

Chi Squared Test and ANOVA Test

Rabail Tahir

What is Chi Test and where is it used?

- ▶ The Chi-Squared Test is a statistical test used to compare observed data with expected data to determine whether they differ significantly.
- ▶ It helps to find out whether a difference between two categorical variables is due to chance or a relationship between them.
- ▶ The goal is to identify any difference between actual and predicted data.
- ▶ The goal is also to identify whether the difference is due to random luck/chance or is there any variable influencing the results.
- ▶ “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”.

<https://www.simplilearn.com/tutorials/statistics-tutorial/chi-square-test>

Formula For Chi-Square Test

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

Where

c = Degrees of freedom

O = Observed Value

E = Expected Value

Types of Chi squared Test

- ▶ There are two types of Chi-Squared Tests:
 - ▶ Goodness-of-Fit Test
 - ▶ Test of Independence

Goodness of fit test

- The Goodness-of-Fit Test is used to determine whether the observed data fits a particular distribution.
- The test involves comparing the observed frequencies of different categories with the expected frequencies of those categories.
- The Observed values are those you gather yourselves.
- The expected values are the frequencies expected, based on the null hypothesis.

Example

- ▶ Example: A study is conducted to test whether the distribution of blood types in a population follows the expected distribution of blood types in the general population.
- ▶ The expected distribution of blood types in the general population is O (45%), A (40%), B (11%), and AB (4%).
- ▶ The observed distribution of blood types in the study population is O (50%), A (35%), B (10%), and AB (5%).
- ▶ We can use the Chi-Squared Test to determine whether the observed distribution of blood types in the study population fits the expected distribution.

Test of Independence

- ▶ The Test of Independence is used to determine whether there is a significant association between two categorical variables.
- ▶ The test involves comparing the observed frequencies of different categories of the two variables with the expected frequencies of those categories if the two variables were independent.

Example

- ▶ Example: A study is conducted to test whether there is a significant association between smoking status (smoker vs. non-smoker) and lung cancer.
- ▶ We can use the Chi-Squared Test to determine whether there is a significant association between smoking status and lung cancer.

Calculation of Chi-Squared Statistic

- The Chi-Squared Test involves calculating the Chi-Squared Statistic (χ^2).
- The formula for the Chi-Squared Statistic is:
- $\chi^2 = \sum (\text{Observed} - \text{Expected})^2 / \text{Expected}$
- Where:
 - Σ = the sum of
 - Observed = the observed frequency of a category
 - Expected = the expected frequency of a category

Calculation of Expected Frequencies

- To calculate the expected frequencies, we use the formula:
- $\text{Expected} = (\text{row total} * \text{column total}) / \text{total}$

Example- Chi Squared test

- 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
 - ▶ H0: 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 $(O-E)^2 / E$ for Each Cell in the Table

- ▶ Now you will calculate the $(O - E)^2 / E$ for each cell in the table.
- ▶ Where
- ▶ O = Observed Value
- ▶ E = Expected Value

$(O - E)^2 / E$				
	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^2

- ▶ X^2 is the sum of all the values in the last table
- ▶ $= 0.743 + 2.05 + 2.33 + 3.33 + 0.384 + 1$
- ▶ $= 9.837$

Results

- ▶ 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.

Steps of Chi squared Test

- ▶ The steps of a chi test are as follows:
 - ▶ State the null and alternative hypotheses
 - ▶ Select the appropriate chi-square test
 - ▶ Calculate the expected frequencies
 - ▶ Calculate the chi-square statistic
 - ▶ Calculate the degrees of freedom
 - ▶ Determine the critical value
 - ▶ Compare the critical value with the calculated chi-square statistic
 - ▶ Draw a conclusion

Pros and Cons

- ▶ Pros:
- ▶ It is easier to compute.
- ▶ It can also be used with nominal data.
- ▶ It does not assume anything about the data distribution.
- ▶ Cons:
- ▶ The number of observations should be more than 20.
- ▶ Data must be frequency data.
- ▶ It assumes random sampling. It means the sample should be selected randomly.
- ▶ It is sensitive to small frequencies, which leads to erroneous conclusions.
- ▶ It is also sensitive to sample size.

ANOVA Test

- ▶ Analysis of Variance Test, is a set of procedures used to analyze the differences in means of different groups.
- ▶ For example:
 - ▶ A group of psychiatric patients are trying three different therapies: counseling, medication and biofeedback. You want to see if one therapy is better than the others.
 - ▶ A manufacturer has two different processes to make light bulbs. They want to know if one process is better than the other.
 - ▶ Students from different colleges take the same exam. You want to see if one college outperforms the other.

One-Way and Two-Way ANOVA Test

- ▶ One-way and Two-way refers to the amount of variables in your data.
- ▶ One-way has one independent variable . For example: brand of cereal,
- ▶ Two-ways has two independent variables. For example: brand of cereal, calories.
- ▶ Groups or levels are different groups within the same independent variable, for example brand of cereal can have 2 groups, lucky charms and Kellogs.

One way ANOVA test example

- ▶ A study is conducted to test whether there are significant differences in the mean height of people from three different regions: North, South, and West.
- ▶ The following table shows the mean height and standard deviation of each group:
 - North | 172.5 | 6.7
 - South | 167.8 | 8.2
 - West | 170.2 | 5.5
 - We can use One-Way ANOVA to determine whether there are significant differences in the mean height of people from these three regions.

Calculation of Sum of Squares Total (SST)

- ▶ The total variation in the data can be represented by the Sum of Squares Total (SST).
- ▶ The formula for SST is:
- ▶ $SST = \sum (X - \bar{X})^2$
- ▶ Where:
- ▶ \sum = the sum of
- ▶ X = an individual value in the dataset
- ▶ \bar{X} = the mean of all the values in the dataset

Calculation of Sum of Squares Within (SSW)

- ▶ The variation within each group can be represented by the Sum of Squares Within (SSW).
- ▶ The formula for SSW is:
- ▶ $SSW = \sum (X_i - \bar{X}_i)^2$
- ▶ Where:
- ▶ \sum = the sum of
- ▶ X_i = an individual value in a particular group
- ▶ \bar{X}_i = the mean of all the values in that group

Calculation of Sum of Squares Between (SSB)

- ▶ The variation between the groups can be represented by the Sum of Squares Between (SSB).
- ▶ The formula for SSB is:
- ▶ $SSB = \sum N_i (\bar{X}_i - \bar{X})^2$
- ▶ Where:
- ▶ \sum = the sum of
- ▶ N_i = the number of values in a particular group
- ▶ \bar{X}_i = the mean of all the values in that group
- ▶ \bar{X} = the overall mean of all the values in the dataset

Degrees of Freedom

- Degrees of Freedom (df) is a measure of the number of independent pieces of information that went into the estimate of a parameter.

Analysis of Variance(ANOVA)

Source of Variation	Sum of Squares	Degrees of Freedom	Mean Squares (MS)	F
Within	$SS_w = \sum_{j=1}^k \sum_{l=1}^l (X - \bar{X}_j)^2$	$df_w = k - 1$	$MS_w = \frac{SS_w}{df_w}$	$F = \frac{MS_b}{MS_w}$
Between	$SS_b = \sum_{j=1}^k (\bar{X}_j - \bar{X})^2$	$df_b = n - k$	$MS_b = \frac{SS_b}{df_b}$	
Total	$SS_t = \sum_{j=1}^n (\bar{X}_j - \bar{X})^2$	$df_t = n - 1$		

ANOVA Test Steps

- ▶ Step 1: State the null and alternative hypotheses.
 - The null hypothesis is that there are no significant differences between the means of the groups.
 - The alternative hypothesis is that at least one group has a significantly different mean than the others.
- ▶ Step 2: Check assumptions.
 - There are several assumptions that need to be checked before performing ANOVA, including:
 - Normality: The data should be normally distributed within each group.
 - Homogeneity of variances: The variances should be equal across all groups.
 - Independence: The observations in each group should be independent of each other.
- ▶ Step 3: Calculate the test statistic.
 - The test statistic for ANOVA is the F-statistic, which is calculated as the ratio of the variance between the groups to the variance within the groups.

► Step 4: Determine the p-value.

- The p-value is the probability of obtaining a test statistic as extreme or more extreme than the observed value, assuming the null hypothesis is true.

► Step 5: Make a decision.

- If the p-value is less than the chosen level of significance (usually 0.05), we reject the null hypothesis and conclude that there are significant differences between the means of the groups.
- If the p-value is greater than the chosen level of significance, we fail to reject the null hypothesis and conclude that there is not enough evidence to suggest that there are significant differences between the means of the groups.

► Step 6: Interpret the results.

- If we reject the null hypothesis, we need to investigate further to determine which groups have significantly different means from each other.
- We can use post-hoc tests, such as Tukey's HSD test or the Bonferroni correction, to identify significant differences between pairs of groups.

Questions?

