Chi-squared test statistic

- The chi-squared test is used when we want to see if two categorical variables are related
- The test statistic for the Chi-squared test uses the sum of the squared differences between each pair of observed (O) and expected values (E)
- the chi-square test for goodness of fit and the chi-square test for independence.

$$\chi^2 = \sum_{i=1}^{n} \frac{\left(O_i - E_i\right)^2}{E_i}$$

There are two common applications of the Chi-square test:

- **Chi-square Goodness of Fit Test**
- **Chi-square Test of Independence**

1. Chi-Square Goodness of Fit Test:Chi-square goodness of fit test is used to determine if a sample data matches a population with a specific distribution. It tests whether the observed frequency distribution of a single categorical variable matches the expected frequency distribution.

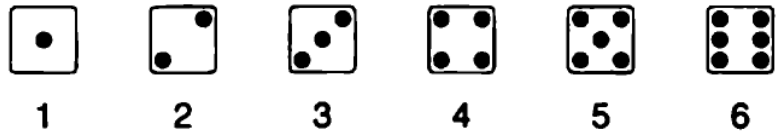
- **Purpose:**
- To test if the observed distribution of data fits a hypothesized distribution (e.g., normal, uniform, etc.).
- It compares the observed data to what would be expected under a given hypothesis.

A die is thrown n times the below table shows the frequency of observed data. Find whether the die is fair or not. With (alpha value =0.05)

Chi-square Test

Observed

#	1	2	3	4	5	6
freq.	22	24	38	30	46	44



Expected

#	1	2	3	4	5	6
freq.						

Chi-square Test

Observed

#	1	2	3	4	5	6	Total
freq.	22	24	38	30	46	44	204

Expected

#	1	2	3	4	5	6
freq.						

$$E(1) = (204)\left(\frac{1}{6}\right) = 34$$

Chi-square Test

Observed

#	1	2	3	4	5	6	Total
freq.	22	24	38	30	46	44	204

Expected

#	1	2	3	4	5	6
freq.	34	34	34	34	34	34

Chi-square Test

Observed

#	1	2	3	4	5	6
freq.	22	24	38	30	46	44

Expected

#	1	2	3	4	5	6
freq.	34	34	34	34	34	34

H_0: Die is fair

H_1: Die is unfair

Chi-square table

df	$\chi^2_{.995}$	$\chi^2_{.990}$	$\chi^2_{.975}$	$\chi^2_{.950}$	$\chi^2_{.900}$	$\chi^2_{.100}$	$\chi^2_{.050}$	$\chi^2_{.025}$	$\chi^2_{.010}$	$\chi^2_{.005}$
1	0.000	0.000	0.001	0.004	0.016	2.706	3.841	5.024	6.635	7.879
2	0.010	0.020	0.051	0.103	0.211	4.605	5.991	7.378	9.210	10.597
3	0.072	0.115	0.216	0.352	0.584	6.251	7.815	9.348	11.345	12.838
4	0.207	0.297	0.484	0.711	1.064	7.779	9.488	11.143	13.277	14.860
5	0.412	0.554	0.831	1.145	1.610	9.236	11.070	12.833	15.086	16.750

$$\chi^2 = \sum \frac{(Oij - Eij)^2}{Eij}$$

$$\chi^2 = \frac{(22-34)^2}{34} + \frac{(24-34)^2}{34} + \frac{(38-34)^2}{34} + \frac{(30-34)^2}{34} +$$

$$\frac{(46-34)^2}{34} + \frac{(44-34)^2}{34} = 15.29$$

15.29>11.07 So H0 is rejected.

There is sufficient evidence to suggest that the die is **not fair** at the 0.05 significance level.

2. Chi-Square Test of Independence: The **Chi-square test of independence** is used to determine whether there is an association between two categorical variables. It tests whether the distribution of one variable is independent of the distribution of the other variable.

Purpose:

- To test if two categorical variables are independent or if there is a relationship between them.
- It is used when you have two categorical variables and want to see if they are associated with each other.

- A survey was conducted among 200 students to see if there is an association between gender and choice of study program (Science or Arts). The survey results are summarized in the table below:

Gender	Science (S)	Arts (A)	Total
Male	60	40	100
Female	50	50	100
Total	110	90	200

- We will apply the **chi-square test** to test if there is an association between gender and study program

- **Step 1: Define the Hypotheses**
- **Null Hypothesis (H_0):**There is **no association** between gender and choice of study program. In other words, the two variables (gender and study program) are **independent**.

  H0:Gender and Study Program are independent.
- **Hypothesis (H_1):**There is **an association** between gender and choice of study program. The two variables are **not independent**.

  H1:Gender and Study Program are not independent.

Step 2: Calculate Expected Frequencies

The expected frequency for each cell is calculated using the formula:

$$E_i = \frac{(\text{Row Total}) \times (\text{Column Total})}{\text{Grand Total}}$$

For Male, Science:

$$E_{\text{Male, S}} = \frac{100 \times 110}{200} = 55$$

For Male, Arts:

$$E_{\text{Male, A}} = \frac{100 \times 90}{200} = 45$$

For Female, Science:

$$E_{\text{Female, S}} = \frac{100 \times 110}{200} = 55$$

For Female, Arts:

$$E_{\text{Female, A}} = \frac{100 \times 90}{200} = 45$$

- Now, the expected frequency table looks like this:

Gender	Science (S)	Arts (A)	Total
Male	55	45	100
Female	55	45	100
Total	110	90	200

Step 3: Calculate the Chi-Square Statistic

The chi-square statistic is calculated as:

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

Where O_i is the observed frequency, and E_i is the expected frequency.

For Male, Science:

$$\frac{(60-55)^2}{55} = \frac{5^2}{55} = \frac{25}{55} \approx 0.45$$

For Male, Arts:

$$\frac{(40-45)^2}{45} = \frac{(-5)^2}{45} = \frac{25}{45} \approx 0.56$$

For Female, Science:

$$\frac{(50-55)^2}{55} = \frac{(-5)^2}{55} = \frac{25}{55} \approx 0.45$$

For Female, Arts:

$$\frac{(50-45)^2}{45} = \frac{5^2}{45} = \frac{25}{45} \approx 0.56$$

Now, sum up the values:

$$\chi^2 = 0.45 + 0.56 + 0.45 + 0.56 = 2.02$$

Step 4: Degrees of Freedom

The degrees of freedom (df) for this test is calculated as:

$$df = (R - 1)(C - 1) = (2 - 1)(2 - 1) = 1$$

Step 5: Compare with the Critical Value

At a 5% significance level ($\alpha = 0.05$) and $df = 1$, the critical value from the chi-square table is **3.841**.

Since the calculated chi-square value $\chi^2 = 2.02$ is **less** than the critical value 3.841, we **fail to reject** the null hypothesis.

Conclusion:

There is **no sufficient evidence** to suggest that gender and choice of study program are associated. Therefore, we conclude that **gender and study program are independent**.

T-test

- A **t-test** is a statistical method used to determine if there is a significant difference between the means of two groups or a sample mean and a population mean. The t-test is used when the sample size is small (typically less than 30), and the population standard deviation is unknown.
- There are different types of t-tests, but the most common ones are:
- **One-sample t-test**: Compares the sample mean to a known population mean.
- **Independent two-sample t-test**: Compares the means of two independent groups.
- **Paired t-test**: Compares the means of two related groups (e.g., before and after treatment on the same group of subjects).

Steps for Conducting a t-test

- **State the hypotheses**:
  - **Null hypothesis (H_0)**: There is no difference (or no effect).
  - **Alternative hypothesis (H_1)**: There is a difference (or effect).
- **Select the significance level (α)**: This is usually 0.05 (5%).
- **Calculate the test statistic (t)**: Using the sample data, you calculate the t-value using the appropriate formula.
- **Determine the degrees of freedom (df)**: For a one-sample t-test, the degrees of freedom are calculated as n−1, where n is the sample size. For two-sample t-tests, degrees of freedom depend on the sample sizes of both groups.
- **Compare the calculated t-value to the critical t-value**: The critical t-value is obtained from the t-distribution table based on the significance level (α) and degrees of freedom.
- **Make a decision**:
  - If the calculated t-value is greater than the critical value (for one-tailed tests) or falls in the rejection region (for two-tailed tests), reject the null hypothesis.
  - If the calculated t-value is less than the critical value, fail to reject the null hypothesis.

Types of t-tests

1. One-Tailed t-test (Directional)

- A one-tailed t-test is used when you are testing for the possibility of an effect in only one direction. This means that you're interested in whether the sample mean is either greater than or less than the population mean, but not both.
- **Right-tailed test**: Tests if the sample mean is significantly greater than the population mean.
- **Left-tailed test**: Tests if the sample mean is significantly smaller than the population mean.

2. Two-Tailed t-test (Non-directional)

- A two-tailed t-test is used when you're testing for the possibility of an effect in either direction — that is, you are testing whether the sample mean is significantly different from the population mean, either higher or lower.

Example of a One-Tailed t-test:

- Let's say a teacher believes that a new teaching method will increase the average test score from 70 to 75. She tests this with a sample of 25 students, and the sample mean is 73 with a sample standard deviation of 5. We want to test if the new method has increased the average score at the 5% significance level.
- **Null hypothesis (H_0)**: The new method has no effect (mean = 70).
- **Alternative hypothesis (H_1)**: The new method increases the mean (mean > 70).

- **t-test formula**:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

  Where:

  - $\bar{x}$ = sample mean (73)
  - μ = population mean (70)
  - s = sample standard deviation (5)
  - n = sample size (25)

- **Calculate the t-value**:

$$t = \frac{73 - 70}{\frac{5}{\sqrt{25}}} = \frac{3}{1} = 3.00$$

- **Degrees of freedom**: $n - 1 = 25 - 1 = 24$
- **Critical t-value** for a right-tailed test with $\alpha = 0.05$ and $df = 24$ is approximately 1.711 (from the t-distribution table).

Since $t = 3.00$ is greater than the critical t-value of 1.711, we **reject the null hypothesis**. This means that the new method significantly increased the test score.

Example of a Two-Tailed t-test:

- Suppose a company claims that their average employee work hours per week is 40 hours. A sample of 30 employees is taken, and their average work hours are found to be 42 with a standard deviation of 4 hours. We want to test whether the average work hours differ from 40 at the 5% significance level.
- **Null hypothesis (H_0)**: The average work hours are 40 hours.
  Alternative hypothesis (H_1): The average work hours are not 40 hours.

- **t-test formula**:

$$t = \frac{\bar{x} - \mu}{\frac{s}{\sqrt{n}}}$$

  Where:

  - $\bar{x}$ = sample mean (42)
  - μ = population mean (40)
  - s = sample standard deviation (4)
  - n = sample size (30)

- **Calculate the t-value**:

$$t = \frac{42 - 40}{\frac{4}{\sqrt{30}}} = \frac{2}{0.730} \approx 2.74$$

- **Degrees of freedom**: $n - 1 = 30 - 1 = 29$
- **Critical t-value** for a two-tailed test with $\alpha = 0.05$ and $df = 29$ is approximately ±2.045 (from the t-distribution table).

Since the calculated t-value (2.74) is greater than the critical value (2.045), we **reject the null hypothesis**. This means there is a significant difference between the observed mean and the hypothesized mean of 40 hours.

Z test

The **Z-test** is a statistical test used to determine if there is a significant difference between the sample mean and the population mean (or between two sample means). It is particularly useful when the sample size is large (typically $n>30$), and the population variance is known or the sample variance is a good approximation of the population variance.

Types of Z-tests:

- **One-Sample Z-test**: Used to compare the mean of a single sample to the population mean when the population variance is known.
- **Two-Sample Z-test**: Used to compare the means of two independent samples when the population variances are known.
- **Z-test for Proportions**: Used to compare sample proportions to population proportions or to compare the proportions between two groups.

Z-test Formula:

The general formula for the **Z-statistic** is:

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

Where:

- $\overline{X}$ = Sample mean
- μ = Population mean (or hypothesized population mean)
- σ = Population standard deviation
- n = Sample size
- $\frac{\sigma}{\sqrt{n}}$ = Standard error of the sample mean

One-Sample Z-Test

Problem: You want to test if the average height of students in a school is different from the national average of 160 cm. A random sample of 100 students from the school has a mean height of 162 cm, and the population standard deviation of height is known to be 10 cm.

- **Population mean (μ)** = 160 cm
- **Sample mean ($\overline{X}$)** = 162 cm
- **Population standard deviation (σ)** = 10 cm
- **Sample size (n)** = 100

1. **State the Hypotheses**:
   - H_0: $\mu = 160$ (The mean height is 160 cm)
   - H_1: $\mu \neq 160$ (The mean height is different from 160 cm)
2. **Set the significance level**:
   - $\alpha = 0.05$

3. **Calculate the Z-Statistic**:

$$Z = \frac{\overline{X} - \mu}{\frac{\sigma}{\sqrt{n}}}$$

$$Z = \frac{162 - 160}{\frac{10}{\sqrt{100}}}$$

$$Z = \frac{2}{\frac{10}{10}} = \frac{2}{1} = 2$$

4. **Find the critical value**: For a two-tailed test with $\alpha = 0.05$, the critical value is ± 1.96.
5. **Make a Decision**: Since the calculated $Z = 2$ is greater than the critical value 1.96, we reject the null hypothesis. This indicates that the mean height of students in this school is significantly different from the national average of 160 cm.

Example of a Two-Sample Z-Test:

Problem:

A researcher wants to test whether the average heights of students in two different schools are significantly different. School 1 has a sample of 100 students with a sample mean height of 170 cm and a population standard deviation of 10 cm. School 2 has a sample of 120 students with a sample mean height of 172 cm and a population standard deviation of 12 cm. We will conduct a two-sample Z-test at a significance level of $\alpha = 0.05$.

- **Sample 1 (School 1)**:
  - Sample size (n_1) = 100
  - Sample mean ($\overline{X_1}$) = 170 cm
  - Population standard deviation (σ_1) = 10 cm
- **Sample 2 (School 2)**:
  - Sample size (n_2) = 120
  - Sample mean ($\overline{X_2}$) = 172 cm
  - Population standard deviation (σ_2) = 12 cm

1. **State the Hypotheses**:
   - H_0: The mean height of students in School 1 is equal to the mean height of students in School 2 ($\mu_1 = \mu_2$).
   - H_1: The mean height of students in School 1 is not equal to the mean height of students in School 2 ($\mu_1 \neq \mu_2$).
2. **Set the Significance Level**:
   - $\alpha = 0.05$.
3. **Calculate the Z-Statistic**:

$$Z = \frac{\overline{X_1} - \overline{X_2}}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_2^2}{n_2}}}$$

Substitute the values:

$$Z = \frac{170 - 172}{\sqrt{\frac{10^2}{100} + \frac{12^2}{120}}}$$

$$Z = \frac{-2}{\sqrt{\frac{100}{100} + \frac{144}{120}}}$$

$$Z = \frac{-2}{\sqrt{1 + 1.2}} = \frac{-2}{\sqrt{2.2}} = \frac{-2}{1.483}$$

$$Z = -1.35$$

4. **Determine the Critical Value**: For a two-tailed test with $\alpha = 0.05$, the critical value is $Z_{\text{critical}} = \pm 1.96$ (from the standard normal distribution).
5. **Make a Decision**: The calculated Z-value is -1.35, which is between -1.96 and +1.96. Since the calculated Z-value does not fall in the rejection region, we **fail to reject** the null hypothesis.

Conclusion:

Based on the two-sample Z-test, there is no significant difference between the mean heights of students in School 1 and School 2 at the 5% significance level.

ANOVA

- ANOVA is a parametric statistical technique that helps in finding out if there is a significant difference between the mean of two or more groups.
- Steps

1. Define the null and alternative hypothesis.

H0 -> μ1 = μ2 = μ3 (where μ = mean)

Ha -> At least one difference among the means.

2. Find the degree of freedom between and within the groups.

$$df_{between} = k - 1 \qquad df_{within} = n - k$$

where
n = number of samples in all groups combined.
k = number of groups.

- It is defined as the ratio of the variance between samples to variance within samples. It is obtained while performing ANOVA test

$$F_{value} = \frac{variance_{between-samples}}{variance_{within-samples}}$$

Step 3 - Refer the F-Distribution table and find F_{table} using $df_{between}$ and df_{within}.

As per the given F-Distribution table,

$df_1 = df_{between}$

$df_2 = df_{within}$

Step 4 - Find the mean of all samples in each group.
Then use **Eq-5** to find the Grand mean. (μ_{Grand})

$$\mu_{Grand} = \frac{\sum G}{n}$$

Step 5 - Find the sum of squares total using **Eq-6** and sum of squares within using
Then find sum of squares between using **Eq-8.**

$$SS_{total} = \sum (x_i - \mu_{grand})^2$$

Eq-6

where, x_i = i_{th} sample

$$SS_{within} = \sum (x_i - \mu_i)^2$$

Eq-7

where,
x_i = i_{th} sample.
μ_i = mean of i_{th} group.

$$SS_{between} = SS_{total} - SS_{within}$$

$$S^2_{between} = \frac{SS_{between}}{df_{between}}$$

Eq-9

$$S^2_{within} = \frac{SS_{within}}{df_{within}}$$

Eq-10

Step 7 - Find Fcalc using Eq-11.

$$F_{calc} = \frac{S^2_{between}}{S^2_{within}}$$

Interpreting the results

if Fcalc < Ftable :

    Don't rejct null hypothesis.

    $\mu_1 = \mu_2 = \mu_3$

if Fcalc > Ftable :

    Reject null hypothesis.

Consider the example given below to understand step by step how to perform this test. The marks of 3 subjects (out of 5) for a group of students is recorded. (as given in the table below)

[Take $\alpha = 0.05$]

Std/Sub	English (e)	Math (m)	Science (s)
Student 1	2	2	1
Student 2	4	3	2
Student 3	2	4	5

Step 1 -

Null hypothesis, H_0 -> $\mu_E = \mu_M = \mu_S$ (where μ = mean)

Alternate hypothesis, H_a -> At least one difference among the means of the 3 subje

Step 2 -

As per the table,

k = 3

n = 9

$df_{between} = 3 - 1 = 2$ **[Eq-1]**

$df_{within} = 9 - 3 = 6$ **[Eq-2]**

$df_{total} = 2 + 6 = 8$ **[Eq-3]**

Step 3 - On referring to the F-Distribution table (link), using $df_1 = 2$ and $df_2 = 6$ at $\alpha = 0.05$: we get, $F_{table} = 5.14$

Step 4 - $\mu_e = (2 + 4 + 2)/3 = (8/3) = 2.67$

$\mu_m = (2 + 3 + 4)/3 = (9/3) = 3.00$

$\mu_s = (1 + 2 + 5)/3 = (8/3) = 2.67$

$\mu_{grand} = (8 + 8 + 9)/9 = (25/9) = 2.78$

Step 5 - $SS_{total} = (2 - 2.78)^2 + (4 - 2.78)^2 + (2 - 2.78)^2 + (2 - 2.78)^2 + (3 - 2.78)^2 + (4 - 2.78)^2 + (1 - 2.78)^2 + (2 - 2.78)^2 + (5 - 2.78)^2$

$= \mathbf{13.60}$

$SS_{within} = (2 - 2.67)^2 + (4 - 2.67)^2 + (2 - 2.67)^2 + (2 - 3.00)^2 + (3 - 3.00)^2 + (4 - 3.00)^2 + (1 - 2.67)^2 + (2 - 2.67)^2 + (5 - 2.67)^2$

$= \mathbf{13.34}$

$SS_{between} = 13.60 - 13.34 = \mathbf{0.23}$

Step 6 - $S^2_{between} = (0.23/2) = 0.12$

$S^2_{within} = (13.34/6) = 2.22$

Step 7 - $F_{calc} = (0.12/2.22) = 0.05$

Since, $F_{calc} < F_{table}$ $(0.05 < 5.14)$

we **cannot reject** the null hypothesis.

ANOVA Test

Example:
The following table shows the salaries of randomly selected individuals from four large metropolitan areas. At $\alpha = 0.05$, can you conclude that the mean salary is different in at least one of the areas? (Adapted from US Bureau of Economic Analysis)

Pittsburgh	Dallas	Chicago	Minneapolis
27,800	30,000	32,000	30,000
28,000	33,900	35,800	40,000
25,500	29,750	28,000	35,000
29,150	25,000	38,900	33,000
30,295	34,055	27,245	29,805

Continued.

Example continued:

H_0: $\mu_1 = \mu_2 = \mu_3 = \mu_4$

H_a: At least one mean is different from the others. (Claim)

Because there are k = 4 samples, $\text{d.f.}_N = k - 1 = 4 - 1 = 3$.

The sum of the sample sizes is
$N = n_1 + n_2 + n_3 + n_4 = 5 + 5 + 5 + 5 = 20$.

$\text{d.f.}_D = N - k = 20 - 4 = 16$

Using $\alpha = 0.05$, $\text{d.f.}_N = 3$, and $\text{d.f.}_D = 16$,
the critical value is $F_0 = 3.24$.

Continued.

Example continued:

To find the test statistic, the following must be calculated.

$$\bar{\bar{x}} = \frac{\sum x}{N} = \frac{140745 + 152705 + 161945 + 167805}{20} = 31160$$

$$MS_B = \frac{SS_B}{\text{d.f.}_\text{N}} = \frac{\sum n_i(\bar{x}_i - \bar{\bar{x}})^2}{k-1}$$

$$= \frac{5(28149 - 31160)^2 + 5(30541 - 31160)^2}{4-1} +$$

$$\frac{5(32389 - 31160)^2 + 5(33561 - 31160)^2}{4-1}$$

$$\approx 27874206.67$$

Example continued:

$$MS_W = \frac{SS_W}{\text{d.f.}_D} = \frac{\sum(n_i - 1)s_i^2}{N - k}$$

$$\approx \frac{(5-1)(3192128.94) + (5-1)(13813030.08)}{20-4} +$$

$$\frac{(5-1)(24975855.83) + (5-1)(17658605.02)}{20-4}$$

$$= 14909904.97$$

$$F = \frac{MS_B}{MS_W} = \frac{27874206.67}{14909904.34} \approx 1.870$$

Test statistic: 1.870 < Critical value: 3.24.

Fail to reject H_0.

There is not enough evidence at the 5% level to conclude that the mean salary is different in at least one of the areas.