The Chi – Squared Test
(X^2 – test)

The chi - square test is a non-parametric test not based on anyassumption or distribution of any variable.

This statistical test follows a specific distribution known as chi square distribution.

Chi square distribution

1. Positively skewed distribution
2. X^2 values will never be negative; minimum is zero
3. X^2 of close to 0 indicate that the variables are independent of one another.

The chi - square test

- It is a test for qualitative data.
- Based on counts or frequencies.
- Chi – squared test measures the difference between **actual frequencies** and **expected frequencies** (as expected under the null hypothesis)

$$X^2 \text{ – test} = \text{sum of } \frac{(\text{Observed frequency} - \text{Expected frequency})^2}{\text{Expected frequency}}$$

$$X^2 \text{ – test} = \Sigma \frac{(O - E)^2}{E}$$

Applications of a chi-square test

This test can be used in:

1. **Goodness of fit of distributions**

   This test enables us to see how well does the assumed theoretical distribution fit to the observed data.

   e.g. A researcher has chosen 25 participants (10 of whom are males and 15 are females) and wishes to know if there are significantly more female than male participants. Assuming that the participants were chosen from a population where the number of females and males is equal.

2. **Test of independence of attributes**

   This test enables us to explain whether or not two attributes are associated.

   e.g. we may be interested in knowing whether a new medicine is effective in controlling fever or not.

3. **Test of homogensity**

   This test can be also used to test whether the occurrence of events follows uniformity or not.

   e.g. the admission of patients in a governmental hospital in all days of the week is uniform or not can be tested with the help of chi square test.

Procedure:

1. State the null hypothesis (Ho):

There is no relationship between the two variables.

2. Arrange the data in a table.

3. Calculate the expected frequencies:

$$\textbf{Expected frequency (E)} = \frac{\textbf{Row total X Column total}}{\textbf{Grand total}}$$

4. Calculate X^2 value:

$$X^2 - \text{test} = \Sigma \frac{(O - E)^2}{E}$$

5. Determine degree of freedom:

df = (Rows – 1) (Columns – 1)

6. Compare the **calculated X^2 value** with the **tabulated critical value**.

7. Conclusion:

At 95% level

If the **calculated X^2 value < tabulated critical value**

$P > 0.05$

So **accept** the null hypothesis

If the **calculated X^2 value > tabulated critical value**

$P < 0.05$

So **reject** the null hypothesis

Example: The following data were obtained from a study on the association between smoking and lung cancer in men:

Smoking status	No. of persons who developed lung cancer	No. of persons who did not develop lung cancer	Total
Smokers	30	120	150
Non – smokers	10	100	110
Total	40	220	260

Perform a complete X^2 - test on the data in the table above to show whether an association does exist between smoking and lung cancer.

1. **Null hypothesis:** There is no relationship or association between smoking and lung cancer, and if there is association is due to chance or sampling error.
2. **Arrange the table.**
3. **Calculate the expected frequency for each cell.**

$$\textbf{Expected frequency (E)} = \frac{\textbf{Row total X Column total}}{\textbf{Grand total}}$$

$$E(30) = \frac{150 \text{ X } 40}{260} = 23.08$$

$$E(120) = \frac{150 \text{ X } 220}{260} = 126.92$$

$$E(10) = \frac{110 \text{ X } 40}{260} = 16.92$$

$$E(100) = \frac{110 \text{ X } 220}{260} = 93.08$$

4. **Calculate X^2 value:**

$$X^2 - \text{test} = \Sigma \frac{(O - E)^2}{E}$$

$$= \frac{(30 \; - 23.08)^2}{23.08} + \frac{(120 - 126.92)^2}{126.92} - \_ + \frac{(10 - 16.92)^2}{16.92} -$$

$$+ \frac{(100 - 93.08)^2}{93.08} -$$

$= 2.08 + 0.37 + 2.83 + 0.51$

$= 5.79$

5. **Calculate degree of freedom:**

df = (Rows – 1) (Columns – 1)

= (2 – 1) (2 – 1)

= 1

6. Tabulated critical X^2 value:

Df	0.05	0.01
1	3.84	6.63

At 95% level

$5.79 > 3.84$

$P < 0.05$

So **reject** the null hypothesis

There is **a significant relationship** between smoking and the development of lung cancer.

At 99.7% level

$5.79 < 6.63$

$P > 0.01$

So **accept** the null hypothesis

No highly significant relationship between smoking and development of lung cancer.

$0.05 > P > 0.01$

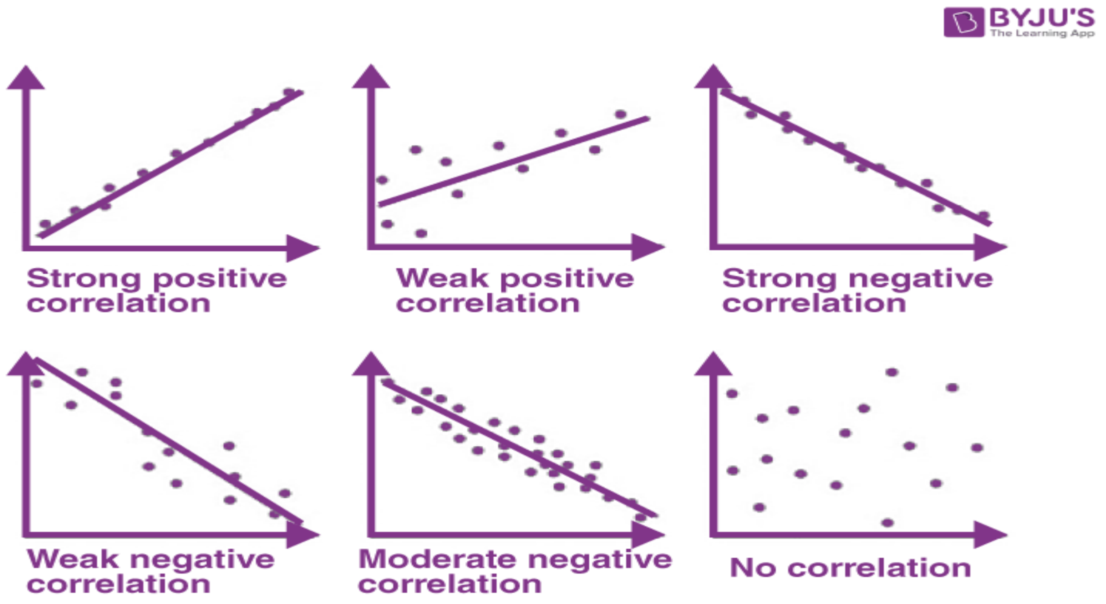
Correlation

Correlation coefficients are used to measure how strong a relationship is between two variables. There are several types of correlation coefficient, but the most popular is Pearson’s. Pearson’s correlation (also called Pearson’s R) is a correlation coefficient commonly used in linear regression. If you’re starting out in statistics, you’ll probably learn about Pearson’s R first. In fact, when anyone refers to the correlation coefficient, they are usually talking about Pearson’s.

Correlation Coefficient Formula: Definition

Correlation coefficient formulas are used to find how strong a relationship is between data. The formulas return a value between -1 and 1, where:

1 indicates a strong positive relationship.
-1 indicates a strong negative relationship.
A result of zero indicates no relationship at all.
Types of correlation coefficient formulas.
There are several types of correlation coefficient formulas.



correlation coefficient formulas:

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

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 × 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 (Σ) is a short way of saying "sum of" or summation.

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
Σ	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 × 486) / [√[[6(11,409) – (247^2)] × [6(40,022) – 486^2]]]
  = 0.5298

The range of the correlation coefficient is from -1 to 1. Our result is 0.5298 or 52.98%, which means the variables have a moderate positive correlation.