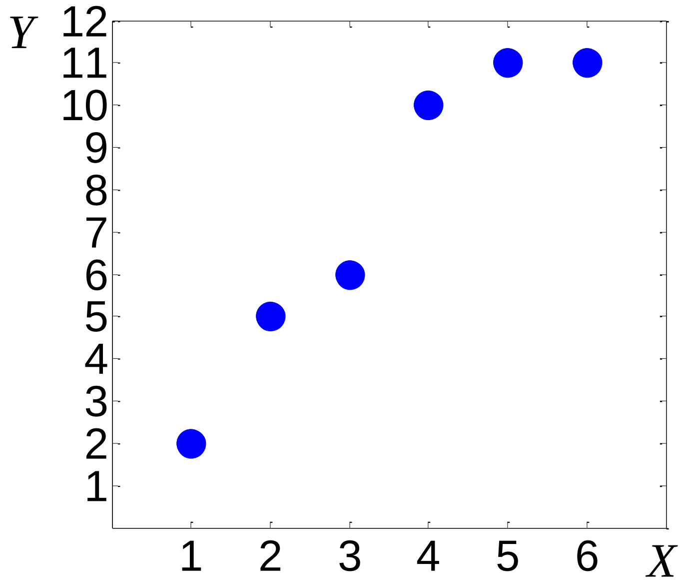
- If *X* and *Y* are not linearly related, SP will be close to 0.

- These properties will be demonstrated using a sample data. However, there properties will hold for the population as well.

1. POSITIVE RELATIONSHIP

- How to calculate SP?

X	Y
1	2
2	5
3	6
4	10
5	11
6	11



1. POSITIVE RELATIONSHIP

- Step 1: Calculate the mean for each variable.

X	Y
1	2
2	5
3	6
4	10
5	11
6	11

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{2+5+6+10+11+11}{6} = 7.5$$

1. POSITIVE RELATIONSHIP

- Step 2 : Calculate the deviation scores for each variable.

X	Y	$X - \bar{X}$	$Y - \bar{Y}$
1	2	1 – 3.5 = -2.5	2 – 7.5 = -5.5
2	5	2 – 3.5 = -1.5	5 – 7.5 = -2.5
3	6	3 – 3.5 = -0.5	6 – 7.5 = -1.5
4	10	4 – 3.5 = 0.5	10 – 7.5 = 2.5
5	11	5 – 3.5 = 1.5	11 – 7.5 = 3.5
6	11	6 – 3.5 = 2.5	11 – 7.5 = 3.5

1. POSITIVE RELATIONSHIP

- Step 3 : Calculate the products of deviations.

X	Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$
1	2	1 – 3.5 = -2.5	2 – 7.5 = -5.5	(-2.5)(-5.5) = 13.75
2	5	2 – 3.5 = -1.5	5 – 7.5 = -2.5	(-1.5)(-2.5) = 3.75
3	6	3 – 3.5 = -0.5	6 – 7.5 = -1.5	(-0.5)(-1.5) = 0.75
4	10	4 – 3.5 = 0.5	10 – 7.5 = 2.5	(0.5)(2.5) = 1.25
5	11	5 – 3.5 = 1.5	11 – 7.5 = 3.5	(1.5)(3.5) = 5.25
6	11	6 – 3.5 = 2.5	11 – 7.5 = 3.5	(2.5)(3.5) = 8.75

1. POSITIVE RELATIONSHIP

- Step 4 : Calculate the sum of products of deviations.

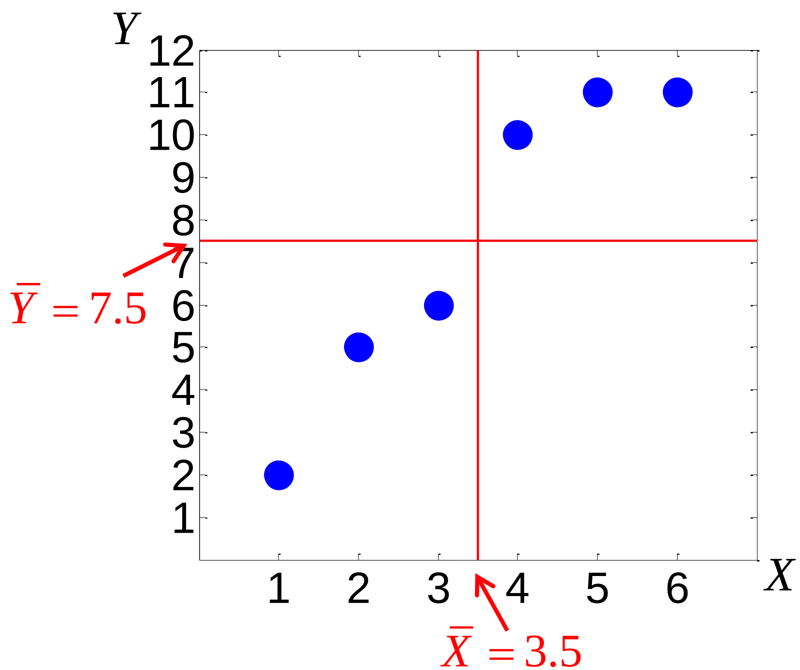
X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$
1	2	1 – 3.5 = -2.5	2 – 7.5 = -5.5	(-2.5)(-5.5) = 13.75
2	5	2 – 3.5 = -1.5	5 – 7.5 = -2.5	(-1.5)(-2.5) = 3.75
3	6	3 – 3.5 = -0.5	6 – 7.5 = -1.5	(-0.5)(-1.5) = 0.75
4	10	4 – 3.5 = 0.5	10 – 7.5 = 2.5	(0.5)(2.5) = 1.25
5	11	5 – 3.5 = 1.5	11 – 7.5 = 3.5	(1.5)(3.5) = 5.25
6	11	6 – 3.5 = 2.5	11 – 7.5 = 3.5	(2.5)(3.5) = 8.75

$$SP=\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})=13.75+3.75+0.75+1.25+5.25+8.75=33.5$$

SP is positive.

POSITIVE SP (SP=33.5)

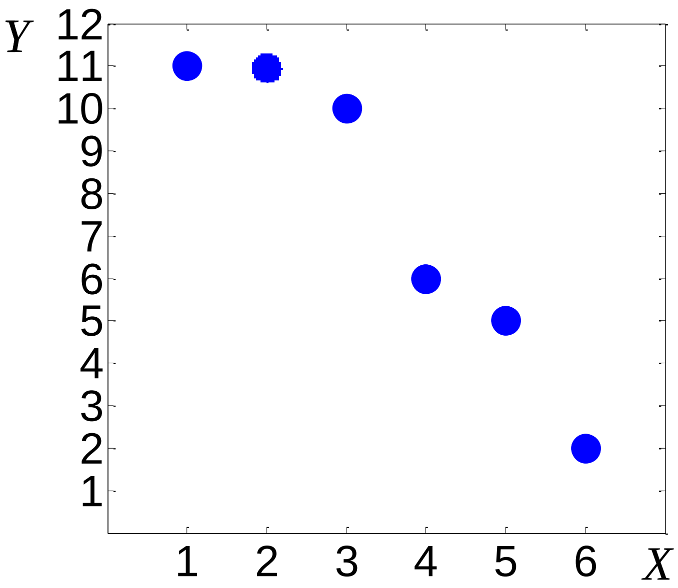
- *X* and *Y* tend to change in the same direction.
- As *X* increases, *Y* also tends to increase.
- When *X* is above (or below) the mean, *Y* also tends to be above (or below) the mean.



2. NEGATIVE RELATIONSHIP

- Let's calculate SP for the following data.

X	Y
1	11
2	11
3	10
4	6
5	5
6	2



2. NEGATIVE RELATIONSHIP

- Step 1: Calculate the mean for each variable

X	Y
1	11
2	11
3	10
4	6
5	5
6	2

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{11+11+10+6+5+2}{6} = 7.5$$

2. NEGATIVE RELATIONSHIP

- Step 2 : Calculate the deviation scores for each variable.

X	Y	$X - \bar{X}$	$Y - \bar{Y}$
1	11	1 – 3.5 = -2.5	11 – 7.5 = 3.5
2	11	2 – 3.5 = -1.5	11 – 7.5 = 3.5
3	10	3 – 3.5 = -0.5	10 – 7.5 = 2.5
4	6	4 – 3.5 = 0.5	6 – 7.5 = -1.5
5	5	5 – 3.5 = 1.5	5 – 7.5 = -2.5
6	2	6 – 3.5 = 2.5	2 – 7.5 = -5.5

2. NEGATIVE RELATIONSHIP

- Step 3 : Calculate the products of deviations.

X	Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$
1	11	1 – 3.5 = -2.5	11 – 7.5 = 3.5	(-2.5)(3.5) = -8.75
2	11	2 – 3.5 = -1.5	11 – 7.5 = 3.5	(-1.5)(3.5) = -5.25
3	10	3 – 3.5 = -0.5	10 – 7.5 = 2.5	(-0.5)(2.5) = -1.25
4	6	4 – 3.5 = 0.5	6 – 7.5 = -1.5	(0.5)(-1.5) = -0.75
5	5	5 – 3.5 = 1.5	5 – 7.5 = -2.5	(1.5)(-2.5) = -3.75
6	2	6 – 3.5 = 2.5	2 – 7.5 = -5.5	(2.5)(-5.5) = -13.75

2. NEGATIVE RELATIONSHIP

- Step 4 : Calculate the sum of products of deviations.

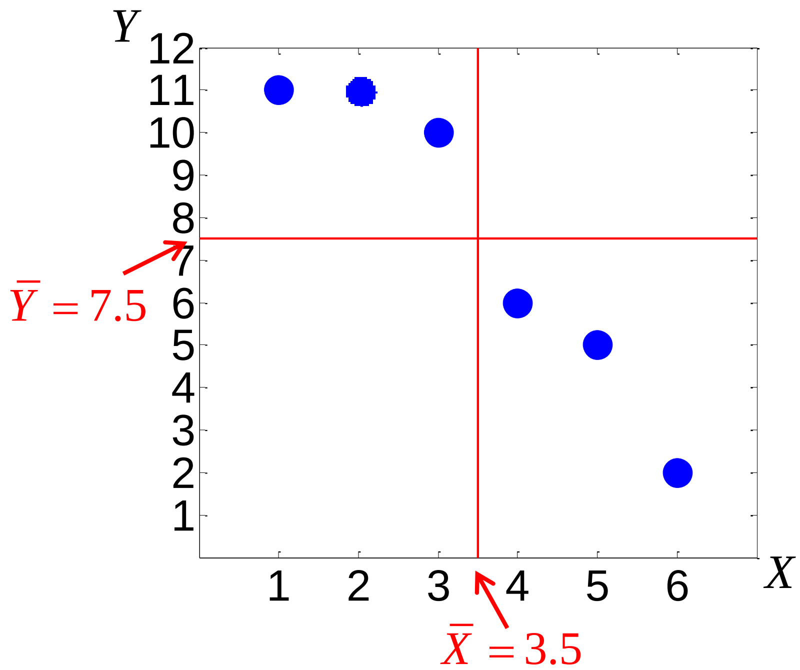
X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$
1	11	1 – 3.5 = -2.5	11 – 7.5 = 3.5	(-2.5)(3.5) = -8.75
2	11	2 – 3.5 = -1.5	11 – 7.5 = 3.5	(-1.5)(3.5) = -5.25
3	10	3 – 3.5 = -0.5	10 – 7.5 = 2.5	(-0.5)(2.5) = -1.25
4	6	4 – 3.5 = 0.5	6 – 7.5 = -1.5	(0.5)(-1.5) = -0.75
5	5	5 – 3.5 = 1.5	5 – 7.5 = -2.5	(1.5)(-2.5) = -3.75
6	2	6 – 3.5 = 2.5	2 – 7.5 = -5.5	(2.5)(-5.5) = -13.75

$$SP=\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})=(-8.75)+(-5.25)+(-1.25)$$
$$+(-0.75)+(-3.75)+(-13.75)=-33.5$$

SP is negative.

NEGATIVE SP (SP=-33.5)

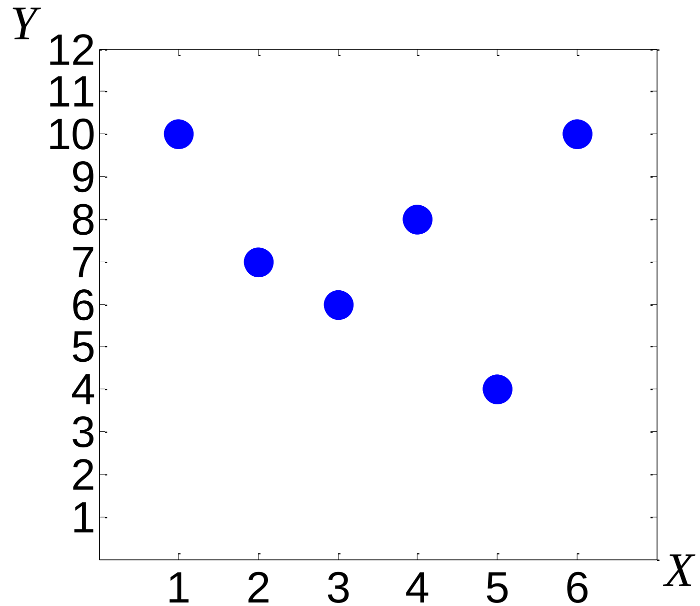
- *X* and *Y* tend to change in the opposite direction.
- As *X* increases, *Y* tends to decrease.
- When *X* is above (or below) the mean, *Y* tends to be below (or above) the mean.



3. NO RELATIONSHIP

- Let's calculate SP for the following data.

X	Y
1	10
2	7
3	6
4	8
5	4
6	10



3. NO RELATIONSHIP

- Step 1. Calculate the mean for each variable.

X	Y
1	10
2	7
3	6
4	8
5	4
6	10

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{1+2+3+4+5+6}{6} = 3.5$$

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{10+7+6+8+4+10}{6} = 7.5$$

3. NO RELATIONSHIP

- Step 2. Calculate the deviation scores for each variable.

X	Y	$X - \bar{X}$	$Y - \bar{Y}$
1	10	1 – 3.5 = -2.5	10 – 7.5 = 2.5
2	7	2 – 3.5 = -1.5	7 – 7.5 = -0.5
3	6	3 – 3.5 = -0.5	6 – 7.5 = -1.5
4	8	4 – 3.5 = 0.5	8 – 7.5 = 0.5
5	4	5 – 3.5 = 1.5	4 – 7.5 = -3.5
6	10	6 – 3.5 = 2.5	10 – 7.5 = 2.5

3. NO RELATIONSHIP

- Step 3. Calculate the products of deviations.

X	Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$
1	10	1 – 3.5 = -2.5	10 – 7.5 = 2.5	(-2.5)(2.5) = -6.25
2	7	2 – 3.5 = -1.5	7 – 7.5 = -0.5	(-1.5)(-0.5) = 0.75
3	6	3 – 3.5 = -0.5	6 – 7.5 = -1.5	(-0.5)(-1.5) = 0.75
4	8	4 – 3.5 = 0.5	8 – 7.5 = 0.5	(0.5)(0.5) = 0.25
5	4	5 – 3.5 = 1.5	4 – 7.5 = -3.5	(1.5)(-3.5) = -5.25
6	10	6 – 3.5 = 2.5	10 – 7.5 = 2.5	(2.5)(2.5) = 6.25

3. NO RELATIONSHIP

- Step 4. Calculate the sum of products of deviations.

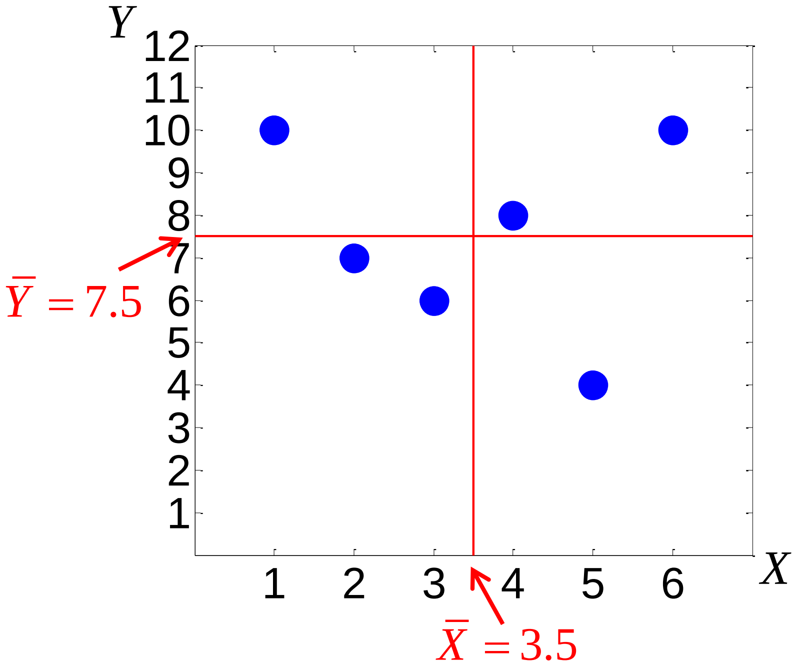
X	Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})(Y - \bar{Y})$
1	10	1 – 3.5 = -2.5	10 – 7.5 = 2.5	(-2.5)(2.5) = -6.25
2	7	2 – 3.5 = -1.5	7 – 7.5 = -0.5	(-1.5)(-0.5) = 0.75
3	6	3 – 3.5 = -0.5	6 – 7.5 = -1.5	(-0.5)(-1.5) = 0.75
4	8	4 – 3.5 = 0.5	8 – 7.5 = 0.5	(0.5)(0.5) = 0.25
5	4	5 – 3.5 = 1.5	4 – 7.5 = -3.5	(1.5)(-3.5) = -5.25
6	10	6 – 3.5 = 2.5	10 – 7.5 = 2.5	(2.5)(2.5) = 6.25

$$SP = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = (-6.25) + 0.75 + 0.75 + 0.25 + (-5.25) + 6.25 = -3.5$$

SP is close to 0.

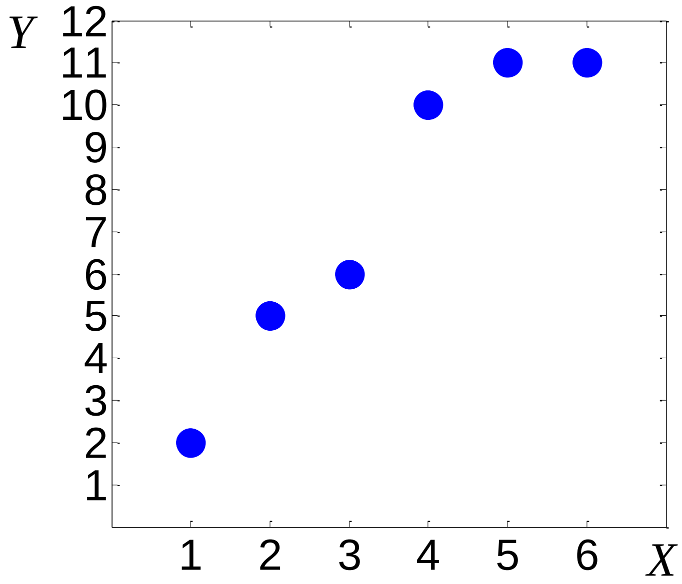
SP CLOSE TO ZERO (SP=-3.5)

- *X* and *Y* do not co-vary in any single direction.
- *X* and *Y* are not linearly related.



LIMITATION OF SP

- SP is the sum and thus affected by the number of observations (or persons).



- If each observation is duplicated, the SP will be doubled. However, this does not indicate that the strength of the relationship between X and Y is doubled.

COVARIANCE

- Covariance (공분산) is the *average* product of deviations.

  - Population covariance

$$\sigma_{XY} = \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{N}$$

  - Sample covariance

$$s_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n - 1}$$

There is an *n* formula as well. However, from now on, we will only use *n*-1 formula for samples.

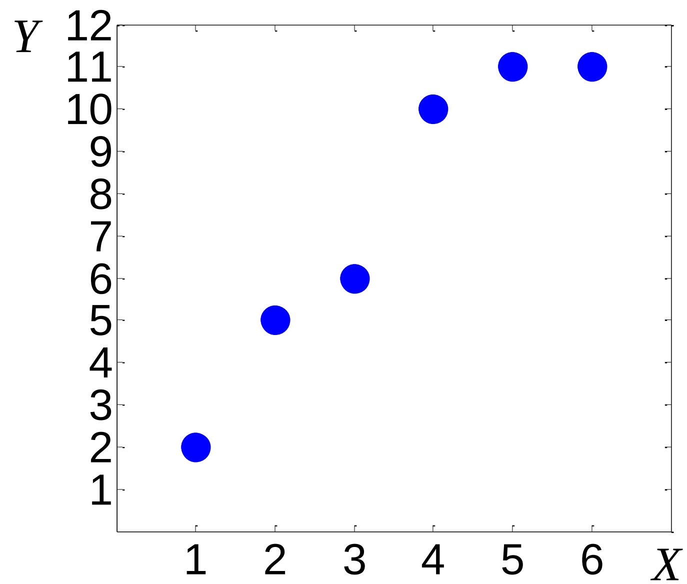
PROPERTIES OF COVARIANCE

- If X and Y are positively related, covariance will be positive.
- If X and Y are negatively related, covariance will be negative.
- If X and Y are not linearly related, covariance will be close to 0.
- These properties will be demonstrated using a sample data. However, there properties will hold for the population as well.

1. POSITIVE RELATIONSHIP

- SP has been already calculated for the following data (SP = 33.5).

X	Y
1	2
2	5
3	6
4	10
5	11
6	11



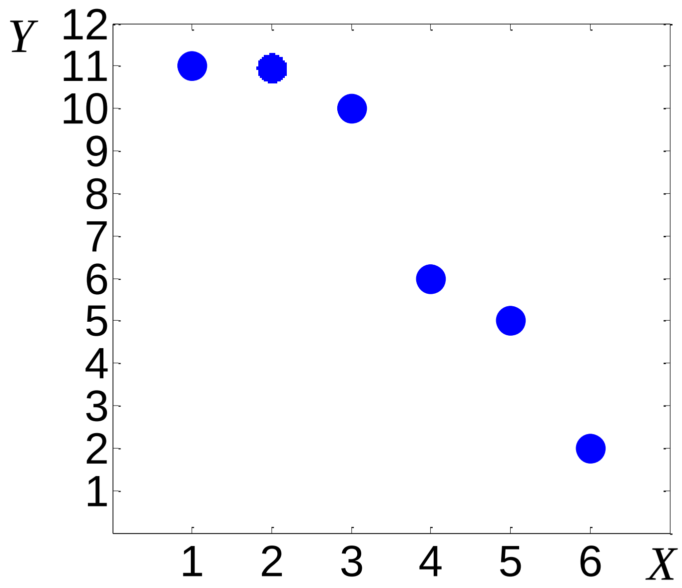
$$s_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{SP}{n-1} = \frac{33.5}{5} = 6.7$$

Covariance is positive.

2. NEGATIVE RELATIONSHIP

- SP has been already calculated for the following data (SP = -33.5).

X	Y
1	11
2	11
3	10
4	6
5	5
6	2



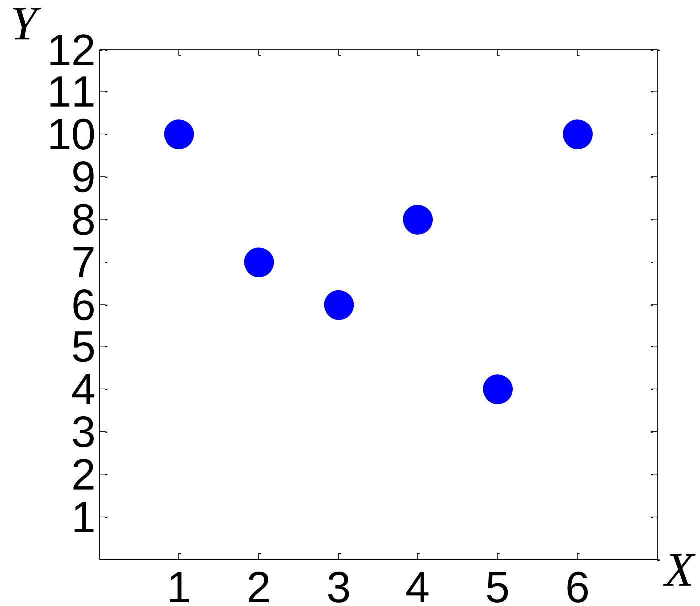
$$s_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{SP}{n-1} = \frac{-33.5}{5} = -6.7$$

Covariance is negative.

3. NO RELATIONSHIP

- SP has been already calculated for the following data (SP = -3.5).

X	Y
1	10
2	7
3	6
4	8
5	4
6	10



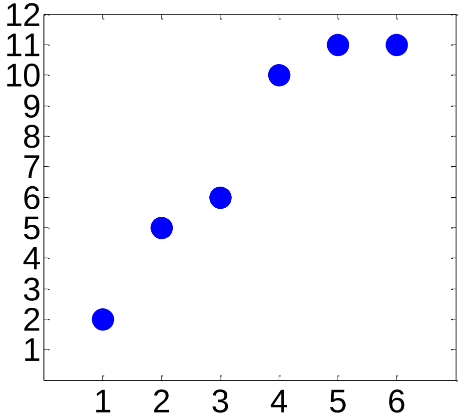
$$s_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{SP}{n-1} = \frac{-3.5}{5} = -0.7$$

Covariance is close to 0.

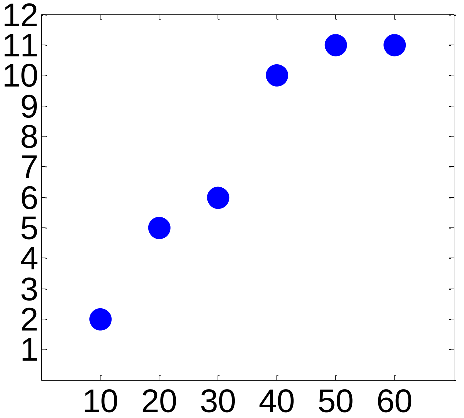
LIMITATION OF COVARIANCE

- The size of covariance is not easy to interpret.
- The units of measurements affect the size of covariance.
- What happens if the unit of measurement changes?

X	Y
1	2
2	5
3	6
4	10
5	11
6	11



$10X$	Y
10	2
20	5
30	6
40	10
50	11
60	11



- Original measurement unit

$$s_{XY} = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1} = \frac{33.5}{5} = 6.7$$

- New measurement unit (X is multiplied by 10)

$$s_{10XY} = \frac{\sum_{i=1}^{n}(10X_i - 10\bar{X})(Y_i - \bar{Y})}{n-1} = \frac{\sum_{i=1}^{n} 10(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}$$

$$= 10\left(\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}\right) = (10)\frac{33.5}{5} = 67$$

- The covariance becomes 10 times larger even though the strength of the relationship between X and Y didn't change!
- A larger covariance does not necessarily indicate a stronger relationship if the measurement units are different.

PEARSON'S CORRELATION COEFFICIENT

- Pearson's correlation coefficient (피어슨 상관 계수) is a standardized covariance that is not affected by unit of measurement.

- Population correlation

$$\boxed{\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y}}$$

- σ_X: population standard deviation of X
- σ_Y: population standard deviation of Y

- Sample correlation

$$\boxed{r = \frac{s_{XY}}{s_X s_Y}}$$

- s_X: sample standard deviation of X
- s_Y: sample standard deviation of Y

- Population correlation

$$\rho = \frac{\sigma_{XY}}{\sigma_X \sigma_Y} = \frac{\frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{N}}{\sqrt{\frac{\sum_{i=1}^{N}(X_i - \mu_X)^2}{N}}\sqrt{\frac{\sum_{i=1}^{N}(Y_i - \mu_Y)^2}{N}}}$$

$$= \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{N}(X_i - \mu_X)^2}\sqrt{\sum_{i=1}^{N}(Y_i - \mu_Y)^2}}$$

$$= \frac{\sum_{i=1}^{N}(X_i - \mu_X)(Y_i - \mu_Y)}{\sqrt{\sum_{i=1}^{N}(X_i - \mu_X)^2 \sum_{i=1}^{N}(Y_i - \mu_Y)^2}} = \frac{SP}{\sqrt{SS_X SS_Y}}$$

- ρ is read "rho" and Greek lower case letter r.

- Sample correlation

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{\frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{n-1}}{\sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}}\sqrt{\frac{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}{n-1}}}$$

$$= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2}\sqrt{\sum_{i=1}^{n}(Y_i - \bar{Y})^2}}$$

$$= \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sqrt{\sum_{i=1}^{n}(X_i - \bar{X})^2 \sum_{i=1}^{n}(Y_i - \bar{Y})^2}} = \frac{SP}{\sqrt{SS_X SS_Y}}$$

PEARSON'S CORRELATION COEFFICIENT

- Pearson's correlation coefficient takes a value between -1 and 1.

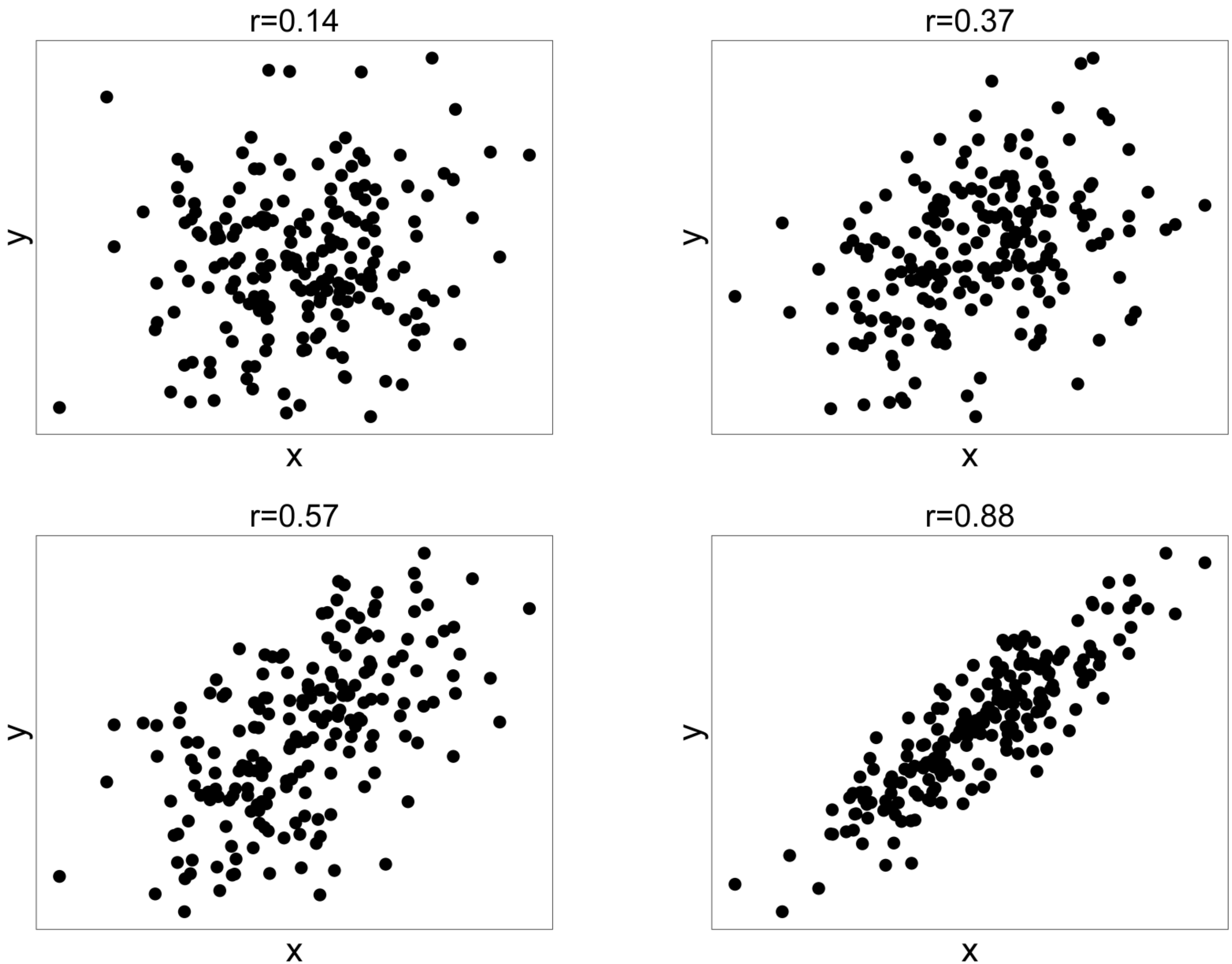
  - The <u>sign</u> of the correlation coefficient indicates the direction of the relationship.
    - Positive correlation : positive relationship
    - Negative correlation : negative relationship

  - The <u>absolute value</u> of the correlation coefficient indicates the strength of the relationship.
    - Absolute value close to 1 : strong relationship
    - Absolute value close to 0: weak relationship

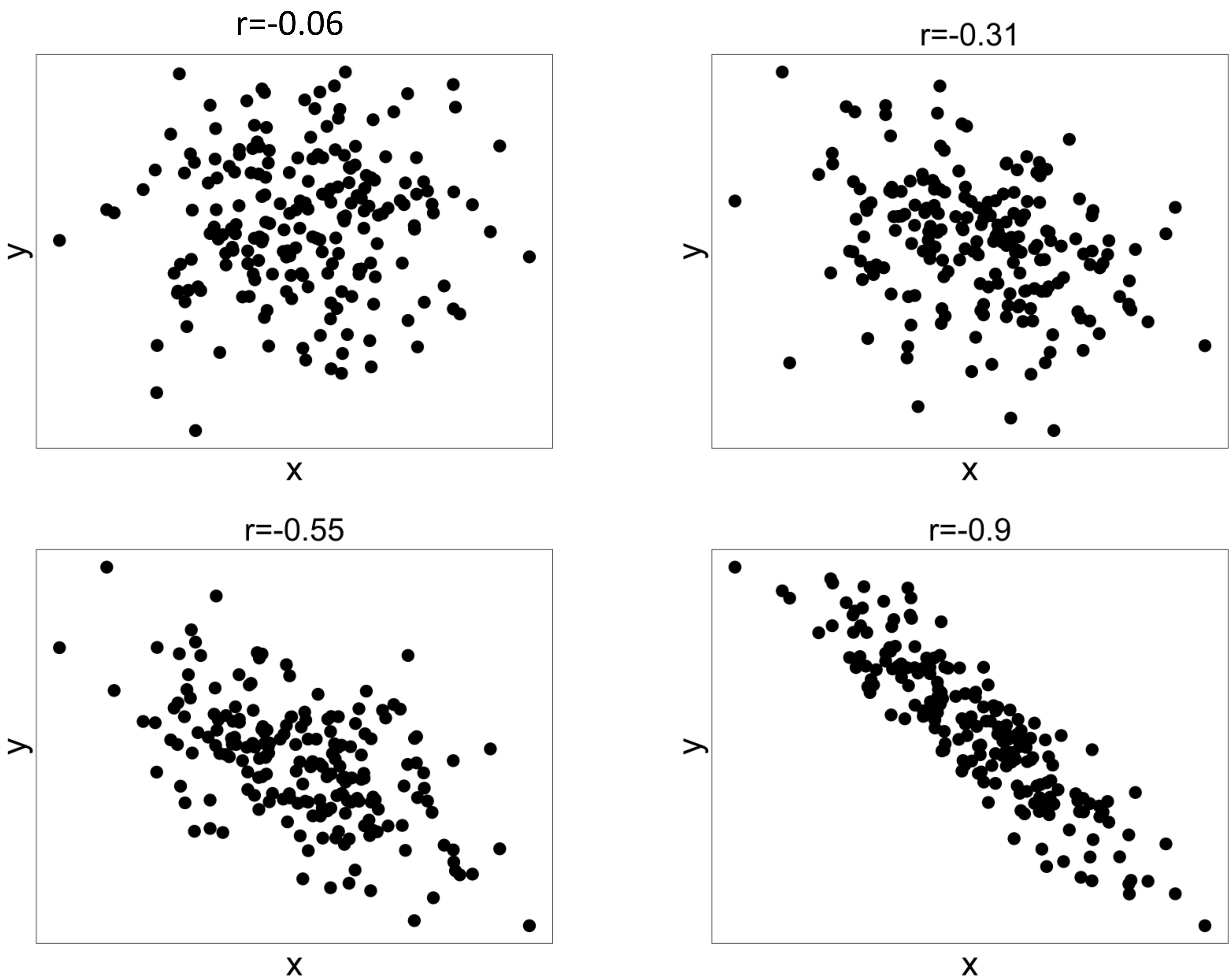
A RULE OF THUMB

- Cohen (1988)
  - Pearson's correlation = .10 : weak association
  - Pearson's correlation = .30 : moderate association
  - Pearson's correlation ≥ .50 : strong association

POSITIVE RELATIONS



NEGATIVE RELATIONS



EXAMPLE

- Calculate Pearson's correlation coefficient between X and Y for the following sample data.

X	Y
0	1
2	7
2	6
4	9
7	12

EXAMPLE

- Step 1. Calculate the mean for each variable.

X	Y
0	1
2	7
2	6
4	9
7	12

$$\bar{X} = \frac{\sum_{i=1}^{n} X_i}{n} = \frac{0+2+2+4+7}{5} = 3$$

$$\bar{Y} = \frac{\sum_{i=1}^{n} Y_i}{n} = \frac{1+7+6+9+12}{5} = 7$$

EXAMPLE

- Step 2. Calculate the deviation scores for each variable.

X	Y	$X - \bar{X}$	$Y - \bar{Y}$
0	1	0 – 3 = -3	1 – 7 = -6
2	7	2 – 3 = -1	7 – 7 = 0
2	6	2 – 3 = -1	6 – 7 = -1
4	9	4 – 3 = 1	9 – 7 = 2
7	12	7 – 3 = 4	12 – 7 = 5

EXAMPLE

- Step 3. Calculate the sum of deviation scores for each variable.

X	Y	$X - \bar{X}$	$Y - \bar{Y}$	$(X - \bar{X})^2$	$(Y - \bar{Y})^2$
0	1	0 – 3 = -3	1 – 7 = -6	9	36
2	7	2 – 3 = -1	7 – 7 = 0	1	0
2	6	2 – 3 = -1	6 – 7 = -1	1	1
4	9	4 – 3 = 1	9 – 7 = 2	1	4
7	12	7 – 3 = 4	12 – 7 = 5	16	25

$$SS_X = \sum_{i=1}^{n} (X_i - \bar{X})^2 = 9 + 1 + 1 + 1 + 16 = 28$$

$$SS_Y = \sum_{i=1}^{n} (Y_i - \bar{Y})^2 = 36 + 0 + 1 + 4 + 25 = 66$$

EXAMPLE

- Step 4. Calculate the products of deviations.

X	Y	$X-\bar{X}$	$Y-\bar{Y}$	$(X-\bar{X})(Y-\bar{Y})$
0	1	0 – 3 = -3	1 – 7 = -6	(-3)(-6)=18
2	7	2 – 3 = -1	7 – 7 = 0	(-1)(0)=0
2	6	2 – 3 = -1	6 – 7 = -1	(-1)(-1)=1
4	9	4 – 3 = 1	9 – 7 = 2	(1)(2)=2
7	12	7 – 3 = 4	12 – 7 = 5	(4)(5)=20

$$SP = \sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y}) = 18 + 0 + 1 + 2 + 20 = 41$$

EXAMPLE

- Step 5. Calculate the Pearson's correlation coefficient.

$$r = \frac{s_{XY}}{s_X s_Y} = \frac{SP}{\sqrt{SS_X SS_Y}} = \frac{41}{\sqrt{(28)(66)}} = 0.95$$

- Step 6. Interpret the correlation.
  - The two variables (X and Y) are positively and very strongly related (r = 0.95).

AN IMPORTANT PROPERTY OF CORRELATION

- Pearson's correlation coefficient is not affected by units of measurement. The correlation between X and Y for the following two sets of data will be equal.

X	Y
0	1
2	7
2	6
4	9
7	12

$1{,}000X$	$10{,}000Y$
0	10,000
2,000	70,000
2,000	60,000
4,000	90,000
7,000	120,000

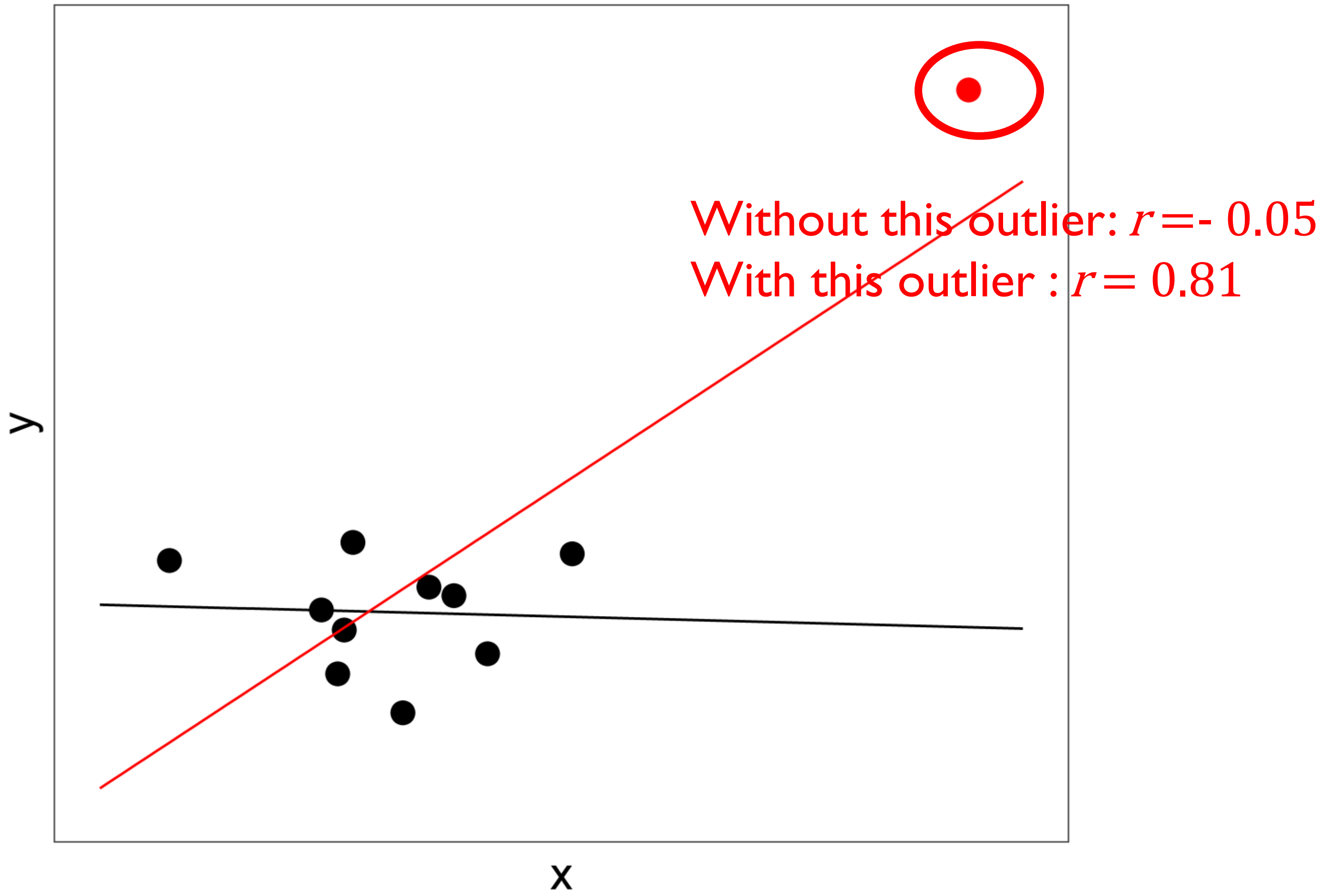
FACTORS AFFECTING CORRELATION

- The following factors might affect Pearson's correlation coefficient. You should carefully consider these factors when interpreting Pearson's correlation coefficient.

  - Outliers
  - Nonlinear relationship
  - Restriction of range

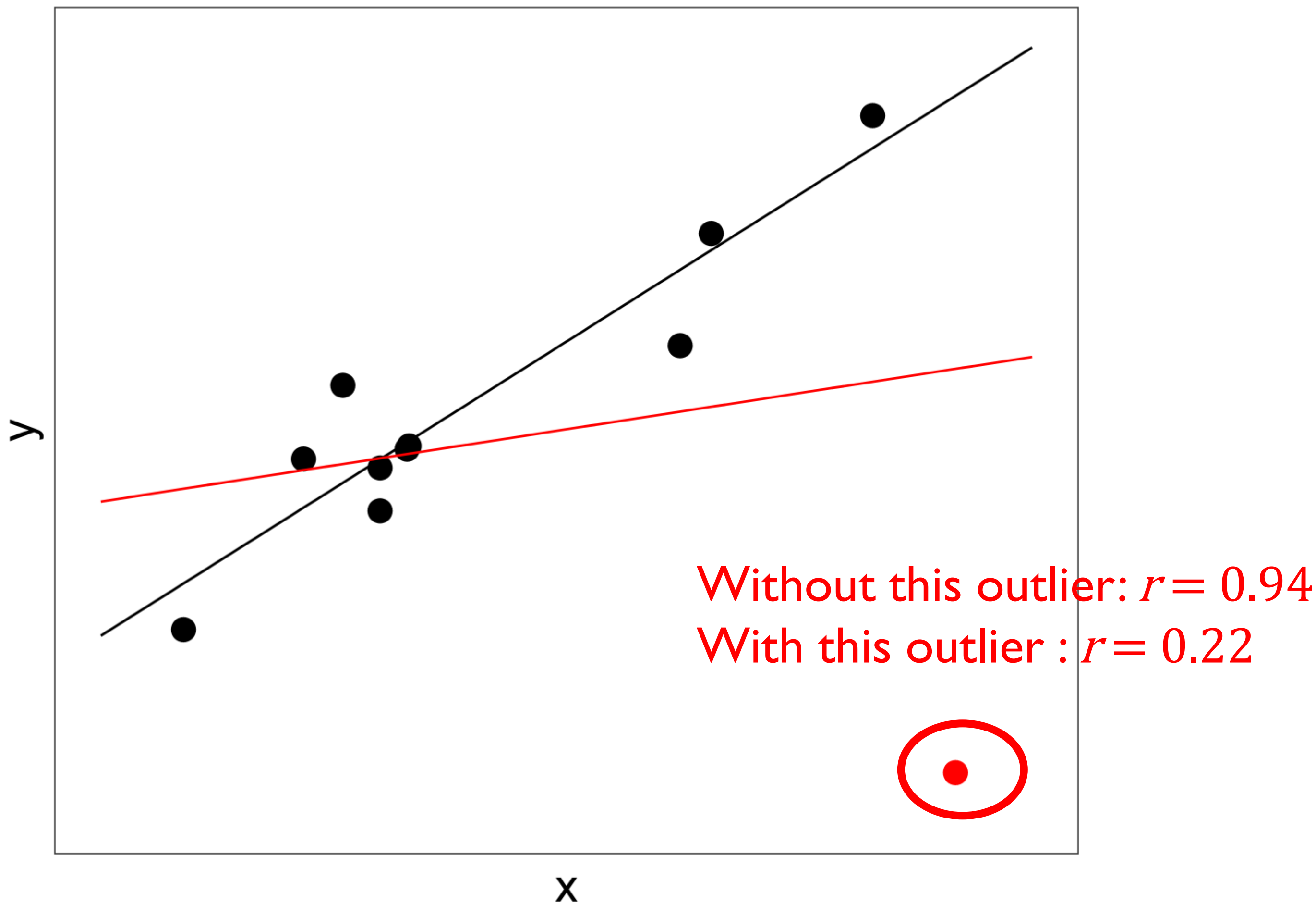
1. OUTLIERS

- Pearson's correlation coefficient can increase or decrease due to an outlier.



1. OUTLIERS

- Pearson's correlation coefficient can increase or decrease due to an outlier.

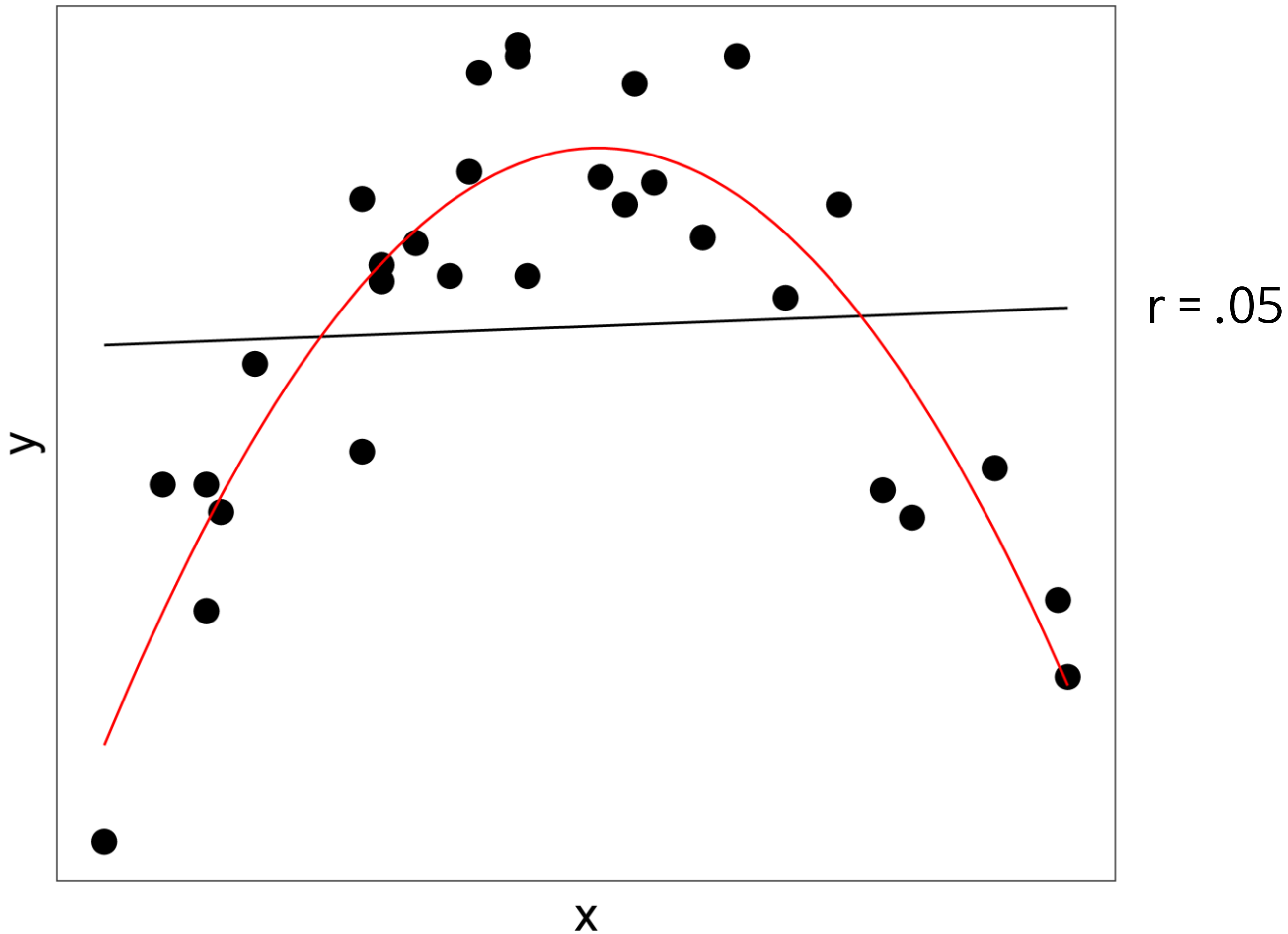


2. NONLIEAR RELATIONSHIP

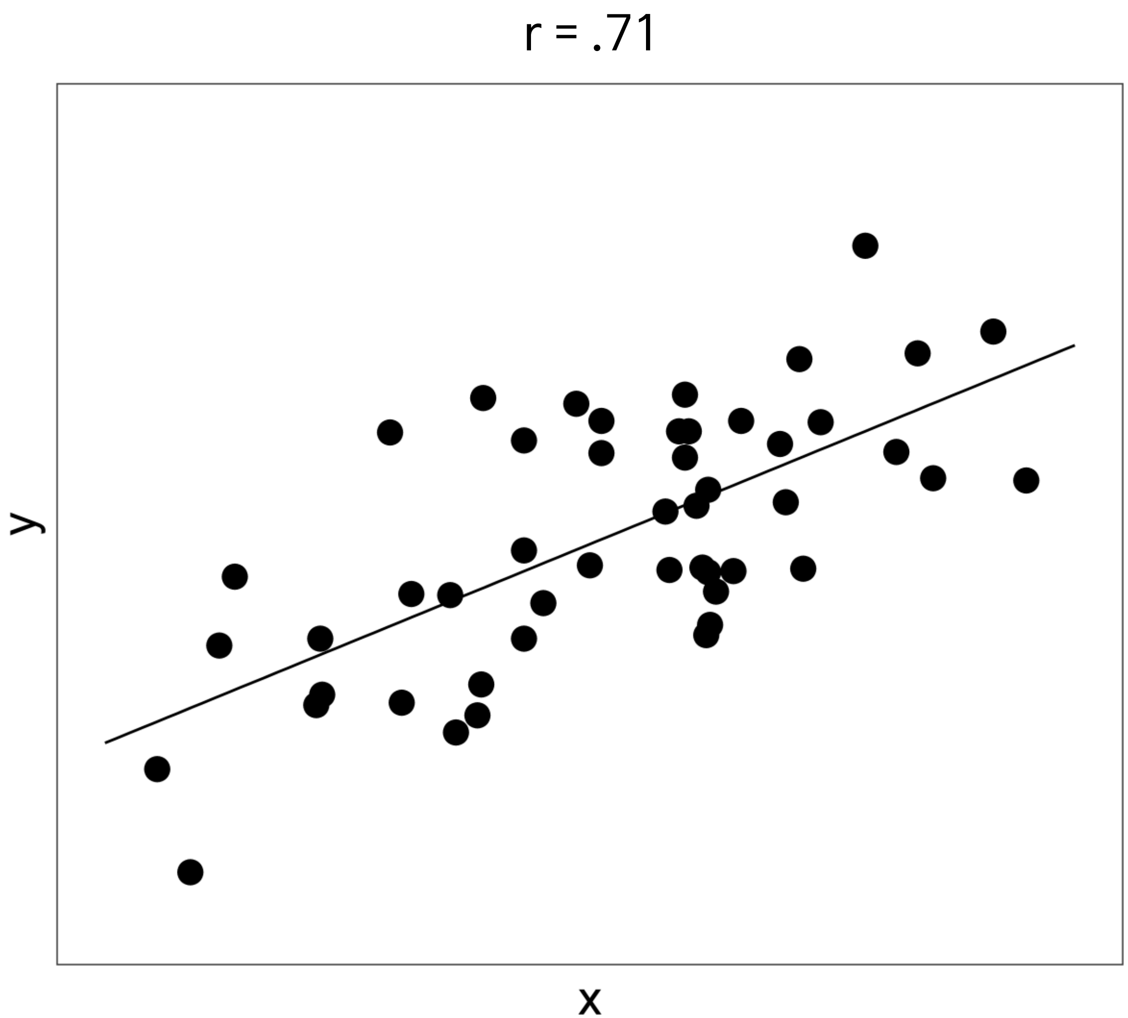
- Pearson's correlation coefficient indicates only the linear relationship between the two variables.

- If the two variables have a nonlinear relationship, Pearson's correlation coefficient can be close to zero even though the two variables are strongly related.

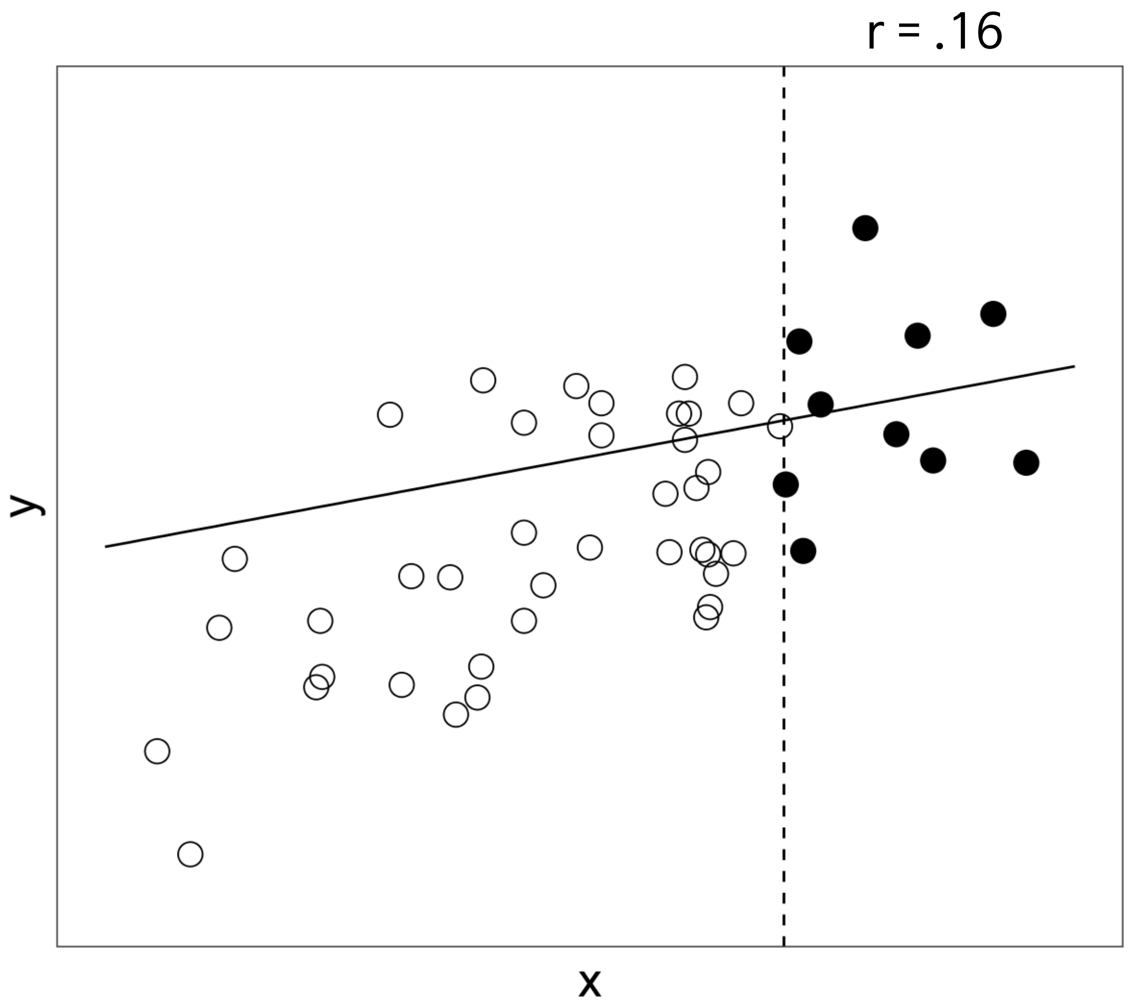
2. NONLIEAR RELATIONSHIP



3. RESTRICTION OF RANGE



3. RESTRICTION OF RANGE



SUMMARY

- How to describe the relationship between two variables?
  - Scatterplot
  - Covariance
  - Correlation

- Factors affecting correlation
  - Outliers
  - Nonlinear relationship
  - Restriction of range
    - When you interpret Pearson's correlation coefficient, you should carefully consider whether these factors are affecting the correlation.