**Fundamental of The Chi Square in Statistics**

[](https://medium.com/@nhan.tran?source=post_page-----37a8eb2f27bb----------------------)

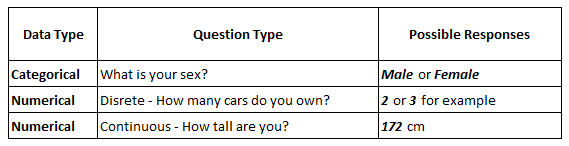
[Nhan Tran](https://medium.com/@nhan.tran?source=post_page-----37a8eb2f27bb----------------------)

[Mar 7](https://medium.com/@nhan.tran/the-chi-square-statistic-p-1-37a8eb2f27bb?source=post_page-----37a8eb2f27bb----------------------) · 8 min read

In this session, I’ll introduce you about **Chi Square** term in statistic. If you don’t know about [*Contingency Table*](https://medium.com/@nhan.tran/contingency-table-55e375a76a8?sk=8922b1162b15e455b89ee6a5910fc5f9&fbclid=IwAR3bO_A6a0ZR_JrrYs_4Hy2tYosfrhwvUPNgRHU7DESPSXkFAG0MLtWasto), please take a look on my previous post before moving forward. It helps.

**Types of Data:**

There are basically two types of random variables and they yield two types of data: numerical and categorical. A chi square ***x²*** statistic is used to investigate whether distributions of categorical variables differ from one another. Basically categorical variable yield data in the categories and numerical variables yield data in numerical form. Responses to such questions as *“What is your major?”* or *“Do you own a car?”* are **categorical** because they yield data such as *“biology”* or *“no”*. In contrast, responses to such questions as *“How tall are you?”* or *“What is your G.P.A.?”* are **numerical**. Numerical data can be either discrete or continuous. The table below may help you see the differences between these two variables.



*Notice that discrete data arise from a counting process, while continuous data arise from a measuring process.*

The Chi Square statistic compares the tallies or counts of categorical responses between two (or more) independent groups. (note: Chi square tests can only be used on actual numbers and not on percentages, proportions, means, etc.)

Chi-square Test is a method that is used to test if there is any relationship between two categorical variables.

In other words, a Chi-square test is a test for independence. Therefore, our Hypothesis statements are going to be the following

* H0: X and Y are **independent**.
* H1: X and Y are **dependent**.

**2 x 2 Contingency Table**

There are several types of chi square tests depending on the way the data was collected and the hypothesis being tested. We’ll begin with the simplest case: a 2 x 2 contingency table. If we set the 2 x 2 table to the general notation shown below in Table 1, using the letters a, b, c, and d to denote the contents of the cells, then we would have the following table:

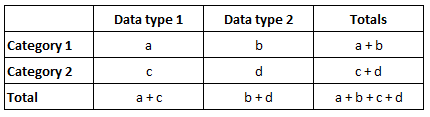
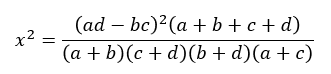


Table 1: General notation for a 2 x 2 contingency table

For a 2 x 2 contingency table the Chi Square statistic is calculated by the formula:



Chi square (x²) formula for 2 x 2 contingency table

*Note: notice that the four components of the denominator are the four totals from the table columns and rows.*

Suppose you conducted a drug trial on a group of animals and you hypothesized that the animals receiving the drug would show increased heart rates compared to those that did not receive the drug. You conduct the study and collect the following data:

* H0: The proportion of animals whose heart rate increased is **independent** of drug treatment.
* H1: The proportion of animals whose heart rate increased is **dependent** with drug treatment.

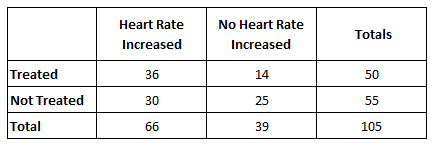
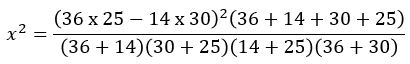


Table 2: Sample notation for a 2 x 2 contingency table

Applying the formula above we get:

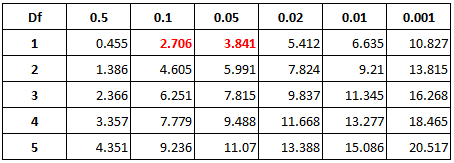


…and the result is***x*² = 3.417673**

Before we can proceed we need to know how many degrees of freedom we have. When a comparison is made between one sample and another, a simple rule is that the degrees of freedom equal (number of columns minus one) x (number of rows minus one) not counting the totals for rows or columns. For our data this gives (2–1) x (2–1) = 1.

We now have our chi square statistic (*x*² = 3.418), our predetermined alpha level of significance (0.05), and our degrees of freedom (df = 1). Entering the Chi square distribution table with 1 degree of freedom and reading along the row we find our value of *x*²(3.418) lies between 2.706 and 3.841. The corresponding probability is between the 0.10 and 0.05 probability levels. That means that the p-value is above 0.05 (it is actually 0.065). Since a p-value of 0.065 is greater than the conventionally accepted significance level of 0.05 (i.e. p > 0.05) we fail to reject the null hypothesis. In other words, there is no statistically significant difference in the proportion of animals whose heart rate increased.

What would happen if the number of control animals whose heart rate increased dropped to 29 instead of 30 and, consequently, the number of controls whose hear rate did not increase changed from 25 to 26? Try it. Notice that the new *x*² value is 4.125 and this value exceeds the table value of 3.841 (at 1 degree of freedom and an alpha level of 0.05). This means that p < 0.05 (it is now 0.04) and we reject the null hypothesis in favor of the alternative hypothesis — the heart rate of animals is different between the treatment groups. When p < 0.05 we generally refer to this as a significant difference.



Probability level (alpha)

**Last but not least, we can calculate chi square of the 2 x 2 contingency table in another way…**

Let’s say we had a random sample of 237 pupils who were asked if they ever got into troubles at school. The result is the table below:

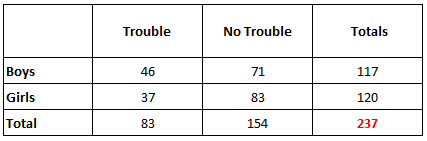


Table 3: Pupil situation observation

Because the number of pupils is varying per gender, it is hard to compare boys and girls in that way. Therefore, let’s standardize joint frequencies by dividing counts within each row with its corresponding row total. In addition, let’s standardize marginal frequencies by dividing each marginal frequency by the overall total (located on the bottom right corner — ***237***)

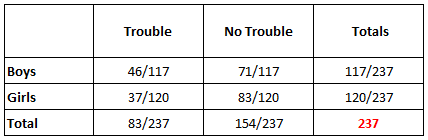


Table 4a: Convert observations to frequencies

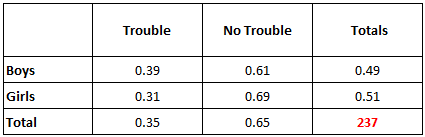
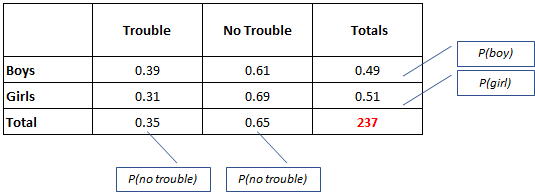


Table 4b: Results in frequencies

This way, joint frequencies become joint (conditional) probabilities, or ***observed*** probabilities (marked in green): it takes into account two categorical variables. For instance, 0.39 is a probability of a boy being in trouble ***P(B, Trouble)***. Marginal frequencies become marginal probabilities — it takes into account only one of the categorical variables. For instance, 0.49 is a probability of being a boy ***P(B)***, or 0.35 — a probability of being in trouble ***P(Trouble)***.



In the beginning, we outlined our hypothesis statements as follows:

* H0: X and Y are **independent**.
* H1: X and Y are **dependent**.

Probabilities in statistics states that if two events are independent, the following equation is satisfied:

P(X, Y) = P(X) x P(Y)

*…where X and Y are some events.*

Chi square test is ***based*** on this assumption. Therefore, if is true, meaning that *X* and *Y* are independent, the following equation will be satisfied:

P(gender, situation) = P(gender) x P(situation)

On the left side of this equation we see our joint probability, and on the right side of this equation, we see two marginal probabilities.

First, Chi-square uses this assumption to calculate expected probabilities (joint probabilities) using its marginal probabilities. For instance, expected probability for boys being in trouble is

*P(Boy) x P(Trouble)  
= 0.35 x 0.49*= *0.17*

or expected probability for girls being in trouble is

*P(Girl) x P(Trouble)  
= 0.35 x 0.51  
= 0.18*

…etc*.* Find expected probabilities in the brackets below:

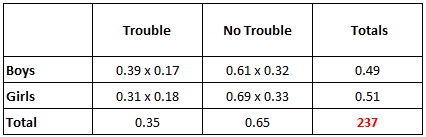
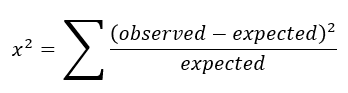


Table 5: Finding P(X,Y), P(X), P(Y)

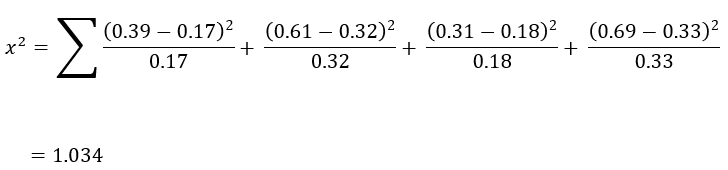
In other words, when we calculate the expected probabilities, we calculate probabilities that we should expect if is true, or, if *X* and *Y* are independent variables. That means that, if boys and trouble status are independent variables, our expected probability for boys not being in trouble is 0.32.

Secondly, we should measure the differences between the actual probabilities (actual joint probabilities in our tables) and those expected probabilities we have just calculated. If we see that the difference between our actual probabilities and the probabilities we expect to have in case two categorical variables are independent, is huge, then our variables are most likely not independent. Similarly, if the difference between our actual probabilities in our table and probabilities that we suppose to get in case of independence is small, our two variables are most likely independent. The difference between these two probabilities is represented by value that we have to calculate using the formula:



Chi square (x²) formula

So, we just plug the values from the table above:

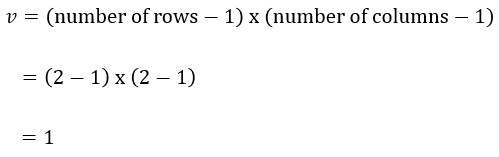


Last step is to compare value with the value in the distribution table (denoted as) to conclude if you should accept or reject . The following procedures applies:

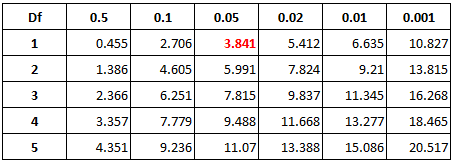
* ***x*² > table value: accept** — meaning that you have enough statistical to conclude that two variables are **dependent**.
* ***x*² ≤ table value: reject** — meaning that you have enough statistical to conclude that two variables are **independent**.

To get a proper value from the table, we have to know two things:*Significance level* and ***degrees of freedom***. In the table, significance level (*α*) is on the top, and degrees of freedom (*υ*) is on the left side.

1. The significance level is something that you choose yourself. Let’s use a significance level of *5%, so α = 0.05.*
2. Degrees of freedom in Chi-square is calculated using a formula:



With **significance level of *0.05*** and **degree of freedom of *1***, we have **table value *= 3.8415***(first row, third column)



Probability level (alpha)

Now we have both ***x*²** and **table values** and can compare these two to make a decision.

So, any estimated from above calculation ***x*² = 1.034** which is below our **table values = 3.8415**, that is, any difference between our actual probabilities and probabilities we expect to have if two variables are independent, that is below 3.8415, means that these **two variables are independent**, or have **no relationships in-between**.

As our ***x*² < table value**, we can reject H1 and conclude that the ***gender and trouble status are NOT correlated*** with each other.

**Conclusion**

So, here are the steps to make a Chi-square test:

1. Add marginal frequencies to a contingency table
2. Translate joint and marginal frequencies into probabilities
3. Estimate the expected probability for each cell
4. Calculate ***x*²**
5. Compare ***x*²** with **table value** and make a decision:

* *x*² > table value = accept = dependent
* *x*² ≤ table value = reject = independent

)