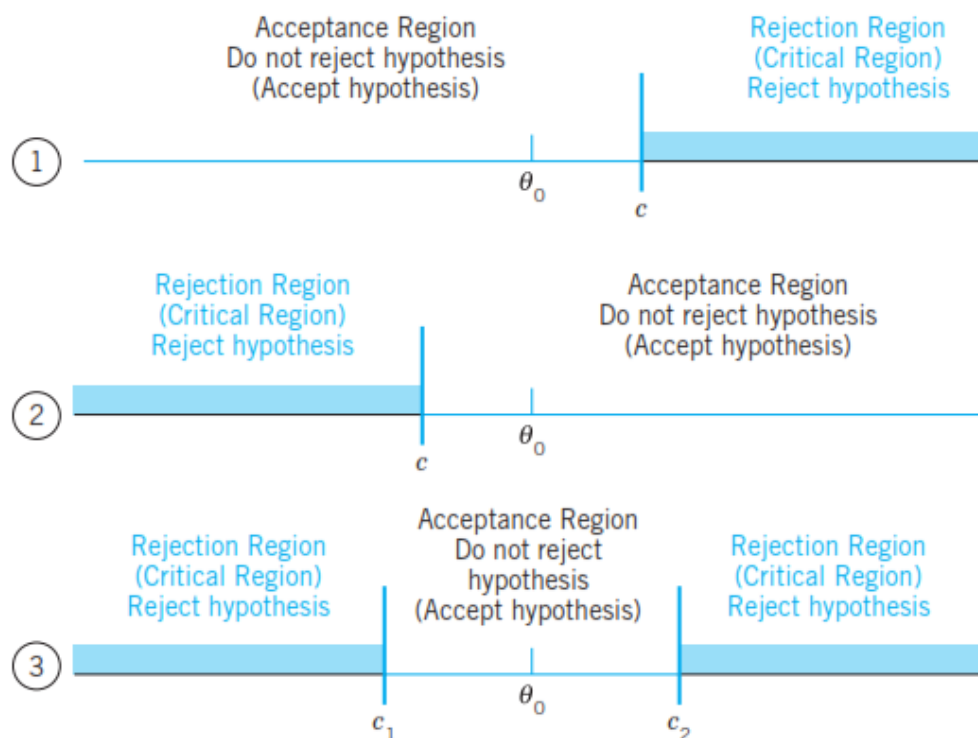


# Module 6, Math-II

## Hypothesis Testing

We want to buy 100 coils of a certain kind of wire, provided we can verify the manufacturer's claim that the wire has a breaking limit  $\mu = \mu_0 = 200$  lb (or more). This is a test of the **hypothesis** (also called *null hypothesis*)  $\mu = \mu_0 = 200$ . We shall not buy the wire if the (statistical) test shows that actually  $\mu = \mu_1 < \mu_0$ , the wire is weaker, the claim does not hold.  $\mu_1$  is called the **alternative** (or *alternative hypothesis*) of the test. We shall **accept** the hypothesis if the test suggests that it is true, except for a small error probability  $\alpha$ , called the **significance level** of the test. Otherwise we **reject** the hypothesis. Hence  $\alpha$  is the probability of rejecting a hypothesis although it is true. The choice of  $\alpha$  is up to us. 5% and 1% are popular values.



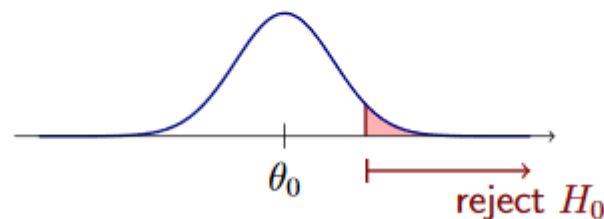
### Steps in hypothesis testing

1. State the null and alternate hypothesis
2. Choose a significance level  $\alpha$
3. Choose the test statistic and establish the critical region
4. Collect the sample and compute the test statistic.  
If the test statistic is in the critical region, reject  $H_0$ .  
Otherwise, do not reject  $H_0$ .

## Hypothesis testing for mean when sample is small or sample variance is unknown .

- $H_0: \theta = \theta_0$
- $H_1: \theta > \theta_0$ .

This is a one-tailed test with the critical region in the right-tail of the test statistic  $X$ .



Q)

If a sample of 25 tires of a certain kind has a mean life of 37,000 miles and a standard deviation of 5000 miles, can the manufacturer claim that the true mean life of such tires is greater than 35,000 miles? Set up and test a corresponding hypothesis at the 5% level, assuming normality.

Ans:

Step-1

We are given the sample of length  $n = 25$ , with both mean and standard deviation being known with values  $\mu = 37000$  and  $\sigma = 5000$  miles, respectively. We need to see if the manufacturer can really claim that the tires' true mean life is greater than 35000 miles. Hence, we state our hypotheses as following:

$$H_0 : \mu_0 = 35000$$

$$H_1 : \mu > \mu_0$$

Step-2

We are given the significance level of 5%.

So  $\alpha=0.05$

Step-3

Test Statistics is  $t = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}$ .

Critical value of t from t table with  $n-1=24$  df and 0.05 level is 1.71

So if calculated t value > critical t value, we must reject the null hypothesis

Step -4

$$\text{Now } t = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{(37000 - 35000)}{5000 / \sqrt{25}} = 2$$

Since Calculated t value > critical t value, we must reject the null hypothesis at 0.05 level of significance.

That means manufacturer's claim is true.

Q)

It is claimed that a vacuum cleaner expends 46 kWh per year. A random sample of 12 homes indicates that vacuum cleaners expend an average of 42 kWh per year with (sample) standard deviation 11.9 kWh. At a 0.05 level of significance, does this suggest that, on average, vacuum cleaner expend less than 46 kWh per year? Assume the population to be normally distributed.

Step-1:

$$H_0 : \mu = 46$$

$$H_1 : \mu < 46$$

Step-2:

$$\alpha = 0.05$$

Step-3:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}}, \text{Critical value of t from t table with } n-1=11 \text{ df and 0.05 level is } -1.8$$

So if calculated t value < critical t value, we must reject the null hypothesis

Step -4:

$$t = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{42 - 46}{11.9 / \sqrt{12}} = -1.16$$

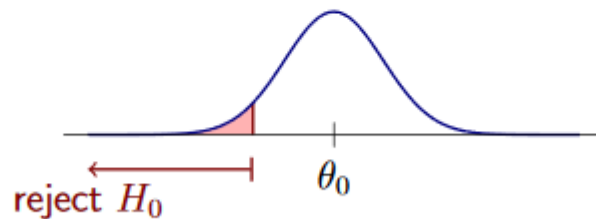
Since Calculated t value does not less than critical t value, we cannot reject the null hypothesis at 0.05 level of significance.

Q)

A quality control engineer finds that a sample of 100 light bulbs had an average life-time of 470 hours. Assuming a population standard deviation of  $\sigma = 25$  hours, test whether the population mean is 480 hours vs. the alternative hypothesis  $\mu < 480$  at a significance level of  $\alpha = 0.05$ .

- $H_0: \theta = \theta_0$
- $H_1: \theta < \theta_0$ ,

in which the critical region is in the left-tail.



#### Hypothesis testing for mean when sample variance is known

Step - 1 :

$$H_0 : \mu = 480$$

$$H_1 : \mu < 480$$

Step - 2 :

$$\alpha = 0.05$$

Step - 3 :

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}, \text{Critical value of } z \text{ from } z \text{ table } 0.05 \text{ level is } -1.645$$

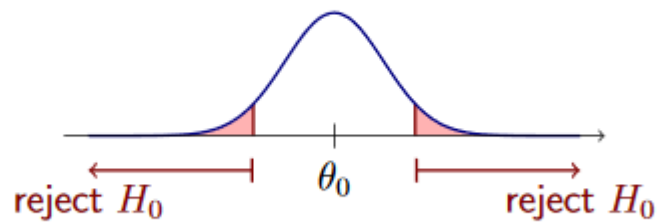
So if calculated  $z$  value  $<$  critical  $z$  value , we must reject the null hypothesis

Step - 4 :

$$z = \frac{\bar{x} - \mu}{s / \sqrt{n}} = \frac{470 - 480}{25 / \sqrt{100}} = -4$$

Since Calculated  $z$  value does not less than critical  $z$  value , we cannot reject the null hypothesis at 0.05 level of significance.

- $H_0: \theta = \theta_0$
- $H_1: \theta \neq \theta_0,$



Q)

A batch of 100 resistors have an average of 102 Ohms. Assuming a population standard deviation of 8 Ohms, test whether the population mean is 100 Ohms at a significance level of  $\alpha = 0.05$ .

Consider a production line of resistors that are supposed to be 100 Ohms. Assume  $\sigma = 8$ . So, the hypotheses are:

- $H_0: \mu = 100$
- $H_1: \mu \neq 100,$

Step-1:

$$H_0 : \mu = 100$$

$$H_1 : \mu \neq 100$$

Step-2:

$$\alpha = 0.05$$

Step-3:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}}, \text{Critical value of } z \text{ from } z \text{ table at } 0.05 \text{ level are } -1.96 \text{ and } 1.96$$

So if calculated  $z$  value  $<$  critical  $z$  value or, calculated  $z$  value  $>$  critical  $z$  value, we must reject the null hypothesis

Step-4:

$$z = \frac{\bar{x} - \mu}{\sigma / \sqrt{n}} = \frac{102 - 100}{8 / \sqrt{100}} = 2.5$$

Since Calculated  $z$  value  $>$  critical  $z$  value, we must reject the null hypothesis at 0.05 level of significance.

### **Hypothesis Testing for single variance**

Q)

The television habits of 30 children were observed. The sample mean was found to be 48.2 hours per week, with a standard deviation of 12.4 hours per week. Test the claim that the standard deviation was at least 16 hours per week.

Step-1:

$$H_0: \sigma = 16$$

$$H_1: \mu < 16$$

Step-2:

$$\alpha = 0.05$$

Step-3:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}, \text{Critical value of } \chi^2 \text{ from } \chi^2 \text{ table at } 0.05 \text{ level with } n-1 = 29 \text{ df is } 17.7$$

So if calculated  $\chi^2$  value < critical  $\chi^2$  value

we must reject the null hypothesis

Step-4:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{29 * (12.4)^2}{16^2} = 17.418$$

Since Calculated  $\chi^2$  value less than critical  $\chi^2$  value, we must reject the null hypothesis at 0.05 level of significance.

so the variation in television watching was less than 16 hours per week.

Q)

The data 159.9, 187.2, 180.1, 158.1, 225.5, 163.7, and 217.3 consists of the weights, in pounds, of a random sample of seven individuals taken from a population that is normally distributed. The variance of this sample is given by 753.04. Let us test the null hypothesis  $H_0: \sigma^2 = 750.0$  against the alternative hypothesis  $H_1: \sigma^2 \neq 750.0$  at a level of significance of .3.

Step-1:

$$H_0: \sigma = 750$$

$$H_1: \mu \neq 750$$

Step-2:

$$\alpha = 0.05$$

Step-3:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2}, \text{Critical value of } \chi^2 \text{ from } \chi^2 \text{ table at } 0.05 \text{ level with } n-1=6 \text{ df are } 1.24, 14.45$$


So if calculated  $\chi^2$  value < critical  $\chi^2$  value or calculated  $\chi^2$  value > critical  $\chi^2$  value we must reject the null hypothesis

Step-4:

$$\chi^2 = \frac{(n-1)s^2}{\sigma^2} = \frac{6*(753.04)}{750} = 6.024$$

Since Calculated  $\chi^2$  value *neither* less than critical  $\chi^2$  value, nor Calculated  $\chi^2$  value *greater* than critical  $\chi^2$  value, so we cannot reject the null hypothesis at 0.05 level of significance.

## **ERROR IN HYPOTHESIS TESTING**

	Null hypothesis is TRUE	Null hypothesis is FALSE
Reject null hypothesis	Type I Error (False positive)	Correct outcome! (True positive)
Fail to reject null hypothesis	Correct outcome! (True negative)	Type II Error (False negative)



## Linear Regression

We have seen equation like below in maths classes.  $y$  is the output we want.  $x$  is the input variable.  $c$  = constant and  $a$  is the slope of the line.

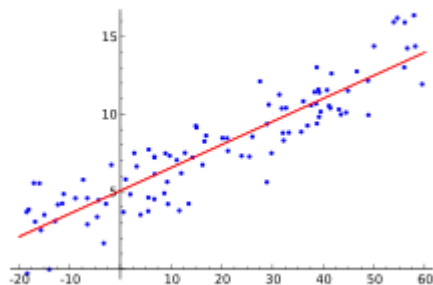
$$y = c + ax$$

$c$  = constant

$a$  = slope

The output varies linearly based upon the input.

- $y$  is the output which is determined by input  $x$ . How much value of  $x$  has impact on  $y$  is determined by “ $a$ ”. In the two dimensional graph having axis ‘ $x$ ’ and ‘ $y$ ’, ‘ $a$ ’ is the slope of the line.
- ‘ $c$ ’ is the constant (value of  $y$  when  $x$  is zero).



From a given sample  $(x_1, y_1), \dots, (x_n, y_n)$  we shall now determine a straight line by least squares. We write the line as

$$y = \beta_0 + \beta_1 x$$

$$\beta_1 = \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2}$$

and then

$$\beta_0 = \frac{\sum_{i=1}^n y_i}{n} - \beta_1 \frac{\sum_{i=1}^n x_i}{n} = \bar{y} - \beta_1 \bar{x}$$

Q)

The manager of a car plant wishes to investigate how the plant's electricity usage depends upon the plant's production.

The linear model

$$y = \beta_0 + \beta_1 x$$

will allow a month's electrical usage to be estimated as a function of the month's production.

	Production (\$ million)	Electricity usage (million kWh)
January	4.51	2.48
February	3.58	2.26
March	4.31	2.47
April	5.06	2.77
May	5.64	2.99
June	4.99	3.05
July	5.29	3.18
August	5.83	3.46
September	4.70	3.03
October	5.61	3.26
November	4.90	2.67
December	4.20	2.53

Ans:

For this example  $n = 12$  and

$$\sum_{i=1}^{12} x_i = 4.51 + \dots + 4.20 = 58.62$$

$$\sum_{i=1}^{12} y_i = 2.48 + \dots + 2.53 = 34.15$$

$$\sum_{i=1}^{12} x_i^2 = 4.51^2 + \dots + 4.20^2 = 291.2310$$

$$\sum_{i=1}^{12} y_i^2 = 2.48^2 + \dots + 2.53^2 = 98.6967$$

$$\sum_{i=1}^{12} x_i y_i = (4.51 \times 2.48) + \dots + (4.20 \times 2.53) = 169.2532$$

The estimates of the slope parameter and the intercept parameter :

$$\begin{aligned}\beta_1 &= \frac{n \sum_{i=1}^n x_i y_i - (\sum_{i=1}^n x_i)(\sum_{i=1}^n y_i)}{n \sum_{i=1}^n x_i^2 - (\sum_{i=1}^n x_i)^2} \\ &= \frac{(12 \times 169.2532) - (58.62 \times 34.15)}{(12 \times 291.2310) - 58.62^2} = 0.49883 \\ \beta_0 &= \bar{y} - \beta_1 \bar{x} = \frac{34.15}{12} - (0.49883 \times \frac{58.62}{12}) = 0.4090\end{aligned}$$

The fitted regression line :

$$y = \beta_0 + \beta_1 x = 0.409 + 0.499x$$

Q)

Example: Sam found how many **hours of sunshine** vs how many **ice creams** were sold at the shop from Monday to Friday:



"x" Hours of Sunshine	"y" Ice Creams Sold
2	4
3	5
5	7
7	10
9	15

Let us find the best **m** (slope) and **b** (y-intercept) that suits that data

$$y = mx + b$$

**Step 1:** For each (x,y) calculate  $x^2$  and  $xy$ :

<b>x</b>	<b>y</b>	<b><math>x^2</math></b>	<b>xy</b>
2	4	4	8
3	5	9	15
5	7	25	35
7	10	49	70
9	15	81	135

**Step 2:** Sum  $x$ ,  $y$ ,  $x^2$  and  $xy$  (gives us  $\Sigma x$ ,  $\Sigma y$ ,  $\Sigma x^2$  and  $\Sigma xy$ ):

<b>x</b>	<b>y</b>	<b><math>x^2</math></b>	<b>xy</b>
2	4	4	8
3	5	9	15
5	7	25	35
7	10	49	70
9	15	81	135
<b><math>\Sigma x</math>: 26</b>	<b><math>\Sigma y</math>: 41</b>	<b><math>\Sigma x^2</math>: 168</b>	<b><math>\Sigma xy</math>: 263</b>

Also **N** (number of data values) = 5

**Step 3:** Calculate Slope **m**:

$$\begin{aligned} m &= \frac{N \sum(xy) - \sum x \sum y}{N \sum(x^2) - (\sum x)^2} \\ &= \frac{5 \times 263 - 26 \times 41}{5 \times 168 - 26^2} \\ &= \frac{1315 - 1066}{840 - 676} \\ &= \frac{249}{164} = 1.5183... \end{aligned}$$

**Step 4:** Calculate Intercept **b**:

$$\begin{aligned} b &= \frac{\sum y - m \sum x}{N} \\ &= \frac{41 - 1.5183 \times 26}{5} \\ &= 0.3049... \end{aligned}$$

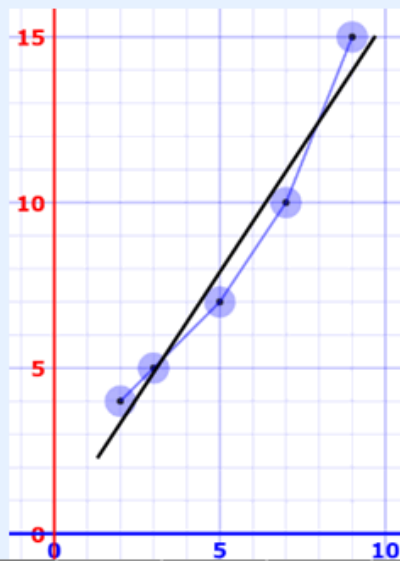
**Step 5:** Assemble the equation of a line:

$$\begin{aligned} y &= mx + b \\ y &= 1.518x + 0.305 \end{aligned}$$

Let's see how it works out:

x	y	$y = 1.518x + 0.305$	error
2	4	3.34	-0.66
3	5	4.86	-0.14
5	7	7.89	0.89
7	10	10.93	0.93
9	15	13.97	-1.03

Here are the (x,y) points and the line  $y = 1.518x + 0.305$  on a graph:



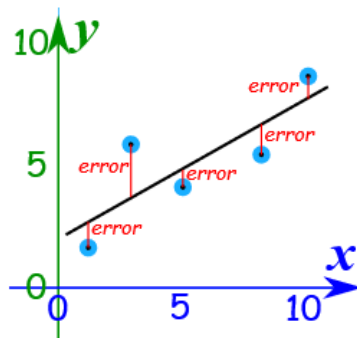
Sam hears the weather forecast which says "we expect 8 hours of sun tomorrow", so he uses the above equation to estimate that he will sell

$$y = 1.518 \times 8 + 0.305 = 12.45 \text{ Ice Creams}$$

Sam makes fresh waffle cone mixture for 14 ice creams just in case. Yum.

## How does it work?

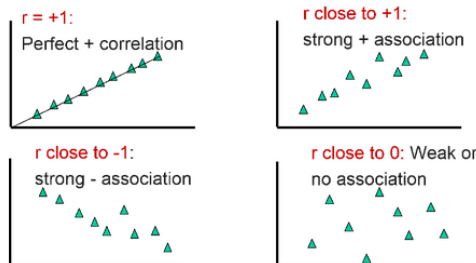
It works by making the total of the **square of the errors** as small as possible (that is why it is called "least squares"):



The straight line minimizes the sum of squared errors

## The Correlation Coefficient (r)

The sample correlation coefficient (r) is a measure of the closeness of association of the points in a scatter plot to a linear regression line based on those points, as in the example above for accumulated saving over time. Possible values of the correlation coefficient range from -1 to +1, with -1 indicating a perfectly linear negative, i.e., inverse, correlation (sloping downward) and +1 indicating a perfectly linear positive correlation (sloping upward).



## Formula to calculate r

$$r = \frac{n(\sum xy) - (\sum x)(\sum y)}{\sqrt{[n\sum x^2 - (\sum x)^2][n\sum y^2 - (\sum y)^2]}}$$



**Example question:** Find the value of the correlation coefficient from the following table:

SUBJECT	AGE X	GLUCOSE LEVEL Y
1	43	99
2	21	65
3	25	79
4	42	75
5	57	87
6	59	81

**Step 1:** *Make a chart.* Use the given data, and add three more columns:  $xy$ ,  $x^2$ , and  $y^2$ .

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	$x^2$	$y^2$
1	43	99			
2	21	65			
3	25	79			
4	42	75			
5	57	87			
6	59	81			

**Step 2:** *Multiply  $x$  and  $y$  together to fill the  $xy$  column. For example, row 1 would be  $43 \times 99 = 4,257$ .*

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	$x^2$	$y^2$
1	43	99	4257		
2	21	65	1365		
3	25	79	1975		
4	42	75	3150		
5	57	87	4959		
6	59	81	4779		

**Step 3:** Take the square of the numbers in the  $x$  column, and put the result in the  $x^2$  column.

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	$x^2$	$y^2$
1	43	99	4257	1849	
2	21	65	1365	441	
3	25	79	1975	625	
4	42	75	3150	1764	
5	57	87	4959	3249	
6	59	81	4779	3481	

**Step 4:** Take the square of the numbers in the  $y$  column, and put the result in the  $y^2$  column.

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	$x^2$	$y^2$
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561

**Step 5:** Add up all of the numbers in the columns and put the result at the bottom of the column. The Greek letter sigma ( $\Sigma$ ) is a short way of saying "sum of."

SUBJECT	AGE X	GLUCOSE LEVEL Y	XY	$x^2$	$y^2$
1	43	99	4257	1849	9801
2	21	65	1365	441	4225
3	25	79	1975	625	6241
4	42	75	3150	1764	5625
5	57	87	4959	3249	7569
6	59	81	4779	3481	6561
$\Sigma$	247	486	20485	11409	40022

**Step 6:** Use the following correlation coefficient formula.

$$r = \frac{n(\Sigma xy) - (\Sigma x)(\Sigma y)}{\sqrt{[n\Sigma x^2 - (\Sigma x)^2][n\Sigma y^2 - (\Sigma y)^2]}}$$

The answer is:  $2868 / 5413.27 = 0.529809$

From our table:

- $\Sigma x = 247$
- $\Sigma y = 486$
- $\Sigma xy = 20,485$
- $\Sigma x^2 = 11,409$
- $\Sigma y^2 = 40,022$
- $n$  is the sample size, in our case = 6

The correlation coefficient =

- $6(20,485) - (247 \times 486) / [\sqrt{[6(11,409) - (247^2)] \times [6(40,022) - 486^2]}]$   
= 0.5298

r value =

+0.70 or higher	Very strong positive relationship
+0.40 to +0.69	Strong positive relationship
+0.30 to +0.39	Moderate positive relationship
+0.20 to +0.29	weak positive relationship
+0.01 to +0.19	No or negligible relationship
0	No relationship [zero correlation]
-0.01 to -0.19	No or negligible relationship
-0.20 to -0.29	weak negative relationship
-0.30 to -0.39	Moderate negative relationship
-0.40 to -0.69	Strong negative relationship
-0.70 or higher	Very strong negative relationship

The theoretical counterpart of  $r$  is the **correlation coefficient**  $\rho$  of  $X$  and  $Y$ ,



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

$X$  and  $Y$  are called **uncorrelated** if  $\rho = 0$ .