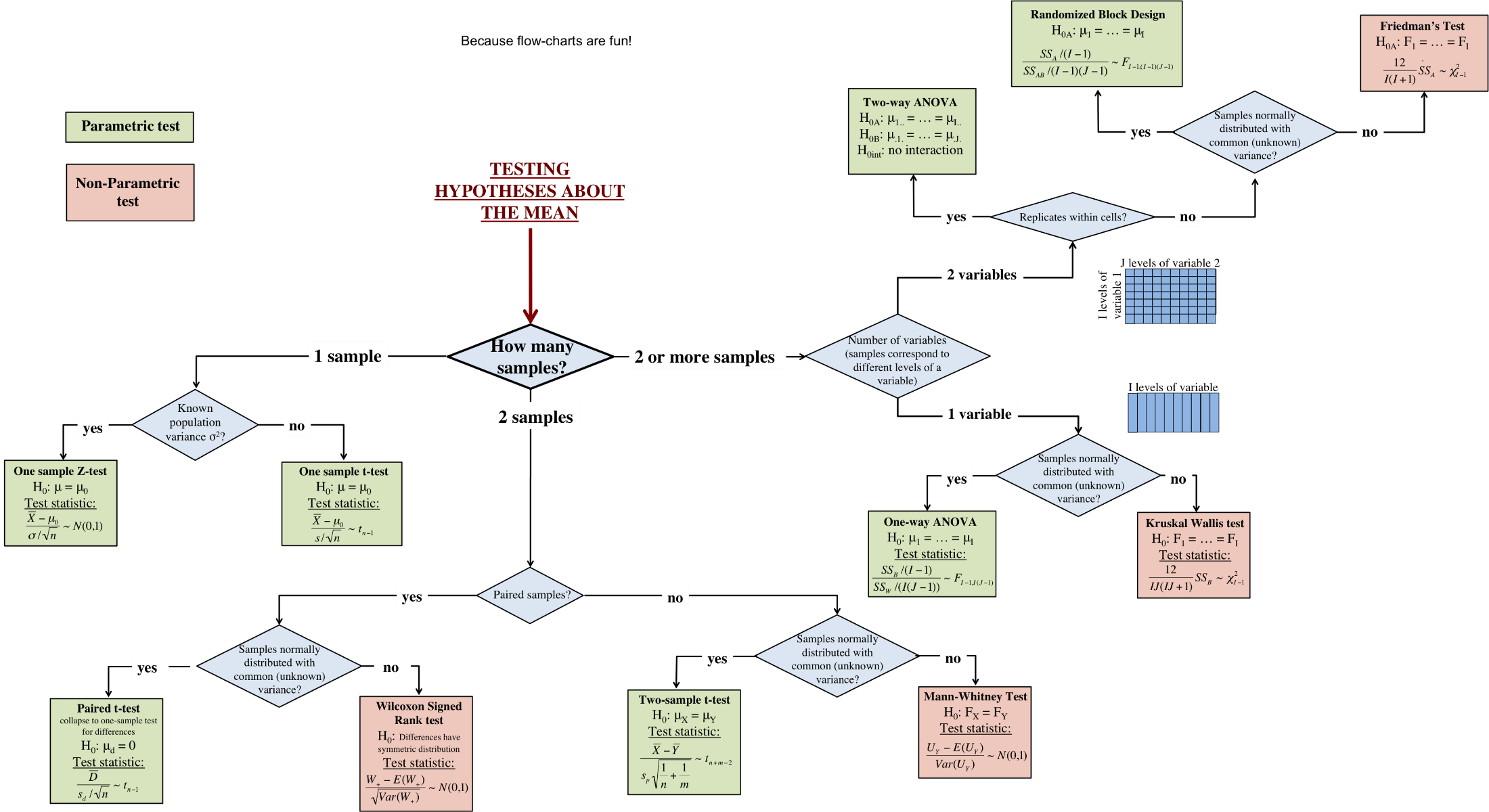
**Inferential Statistics: Understanding Hypothesis Testing Using Chi-Square Test**

[](https://towardsdatascience.com/@pramod.pandey83?source=post_page-----eacf9fcac533----------------------)

[Pramod Chandrayan](https://towardsdatascience.com/@pramod.pandey83?source=post_page-----eacf9fcac533----------------------)

[Jul 11](https://towardsdatascience.com/inferential-statistic-understanding-hypothesis-testing-using-chi-square-test-eacf9fcac533?source=post_page-----eacf9fcac533----------------------) · 11 min read



[src](http://www.rebeccabarter.com/blog/2018-12-04_hypothesis_testing/)

*As data science engineer it’s imperative that the sample data set which you pick from the population data is reliable, clean and well tested for its usability in Machine learning model building.*

**So how do you do that ?** Well, we have multiple statistical techniques like descriptive statistic where we measure the data central value, how it is spread across the mean/median. Is it normally distributed or there is a skew in the data spread. Please refer my previous article on the same for more clarity.

**[Descriptive Statistics Fundamentals For Data Science Aspirants](https://towardsdatascience.com/descriptive-statistics-fundamentals-for-data-science-aspirants-6500e8a0c964)**

**[Applied statistics fundamental for Data Science aspirants](https://towardsdatascience.com/descriptive-statistics-fundamentals-for-data-science-aspirants-6500e8a0c964)**

**[towardsdatascience.com](https://towardsdatascience.com/descriptive-statistics-fundamentals-for-data-science-aspirants-6500e8a0c964)**

As first thing we do is to visualize the data using various data visualization techniques to make some early sense of any data skewness or discrepancies , to identify any kind of relationship between data set variables.

*Data has so much to say and we data engineer give it a voice to express and describe itself using descriptive statistical techniques .*

But to make any prediction or to infer something beyond the given data to find any hidden probability, we rely on inferential statistic techniques .

*Inferential statistics are concerned with making inferences based on relations found in the sample, to relations in the population. Inferential statistics help us decide, for example, whether the differences between groups that we see in our data are strong enough to provide support for our hypothesis that group differences exist in general, in the entire population.*

Today we will cover one of the inferential statistical mechanism to understand the concept of hypothesis testing using popular Chi-Square test.

**What is Chi-Square Test ?**

*Do, remember that ,*

***It is an inferential statistical test which works on categorical data .***

*The Chi-Squared test is a statistical hypothesis test that assumes (the null hypothesis) that the observed frequencies for a categorical variable match the expected frequencies for the categorical variable. The test calculates a statistic that has a chi-squared distribution, named for the Greek capital letter Chi (X) pronounced “ki” as in kite.*

We try to test the likelihood of test data(sample data) to find out whether the observed distribution of data set is a statistical fluke(due to chance ) or not. ***“Goodness of fit” statistic*** in chi-square test, measures how well the observed distribution of data fits with the distribution that is expected if the variables are independent.

**How Does Chi -Square Work ?**

Generally, we try to establish a relationship between given categorical variable in this test. Chi-square evaluates whether given variables in a data set(sample) are independent, called the *Test of Independence.* Chi-square tests are used for testing hypotheses about one or two categorical variables, and are ***appropriate when the data can be summarized by counts in a table***. The variables can have multiple categories.

**Type of Chi-Square Test:**

For One Categorical Variable, we perform

* ***Chi-Square Goodness-of-Fit Test***

The chi square goodness of fit test begins by hypothesizing that the distribution of a variable behaves in a particular manner. For example, in order to determine daily staffing needs of a retail store, the manager may wish to know whether there are an equal number of customers each day of the week.

For Two Categorical Variables, we perform

* ***Chi-Square Test for Association***

Another way we can describe the Chi-square test is that:

*It tests the null hypothesis that the variables are independent.*

The test compares the observed data to a model that distributes the data according to the expectation that the variables are independent. Wherever the observed data doesn’t fit the model, the likelihood that the variables are dependent becomes stronger, thus proving the null hypothesis incorrect!

**Hypothesis In Chi-Square:**

The first thing as a data engineer, you need to establish before performing any Inferential statistic test like Chi-Square, is to establish

* *H0: Null Hypothesis*
* *H1: Alternate Hyptotheis*

**For One Categorical Variable:**

* ***Null hypothesis***: The proportions match an assumed set of proportions
* ***Alternative hypothesis***: At least one category has a different proportion. •

**For Two Categorical Variables:**

* ***Null hypothesis***: There is no association between the two variables
* **Alternative hypothesis**: There is an association between the two variable

Before we jump into understanding how Chi-sqaure works with an example, we need to understand what is Chi-square distribution & some other related concepts. This Chi-squared distribution is what we will analyze going forward in **chi-square** or ***χ*2 test.**

**What Is Chi-Square Distribution?**

The **chi-square distribution** (also **chi-squared** or ***χ*2-distribution**) with *k* degrees of freedom is the distribution of a sum of the squares of *k* independent standard normal random variables.

It is one of the most widely used probability distributions in inferential statistics, notably in hypothesis testing or in the construction of confidence intervals.

*The primary reason that the chi-square distribution is used extensively in hypothesis testing is its relationship to the normal distribution. An additional reason that the chi-square distribution is widely used is that it is a member of the class of likelihood ratio tests (LRT).LRT’s have several desirable properties; in particular, LRT’s commonly provide the highest power to reject the null hypothesis.*

**Degree Of Freedom in Chi Squared Distribution:**

The degrees of freedom in Chi Squared distribution is equal to the number of standard normal deviates being summed. The mean of a Chi-square distribution is its degrees of freedom. A chi-square distribution constructed by squaring a single standard normal distribution is said to have 1 degree of freedom

The **degrees of freedom** ( **df** or **d**) tell you how many numbers in your grid are *actually* independent. For a Chi-square grid, the degrees of freedom can be said to be the number of cells you need to fill in before, given the totals in the margins, you can fill in the rest of the grid using a formula.

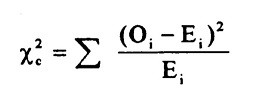
The degrees of freedom for a Chi-square grid are equal to the number of rows minus one times the number of columns minus one: that is, (R-1)\*(C-1).

***Remember!***

*As the degree of freedom (df), increases the Chi-square distribution approaches a normal distribution*

**Chi-Square Statistic:**

The formula for the chi-square statistic used in the chi square test is:



The subscript “**c**” here are the degrees of freedom. “**O**” is your observed value and **E** is your expected value. The summation symbol means that you’ll have to perform a calculation for every single data item in your data set.

***E=(row total×column total) / sample size***

*The Chi-square statistic can only be used on the numbers. They can’t be used for percentages, proportions, means or similar statistical value. For example, if you have 10 percent of 200 people, you would need to convert that to a number (20) before you can run a test statistic.*

Chi-Square test involves calculating a metric called the Chi-square statistic mentioned above, which follows the Chi-square distribution.

Let’s see an example to get clarity on all the above covered topics related to Chi-Square:

**P-Value:**

The null hypothesis provides a probability framework against which to compare our data. Specifically, through the proposed statistical model, the null hypothesis can be represented by a probability distribution called **P-value** , which gives the probability of all possible outcomes if the null hypothesis is true;

*It is a probabilistic representation of our expectations under the null hypothesis.*

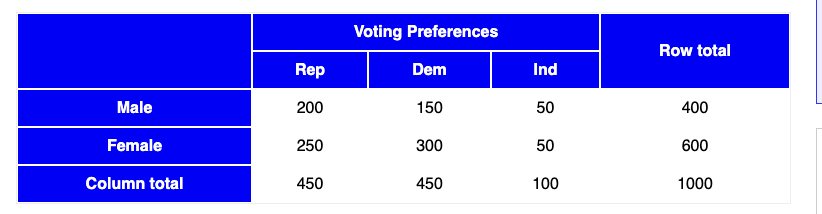
**Chi-Square Test Explained With Example:**

We will cover following important steps in our journey of Chi\_square test for Independence of two variables.

1. *State The Hypothesis*
2. *Formulate Data Analysis Plan*
3. *Analyze The Smaple Data*
4. *Interpret The Outcome*

**Problem: This problem has been sourced from** [**starttrek**](https://stattrek.com/chi-square-test/independence.aspx)

A public opinion poll surveyed a simple random sample of 1000 voters. Respondents were classified by gender (male or female) and by voting preference (Republican, Democrat, or Independent). Results are shown in the contingency table below.



We have to infer, Is there a gender gap? Do the men’s voting preferences differ significantly from the women’s preferences? Use a 0.05 level of significance.

Let’s try to solve this problem using Chi-Square test to find out the P Value.

Here test type which we will employ is :

**Chi-square test for independence.**

So let’s get started by first stating our hypothesis.

**Step 1 : State The Hypothesis:**

Here we need to start by establishing a null hypothesis and counter hypothesis(alternative hypothesis) as given below.

**Null Hypothesis:**

*Ho: Gender and voting preferences are independent.*

**Alternate Hypothesis:**

*H1: Gender and voting preferences are not independent.*

**Step 2: Let’s Build Our Data Analysis Plan :**

Here we will try to find out P Value and match it with the significance level . Let’s take the standard and accepted **level of significance to be 0.05.** Given the sample data in the table above , let’s try to employ *Chi-Square test for independence and deduce the Probability value.*

**Step 3: Let’s Do Sample Analysis:**

Here we will analyze the given sample data to compute

* *Degree of freedom*
* *Expected Frequency Count of sample variable*
* *Calculate Chi-Square test static value*

All the above values will help us find the **P-value**.

**Degree Of Freedom Calculation:** Let’s calculate df = (r — 1) \* (c — 1), so in the given table, we have r(rows)= 2 and c(column) = 3

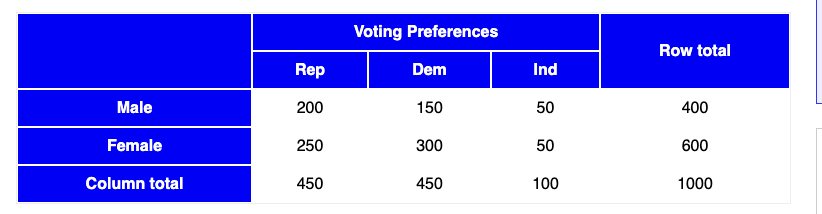
*df= (2–1)\*(3–1) = 1\*2= 2 ;*

**Expected Frequency Count Calculation:**

Let Eij, represent expected values of the two variables are independent of one another.

***Eij = ith (row total X jth column total) / grand total***

Let’s calculate the expected value for each given row and column value by using above mentioned formula, Let me copy the table image again below to help you make calculation easily,



Here, Row 1 total value = 400, total value for column1 = 450, total sample size = 1000,

So ,

***E1,1 = (400 \* 450) / 1000 = 180000/1000 = 180***

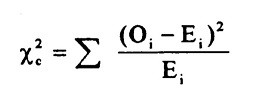
Similarly, lets calculate other expected values as shown below,

***E1,2 = (400 \* 450) / 1000 = 180000/1000 = 180  
E1,3 = (400 \* 100) / 1000 = 40000/1000 = 40  
E2,1 = (600 \* 450) / 1000 = 270000/1000 = 270  
E2,2 = (600 \* 450) / 1000 = 270000/1000 = 270  
E2,3 = (600 \* 100) / 1000 = 60000/1000 = 60***

Time to calculate Chi-Squares for each calculated expected values above using the formula:

**Calculating Chi-Sqaures:**

As already discussed above, the formula for calculating chi-sqaure statistic is



Chi-square statistic formula

The subscript “**c**” here are the degrees of freedom. “**O**” is your observed value (actual values given in the table above)and **E** is your expected value(which we just calculated). The summation symbol means that you’ll have to perform a calculation for every single data item in your data set.

***Χ² = Σ [ (Oi,j — Ei,j)² / Ei,j ]***

Using the above formula our chi-square values comes out to be as given below,

*Χ² = (200–180)²/180 + (150–180)²/180 + (50–40)²/40 + (250–270)²/270 + (300–270)²/270 + (50–60)²/60  
Χ² = 400/180 + 900/180 + 100/40 + 400/270 + 900/270 + 100/60*

So our final chi-square statistic value **,**

**Χ² = 2.22 + 5.00 + 2.50 + 1.48 + 3.33 + 1.67 = 16.2**

Having calculated the chi-square value and degrees of freedom, we consult a chi-square table to check whether the chi-square statistic of 16.2 exceeds the critical value for the Chi-square distribution. The intent is to find *P-value, which is is the probability that a chi-square statistic having 2 degrees of freedom is more extreme than 16.2.*

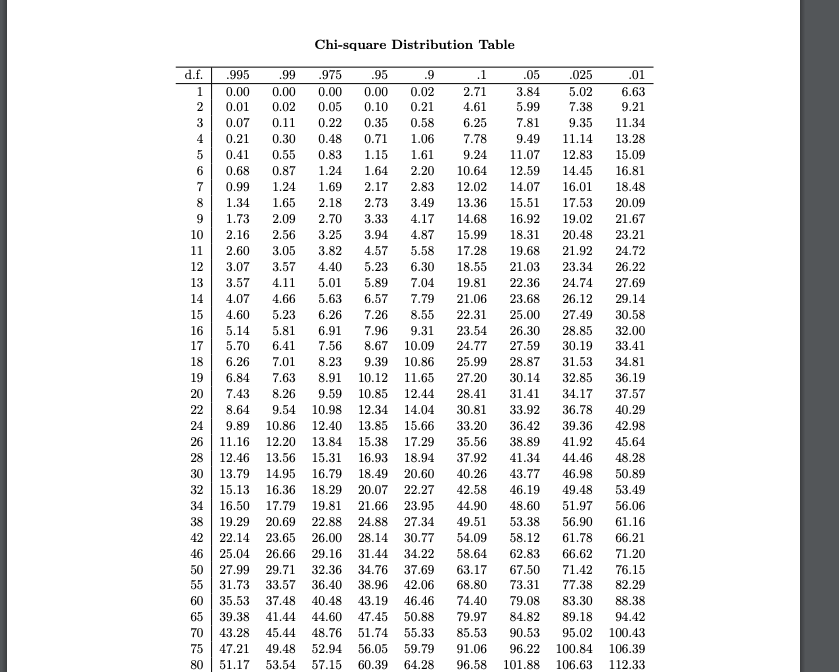
**How to calculate P-value ?**

Given the degree of freedom = 2 & Chi-square statistic value = 16.2 , we can easily find P-value using this given

[Chi-Square Calculator](https://stattrek.com/online-calculator/chi-square.aspx) link , simply enter the Chi-square statistic value & degree of freedom as an input , also keep your significance level as 0.05, you will find the result as given below,

***P-Value is =. 000304. The result is significant at p < .05.***

You can also find P-value using Chi-Square table given below , you can get this table from this [source](https://people.smp.uq.edu.au/YoniNazarathy/stat_models_B_course_spring_07/distributions/chisqtab.pdf)



Chi-square table

Having calculated the chi-square value to be 16.2 and degrees of freedom to be 2 , we consult a chi-square table given above to check whether the chi-square statistic of 16.2 exceeds the critical value for the Chi-square distribution. **The critical value for alpha of .05 (95% confidence) for df=2 comes out to be 5.99**

**Step 4 : Interpreting the result**

***A: Inference From The P-value:***

Since we have got the P-value of 0.000304 we can interpret the result where it signifies that

As the P-value (0.000304) is less than the significance level (0.05),

So we have to reject the below given

**Null Hypothesis**, which says , g*ender and voting preferences are independent.*

& accept **Alternate Hypothesis**:

*Which says , gender and voting preferences are not independent.*

Hence we can conclude that,

*There is a relationship between gender and voting preference.*

***B: Interpreting from Chi-Square Table:***

Since the critical value for alpha of .05 (95% confidence) for df=2 is 5.99 and our chi-square statistic value 16.3, is much larger than 5.99, we have sufficient evidence to reject our Null hypothesis which we covered above.

So we accept Alternate Hypothesis:

*Which says , gender and voting preferences are not independent.*

Hence we conclude that,

*There is a relationship between gender and voting preference.*

**What’s Next ?**

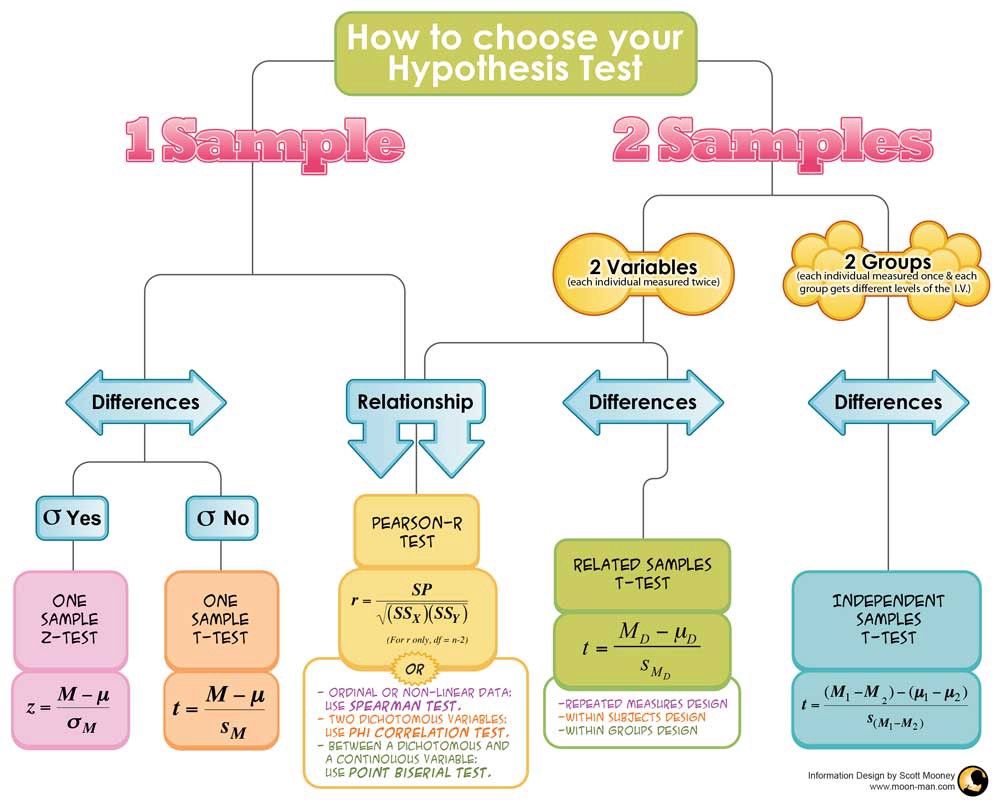
We will understand how to perform Chi-Square test using python & Jupyter notebook in the part 2 of this series of ***Inferential Statistic : Hypothesis testing Using Chi-Square*** *a*nd will further explore

* *Normal Devitate Z Test:*
* *Two Sample T Test*
* ANOVA Test

& also will introduce one of the key topic : “***Power of Statistical Test “***

The power of any test of statistical significance is defined as the probability that it will reject a false null hypothesis.

Summing up this part, with a very helpful infographic which guides you to choose your hypothesis test type:



credit: scott Mooney

So choose your test data wisely and make sure you are interpreting sample data right, so that you can go ahead to design your ML models with required accuracy & confidence.

Your ability to be an effective data scientist will largely become a reality only & only if you know how to analyze the given sample data with minimum deviation. The more you treat data with required precision and clean them in the preliminary stage of EDA, the more reliable and productive your model building effort will become.

**Thanks For Being There As Always & Inspiring Me To Write More…**

**[Towards Data Science](https://towardsdatascience.com/?source=post_sidebar--------------------------post_sidebar-)**

**Sharing concepts, ideas, and codes.**

)