

# What Is a Chi-Square Test?

---

 [simplilearn.com/tutorials/statistics-tutorial/chi-square-test](https://simplilearn.com/tutorials/statistics-tutorial/chi-square-test)

## What is a Chi-Square Test? Formula, Examples & Application

---

The world is constantly curious about the Chi-Square test's application in machine learning and how it makes a difference. Feature selection is a critical topic in machine learning, as you will have multiple features in line and must choose the best ones to build the model. By examining the relationship between the elements, the chi-square test aids in the solution of feature selection problems. In this tutorial, you will learn about the chi-square test and its application.

## Fundamentals of Hypothesis Testing

---

Hypothesis testing is a technique for interpreting and drawing inferences about a population based on sample data. It aids in determining which sample data best support mutually exclusive population claims.

Null Hypothesis ( $H_0$ ) - The Null Hypothesis is the assumption that the event will not occur. A null hypothesis has no bearing on the study's outcome unless it is rejected.

$H_0$  is the symbol for it, and it is pronounced H-naught.

Alternate Hypothesis ( $H_1$  or  $H_a$ ) - The Alternate Hypothesis is the logical opposite of the null hypothesis. The acceptance of the alternative hypothesis follows the rejection of the null hypothesis.  $H_1$  is the symbol for it.

## What Are Categorical Variables?

---

Categorical variables belong to a subset of variables that can be divided into discrete categories. Names or labels are the most common categories. These variables are also known as qualitative variables because they depict the variable's quality or characteristics.

Categorical variables can be divided into two categories:

1. Nominal Variable: A nominal variable's categories have no natural ordering.  
Example: Gender, Blood groups
2. Ordinal Variable: A variable that allows the categories to be sorted is ordinal variables. Customer satisfaction (Excellent, Very Good, Good, Average, Bad, and so on) is an example.

The Chi-Square test is a statistical procedure for determining the difference between observed and expected data. This test can also be used to determine whether it correlates to the categorical variables in our data. It helps to find out whether a difference between two categorical variables is due to chance or a relationship between them.

## Chi-Square Test Definition

---

A chi-square test is a statistical test that is used to compare observed and expected results. The goal of this test is to identify whether a disparity between actual and predicted data is due to chance or to a link between the variables under consideration. As a result, the chi-square test is an ideal choice for aiding in our understanding and interpretation of the connection between our two categorical variables.

A chi-square test or comparable nonparametric test is required to test a hypothesis regarding the distribution of a categorical variable. Categorical variables, which indicate categories such as animals or countries, can be nominal or ordinal. They cannot have a normal distribution since they can only have a few particular values.

For example, a meal delivery firm in India wants to investigate the link between gender, geography, and people's food preferences.

It is used to calculate the difference between two categorical variables, which are:

- As a result of chance or
- Because of the relationship

## Data Scientist Master's Program

---

In Collaboration with IBM [Explore Course](#)



## Formula For Chi-Square Test

---

Where

c = Degrees of freedom

O = Observed Value

E = Expected Value

$$\chi^2_c = \frac{\sum (O_i - E_i)^2}{E_i}$$

The degrees of freedom in a statistical calculation represent the number of variables that can vary in a calculation. The degrees of freedom can be calculated to ensure that chi-square tests are statistically valid. These tests are frequently used to compare observed data with data that would be expected to be obtained if a particular hypothesis were true.

The Observed values are those you gather yourselves.

The expected values are the frequencies expected, based on the null hypothesis.

## Why Do You Use the Chi-Square Test?

---

Chi-square is a statistical test that examines the differences between categorical variables from a random sample in order to determine whether the expected and observed results are well-fitting.

Here are some of the uses of the Chi-Squared test:

- The Chi-squared test can be used to see if your data follows a well-known theoretical probability distribution like the Normal or Poisson distribution.
- The Chi-squared test allows you to assess your trained regression model's goodness of fit on the training, validation, and test data sets.

## What Does A Chi-Square Statistic Test Tell You?

---

A Chi-Square test ( symbolically represented as  $\chi^2$  ) is fundamentally a data analysis based on the observations of a random set of variables. It computes how a model equates to actual observed data. A Chi-Square statistic test is calculated based on the data, which must be raw, random, drawn from independent variables, drawn from a wide-ranging sample and mutually exclusive. In simple terms, two sets of statistical data are compared - for instance, the results of tossing a fair coin. Karl Pearson introduced this test in 1900 for categorical data analysis and distribution. This test is also known as 'Pearson's Chi-Squared Test'.

Chi-Squared Tests are most commonly used in hypothesis testing. A hypothesis is an assumption that any given condition might be true, which can be tested afterwards. The Chi-Square test estimates the size of inconsistency between the expected results and the actual results when the size of the sample and the number of variables in the relationship is mentioned.

These tests use degrees of freedom to determine if a particular null hypothesis can be rejected based on the total number of observations made in the experiments. Larger the sample size, more reliable is the result.

There are two main types of Chi-Square tests namely -

1. Independence
2. Goodness-of-Fit

### Independence

---

The Chi-Square Test of Independence is a derivable ( also known as inferential ) statistical test which examines whether the two sets of variables are likely to be related with each other or not. This test is used when we have counts of values for two nominal or categorical variables and is considered as non-parametric test. A relatively large sample size and independence of observations are the required criteria for conducting this test.

For Example-

In a movie theatre, suppose we made a list of movie genres. Let us consider this as the first variable. The second variable is whether or not the people who came to watch those genres of movies have bought snacks at the theatre. Here the null hypothesis is that the genre of the film and whether people bought snacks or not are unrelatable. If this is true, the movie genres don't impact snack sales.

>

## Goodness-Of-Fit

---

In statistical hypothesis testing, the Chi-Square Goodness-of-Fit test determines whether a variable is likely to come from a given distribution or not. We must have a set of data values and the idea of the distribution of this data. We can use this test when we have value counts for categorical variables. This test demonstrates a way of deciding if the data values have a “good enough” fit for our idea or if it is a representative sample data of the entire population.

For Example-

Suppose we have bags of balls with five different colours in each bag. The given condition is that the bag should contain an equal number of balls of each colour. The idea we would like to test here is that the proportions of the five colours of balls in each bag must be exact.

## Who Uses Chi-Square Analysis?

---

Chi-square is most commonly used by researchers who are studying survey response data because it applies to categorical variables. Demography, consumer and marketing research, political science, and economics are all examples of this type of research.

## Example

---

Let's say you want to know if gender has anything to do with political party preference. You poll 440 voters in a simple random sample to find out which political party they prefer. The results of the survey are shown in the table below:

	Republican	Democrat	Independent	Total
Male	100	70	30	200
Female	140	60	20	220
Total	240	130	50	440

To see if gender is linked to political party preference, perform a Chi-Square test of independence using the steps below.

### Step 1: Define the Hypothesis

---

Ho: There is no link between gender and political party preference.

H1: There is a link between gender and political party preference.

## Step 2: Calculate the Expected Values

---

Now you will calculate the expected frequency.

$$\text{Expected Value} = \frac{(\text{Row Total}) * (\text{Column Total})}{\text{Total Number Of Observations}}$$

For example, the expected value for Male Republicans is:

$$= \frac{(240) * (200)}{440} = 109$$

Similarly, you can calculate the expected value for each of the cells.

Expected Values				
	Republican	Democrat	Independent	Total
Male	109	59	22.72	200
Female	120	65	25	220
Total	240	130	50	440

## Step 3: Calculate (O-E)<sup>2</sup> / E for Each Cell in the Table


---

Now you will calculate the (O - E)<sup>2</sup> / E for each cell in the table.

Where

O = Observed Value

E = Expected Value

(O - E) <sup>2</sup> /E				
	Republican	Democrat	Independent	Total
Male	0.74311927	2.050847	2.332676056	200
Female	3.333333333	0.384615	1	220
Total	240	130	50	 440

## Step 4: Calculate the Test Statistic X<sup>2</sup>

---

X<sup>2</sup> is the sum of all the values in the last table

$$= 0.743 + 2.05 + 2.33 + 3.33 + 0.384 + 1$$

$$= 9.837$$

Before you can conclude, you must first determine the critical statistic, which requires determining our degrees of freedom. The degrees of freedom in this case are equal to the table's number of columns minus one multiplied by the table's number of rows minus one, or  $(r-1)(c-1)$ . We have  $(3-1)(2-1) = 2$ .

Finally, you compare our obtained statistic to the critical statistic found in the chi-square table. As you can see, for an alpha level of 0.05 and two degrees of freedom, the critical statistic is 5.991, which is less than our obtained statistic of 9.83. You can reject our null hypothesis because the critical statistic is higher than your obtained statistic.

This means you have sufficient evidence to say that there is an association between gender and political party preference.

Critical values of the Chi-square distribution with $d$ degrees of freedom							
Probability of exceeding the critical value							
$d$	0.05	0.01	0.001	$d$	0.05	0.01	0.001
1	3.841	6.635	10.828	11	19.675	24.725	31.264
2	5.991	9.210	13.816	12	21.026	26.217	32.910
3	7.815	11.345	16.266	13	22.362	27.688	34.528
4	9.488	13.277	18.467	14	23.685	29.141	36.123
5	11.070	15.086	20.515	15	24.996	30.578	37.697
6	12.592	16.812	22.458	16	26.296	32.000	39.252
7	14.067	18.475	24.322	17	27.587	33.409	40.790
8	15.507	20.090	26.125	18	28.869	34.805	42.312
9	16.919	21.666	27.877	19	30.144	36.191	43.820
10	18.307	23.209	29.588	20	31.410	37.566	45.315

INTRODUCTION TO POPULATION GENETICS, Table D.1  
© 2013 Sinauer Associates, Inc.

## When to Use a Chi-Square Test?

A Chi-Square Test is used to examine whether the observed results are in order with the expected values. When the data to be analysed is from a random sample, and when the variable is the question is a categorical variable, then Chi-Square proves the most appropriate test for the same. A categorical variable consists of selections such as breeds of dogs, types of cars, genres of movies, educational attainment, male v/s female etc. Survey responses and questionnaires are the primary sources of these types of data. The

Chi-square test is most commonly used for analysing this kind of data. This type of analysis is helpful for researchers who are studying survey response data. The research can range from customer and marketing research to political sciences and economics.

## Chi-Square Distribution

---

Chi-square distributions ( $X^2$ ) are a type of continuous probability distribution. They're commonly utilized in hypothesis testing, such as the chi-square goodness of fit and independence tests. The parameter  $k$ , which represents the degrees of freedom, determines the shape of a chi-square distribution.

A chi-square distribution is followed by very few real-world observations. The objective of chi-square distributions is to test hypotheses, not to describe real-world distributions. In contrast, most other commonly used distributions, such as normal and Poisson distributions, may explain important things like baby birth weights or illness cases per year.

Because of its close resemblance to the conventional normal distribution, chi-square distributions are excellent for hypothesis testing. Many essential statistical tests rely on the conventional normal distribution.

In statistical analysis, the Chi-Square distribution is used in many hypothesis tests and is determined by the parameter  $k$  degree of freedoms. It belongs to the family of continuous probability distributions. The Sum of the squares of the  $k$  independent standard random variables is called the Chi-Squared distribution. Pearson's Chi-Square Test formula is -

$$X^2 = \sum \frac{(O-E)^2}{E}$$

Where  $X^2$  is the Chi-Square test symbol

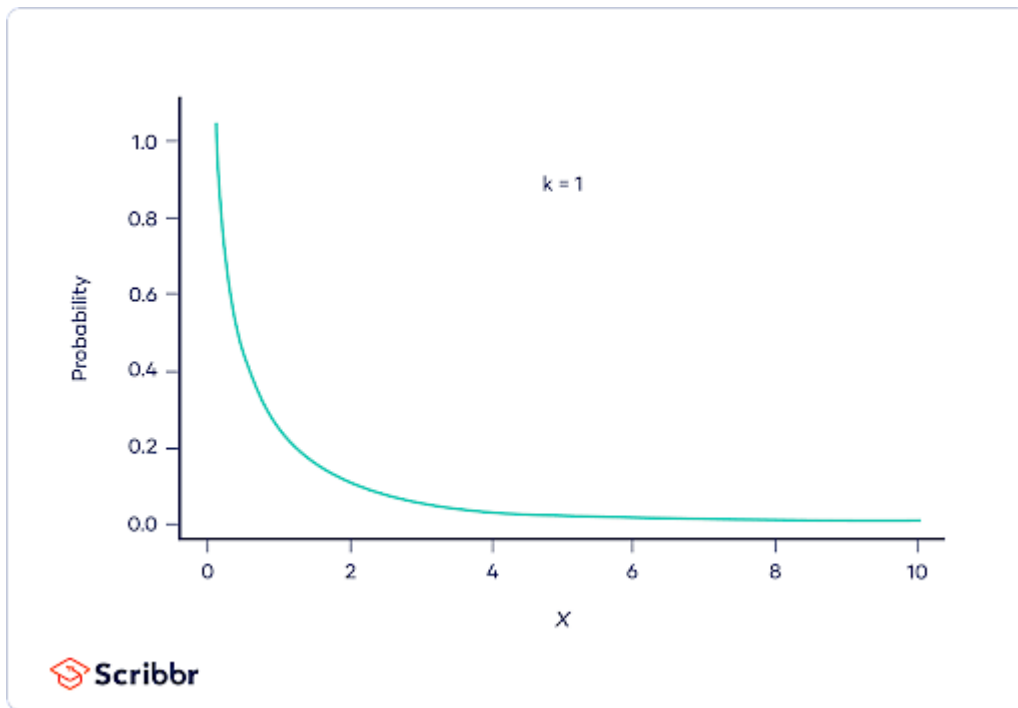
$\Sigma$  is the summation of observations

$O$  is the observed results

$E$  is the expected results

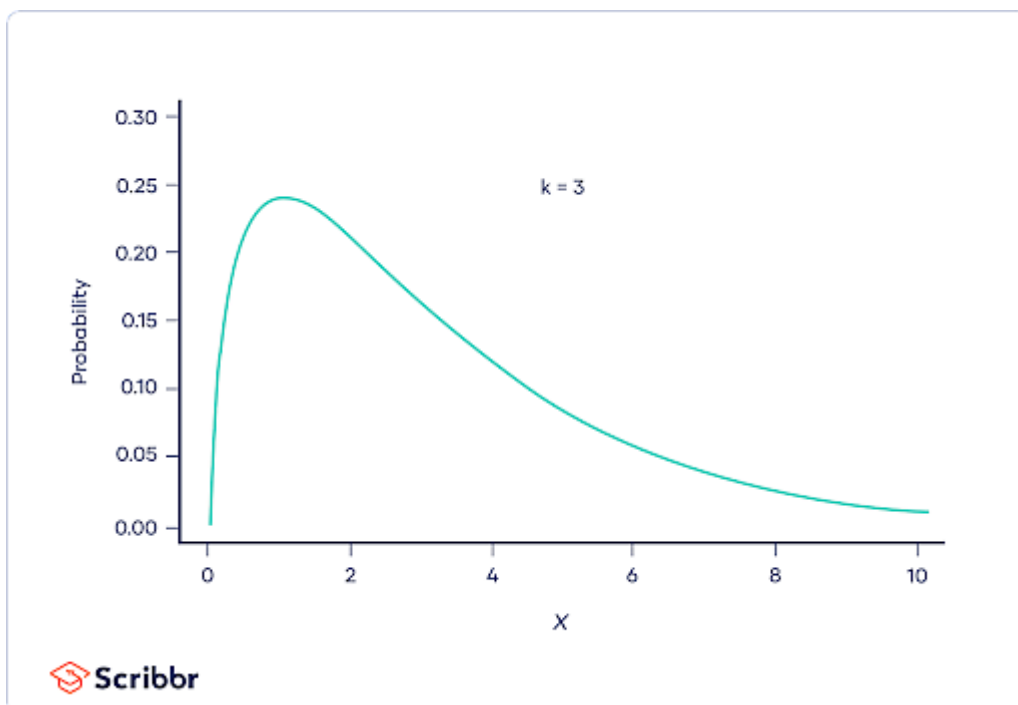
The shape of the distribution graph changes with the increase in the value of  $k$ , i.e. degree of freedoms.

When  $k$  is 1 or 2, the Chi-square distribution curve is shaped like a backwards 'J'. It means there is a high chance that  $X^2$  becomes close to zero.



Courtesy: Scribbr

When  $k$  is greater than 2, the shape of the distribution curve looks like a hump and has a low probability that  $X^2$  is very near to 0 or very far from 0. The distribution occurs much longer on the right-hand side and shorter on the left-hand side. The probable value of  $X^2$  is  $(X^2 - 2)$ .



Courtesy: Scribbr

When  $k$  is greater than ninety, a normal distribution is seen, approximating the Chi-square distribution.





---

## Chi-Square P-Values

---

Here P denotes the probability; hence for the calculation of p-values, the Chi-Square test comes into the picture. The different p-values indicate different types of hypothesis interpretations.

1.  $P \leq 0.05$  (Hypothesis interpretations are rejected)
2.  $P > 0.05$  (Hypothesis interpretations are accepted)

The concepts of probability and statistics are entangled with Chi-Square Test. Probability is the estimation of something that is most likely to happen. Simply put, it is the possibility of an event or outcome of the sample. Probability can understandably represent bulky or complicated data. And statistics involves collecting and organising, analysing, interpreting and presenting the data.

---

## Finding P-Value

---

When you run all of the Chi-square tests, you'll get a test statistic called  $X^2$ . You have two options for determining whether this test statistic is statistically significant at some alpha level:

1. Compare the test statistic  $X^2$  to a critical value from the Chi-square distribution table.
2. Compare the p-value of the test statistic  $X^2$  to a chosen alpha level.

Test statistics are calculated by taking into account the sampling distribution of the test statistic under the null hypothesis, the sample data, and the approach which is chosen for performing the test.

The p-value will be as mentioned in the following cases.

- A lower-tailed test is specified by:  $P(TS \leq ts \mid H_0 \text{ is true})$  p-value =  $\text{cdf}(ts)$
- Lower-tailed tests have the following definition:  $P(TS \leq ts \mid H_0 \text{ is true})$  p-value =  $\text{cdf}(ts)$
- A two-sided test is defined as follows, if we assume that the test static distribution of  $H_0$  is symmetric about 0.  $2 * P(TS \geq |ts| \mid H_0 \text{ is true}) = 2 * (1 - \text{cdf}(|ts|))$

Where:

P: probability Event

TS: Test statistic is computed observed value of the test statistic from your sample  $\text{cdf}()$ : Cumulative distribution function of the test statistic's distribution (TS)

## Types of Chi-square Tests

---

Pearson's chi-square tests are classified into two types:

1. Chi-square goodness-of-fit analysis
2. Chi-square independence test

These are, mathematically, the same exam. However, because they are utilized for distinct goals, we generally conceive of them as separate tests.

## Properties

---

The chi-square test has the following significant properties:

1. If you multiply the number of degrees of freedom by two, you will receive an answer that is equal to the variance.
2. The chi-square distribution curve approaches the data is normally distributed as the degree of freedom increases.
3. The mean distribution is equal to the number of degrees of freedom.

## Properties of Chi-Square Test

---

1. Variance is double the times the number of degrees of freedom.
2. Mean distribution is equal to the number of degrees of freedom.
3. When the degree of freedom increases, the Chi-Square distribution curve becomes normal.

## Limitations of Chi-Square Test

---

There are two limitations to using the chi-square test that you should be aware of.

- The chi-square test, for starters, is extremely sensitive to sample size. Even insignificant relationships can appear statistically significant when a large enough sample is used. Keep in mind that "statistically significant" does not always imply "meaningful" when using the chi-square test.
- Be mindful that the chi-square can only determine whether two variables are related. It does not necessarily follow that one variable has a causal relationship with the other. It would require a more detailed analysis to establish causality.

## Chi-Square Goodness of Fit Test

---

When there is only one categorical variable, the chi-square goodness of fit test can be used. The frequency distribution of the categorical variable is evaluated for determining whether it differs significantly from what you expected. The idea is that the categories will have equal proportions, however, this is not always the case.

## SPSS

---

When you want to see if there is a link between two categorical variables, you perform the chi-square test. To acquire the test statistic and its related p-value in SPSS, use the chisq option on the statistics subcommand of the crosstabs command. Remember that the chi-square test implies that each cell's anticipated value is five or greater.

Learn over a dozen of data analytics tools and skills with [PG Program in Data Analytics](#) and gain access to masterclasses by Purdue faculty and IBM experts. Enroll and add a star to your data analytics resume now!

## Conclusion

---

In this tutorial titled 'The Complete Guide to Chi-square test', you explored the concept of Chi-square distribution and how to find the related values. You also take a look at how the critical value and chi-square value is related to each other.

If you want to gain more insight and get a work-ready understanding in statistical concepts and learn how to use them to get into a [career in Data Analytics](#), our Post Graduate Program in Data Analytics in partnership with Purdue University should be your next stop. A comprehensive program with training from top practitioners and in collaboration with IBM, this will be all that you need to kickstart your career in the field.

Was this tutorial on the Chi-square test useful to you? Do you have any doubts or questions for us? Mention them in this article's comments section, and we'll have our experts answer them for you at the earliest!

## Our Learners Also Asked

---

### 1. What is the Chi-square test? Write its formula.

---

ANS. A test used for measuring the size of inconsistency between the expected results and the observed results is called the Chi-Square Test. The formula for the Chi-Square Test is given below-

$$X^2 = \sum \frac{(O-E)^2}{E}$$

Where  $X^2$  is the Chi-Square test symbol

$\Sigma$  is the summation of observations

O is the observed results

E is the expected results

### 2. How do you calculate the Chi-squared?

---

The value of the Chi-squared test can be formulated by using the formula given below-

By following the steps mentioned above, the Chi-Square statistic can be calculated-

- Subtract the expected result from the observed results,i.e. (O-E)
- Square the difference obtained, i.e.  $(O-E)^2$
- Divide the squared difference by the expected result of each observation,i.e.  $(O-E)^2/E$
- Finally, we obtain the Chi-Squared statistic by taking the sum of the whole expression.

### 3. What is a Chi-square test used for?

---

The Chi-Squared statistic is used to examine whether there is a difference between the observed and the expected results.

### 4. How do you interpret a Chi-squared test?

---

The P-value less than or equal to the defined significance level demonstrates adequate proof to conclude that the observed results are the same as the expected results. Therefore, in a Chi-Square test, we can conclude whether there exists a relationship between the categorical variables or not.

### 5. What is a good Chi-square value?

---

5 is assumed to be a good Chi-square value. For a chi-square approach to be valid, at least five must be the expected frequency.

## About the Author

---

### Avijeet Biswal

Avijeet is a Senior Research Analyst at Simplilearn. Passionate about Data Analytics, Machine Learning, and Deep Learning, Avijeet is also interested in politics, cricket, and football.



[View More](#)