



Chi-square Test

Please note that any topics that are not covered in today's lecture will be covered in the next lecture.

✓ Content

1. Degrees of freedom
2. Chi-square Goodness of fit test
3. Chi-square Test for Independence

Let's recall our conceptual learnings in the T-test what was the setup for two variables?

- 1) Numerical Vs Numerical
- 2) Numerical Vs Categorical
- 3) Categorical Vs Categorical

- **Numerical Vs Categorical (2 Categories):** we use Two sample T-test. It helps us see if there's a difference in the numerical values between these two categories.
- **Numerical Vs Categorical (>2 Categories):** we use the ANOVA test. This test helps us figure out if there's a significant difference among the numerical values when we have more than two categories to compare.
 - We will learn about this in future lectures.
- **Numerical Vs Numerical:** if these two variables are related, so we use a correlation test. It helps us figure out if changes in one variable are connected to changes in the other variable.
 - We will learn about this in future lectures.
- **Categorical Vs Categorical:** we use the Chi-square test. It's like a detective tool that helps us find out if there's a relationship or connection between these categories.
- In today's class, we're diving into **Categorical Vs Categorical** using the **Chi-square test**. This test helps us investigate if there's a meaningful connection between two categorical things.

Before starting with the topic let's have a look into one example:

Example: Movie Nights

Consider planning a week of movie nights with your friends.

You have a collection of seven different movie genres to choose from.

To make things interesting, you decide that each movie night will feature a unique genre

- On the first night, you have the freedom to choose from all seven genres.
- On the second night, your options reduce to six, as you've already picked one.
- This pattern continues until the sixth night when you have only one remaining genre to choose from.
- By the seventh night, your choice is predetermined since there's only one genre left.

In this scenario, you had the freedom to choose a movie genre on six of the seven nights.

The restriction you imposed on having unique genres each night influenced your choices, making the last night's choice dependent on the previous selections.

In statistics, degrees of freedom represent the number of values or quantities in a statistical calculation that are free to vary. It's like the flexibility or choices we have when making decisions in our statistical calculations.

Why should we Learn DOF before Chi-square?

- Understanding degrees of freedom (DOF) is crucial before learning the Chi-square test because it helps us grasp the flexibility or constraints in our data analysis.
- In the example above, knowing the degrees of freedom helps us understand how much we can freely determine before the data becomes more fixed or constrained.
- It's like knowing the rules before playing a game.
- The Chi-square test, which investigates connections between categorical variables, involves degrees of freedom to ensure our analysis is reliable and meaningful.

So let's start our today's lecture with the topic Degrees of freedom

✓ Degrees of freedom

Let's dive into the concept of **degrees of freedom**, which might sound complex, but we'll make it simple with an example involving salaries.

Example: Salaries

Setup 1:

Imagine you have information about people's salaries.

You know the first person's salary is 35 lakhs, the second person's salary is 36 lakhs, and the third person's salary is missing.

But, you do know that the average (or mean) salary is 35 lakhs.

Can you figure out the missing salary?

Yes, it's 34 lakhs.

Because if the average is 35 lakhs and two people already have 35 and 36 lakhs, the third one must be 34 lakhs to make the average 35.

Setup 2:

Now, let's say you have more salary data.

You know the first person's salary is 35 lakhs, the second person's salary is 36 lakhs, the third person's salary is missing, and the fourth person's salary is 30 lakhs.

Surprisingly, the average salary is 37 lakhs.

Can you find the missing salary?

Yes, the missing salary is 47 lakhs.

To have an average of 37 with 35, 36, and 30 as the other salaries, the fourth one must be 47.

General Rule: Degrees of Freedom

Now, think about a general rule.

If you have a set of n numbers and you know the average of those numbers, how many of these numbers do you need to know to determine the full set?

It's $n - 1$.

- This number, $n - 1$, is what we call "**degrees of freedom**".

In simpler words, when you know the average and you have a bunch of numbers, you can pick $n - 1$ numbers freely.

The last one will automatically be determined by the average and the other known numbers.

Degrees of freedom help us understand how much flexibility we have in a dataset when we know its average.

Setup 3:

In the scenario where we have data for both height and weight, along with their respective average values.

To completely determine the entire dataset of 10 numbers (5 heights and 5 weights).

Among these 10 numbers, what will be the bare minimum numbers we should know to be able to determine the full dataset?

A handwritten table on a black background with yellow text and lines. The table has two columns, 'H' (Height) and 'W' (Weight), separated by a vertical line. A horizontal line is above the data rows, and another is below the averages. The 'H' column contains five values: 73, 68, 74, 71, and 62. The first four values are enclosed in a yellow rounded rectangle. Below these is the average '71'. The 'W' column contains five values: 85, 73, 96, 82, and 70. The first four values are enclosed in a yellow rounded rectangle. Below these is the average '81.2'. In the bottom left corner, the word 'avg:' is written.

H	W
73	85
68	73
74	96
71	82
62	70
71	81.2

avg:

In this scenario, to fully determine the entire dataset consisting of two arrays, we only need to be aware of a minimum of **8 values** from either of the arrays.

This understanding can be expressed mathematically using degrees of freedom.

For two arrays with lengths n_1 and n_2 , the degrees of freedom (dof) is calculated as the sum of the degrees of freedom for each array: $(n_1 - 1) + (n_2 - 1)$, which can be further simplified to $n_1 + n_2 - 2$.

This concept becomes particularly useful when we delve into hypothesis testing.

In the context of one-sample T-test and two-sample T-test, we often compare data with their respective averages. The associated degrees of freedom play a crucial role here.

For a two-sample T-test, the degrees of freedom become $n_1 + n_2 - 2$, while for a one-sample T-test, it simplifies to $n - 1$, where 'n' is the sample size.

Setup 4:

Now, let's look at the data on Sachin's centuries and victories.

Sachin : Centuries & Victory

		Win	
		F	T
Cent	F		314
	T		46
		176	184
			360

Sachin : Centurio & Victory

Win

Both row sum & column sum

		F	T	
Cent	F	160	154	314
	T	16	30	46
		176	184	360

Single no was enough
dof = 1

When considering the provided data, a single value is all that's required to build the entire table. This means that having just one value allows us to deduce the rest of the values within the table. The difference between this scenario and the previous examples lies in the information we possess.

Here, we are armed not only with the averages of individual values but also with the sum values of both rows and columns.

Why is DOF = 1 here?

In this scenario we have sum of rows and columns

- Let's consider first row which has total 314.
- In this scenario if we have 1 value out of 2 then we can easily calculate the 2nd value also (total - 1st value = 2nd value)
 - We are giving 1st value of first row (160), so 2nd value will be $(314 - 160 = 154)$
 - Now by given only one value (160) we are able to complete 1st row (160, 154) and single single values in both columns also.
- For the first column by using the first value of first row (160) we can able to calculate remaining value of first column $(176 - 160 = 16)$
- Same for the second column by using second value of first row (154) that we calculated and sum of second column (184) we can able to calculate remaining value of second column $(184 - 154 = 30)$

We can observe that in this scenario we only need a **"Single value"** to complete the entire array, so $\text{DOF} = 1$ here.

Given this setup, the degrees of freedom for this context simplify to just 1.

This means that with the sum of information available, we have enough constraints to determine all the remaining values within the table.

Setup 5:

Do you all agree with me when I say different politicians are famous in different regions?

The same politician will not be famous in Karnataka, Andhra or Maharashtra.

Let's take an example of regional support there 4 politicians -> A, B, C, and D and we have done a survey in three different cities -> X, Y, Z.

	A	B	C	D	
X					349
Y					151
Z					150
	150	150	200	150	650

Now, that we have the overall data from the survey we conducted what are the bare minimum values we need to complete the data?

Regional Support for politician

4 politicians \rightarrow A, B, C, D

3 cities \rightarrow X, Y, Z

	A	B	C	D	
X	90	60	104		349
Y	30	50	51		151
Z					150
	150	150	200	150	650

dof = 6

$$(\# \text{ row} - 1)(\# \text{ col} - 1)$$

In this particular situation, having just **6 values** at our disposal is sufficient to construct the entire table.

The key lies in recognizing the pattern formed by the data arrangement.

When we closely examine the data, we see the emergence of a small rectangle.

- The height of this rectangle, which contributes to the determination of the number of rows, is obtained by subtracting 1 from the total number of rows.
- Similarly, the width of the rectangle, which plays a role in the number of columns, is derived by subtracting 1 from the total number of columns.

Degrees of freedom are then calculated by multiplying these two dimensions together: $(\# \text{ rows} - 1) * (\# \text{ columns} - 1)$.

In this case, with 3 rows and 4 columns, it's $(3 - 1) * (4 - 1)$, resulting in $2 * 3$, which gives us 6.

Now, why is this degree of freedom important?

1. In the context of the Chi-square test, degrees of freedom represent the number of categories the value influences the critical values used to determine statistical significance.
 - As degrees of freedom increase, the chi-squared distribution changes shape.

- Higher degrees of freedom lead to higher critical values, requiring a larger test statistic to reject the null hypothesis at a given significance level (α).
2. A higher degree of freedom allows for more variability and flexibility in the distribution of the test statistic.
 3. Degrees of freedom help define the expected distribution of the test statistic under the null hypothesis.
 - The expected distribution is a key reference point for evaluating the observed test statistic and determining whether deviations are statistically significant.

✓ Chi-square Goodness of fit test

✓ Use Case: Coin Toss

The Chi-squared test for goodness of fit is used when you have **one categorical variable**, and you want to see if the observed frequencies match the expected frequencies.

Now, to understand it let's start with a simple example of a coin toss which we all relate to.

50 times → test if coin is fair

	H	T
Expected	25	25
Actual	28	22

def = 1

In the coin toss, we toss the coin 50 times to test whether the coin is fair or not.

Here, **degrees of freedom = 1**

Now, to do the hypothesis test to determine whether the coin is fair or not.

- H_0 : The Coin is Fair.

- H_1 : The Coin is Biased.

A new term comes into play when we're assessing fairness, and this hinges on the assumption of the null hypothesis.

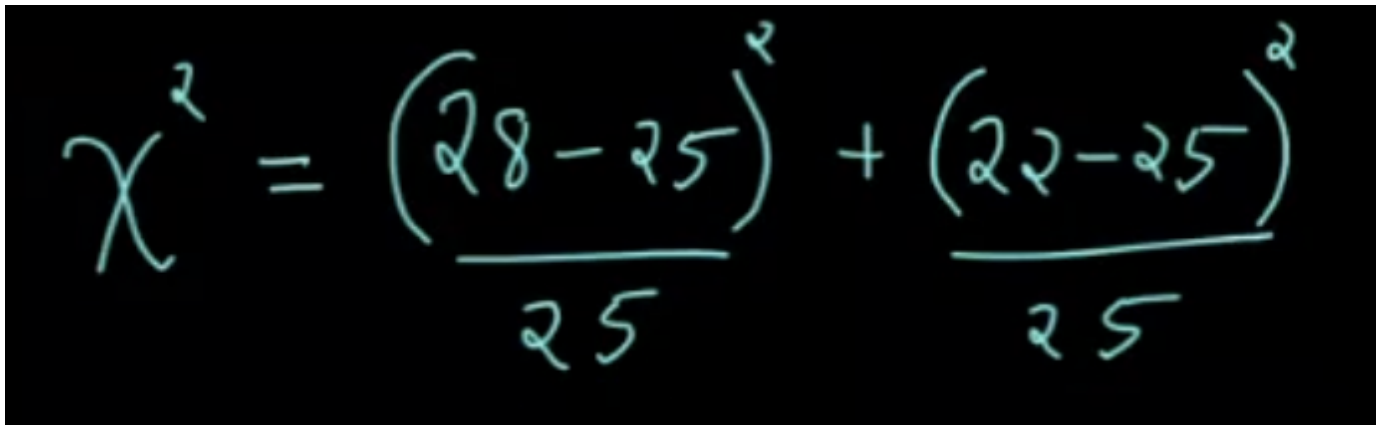
In the context of the null hypothesis, the difference between the actual value and the expected value tends to be **small**.

However, within this small difference, it's important to note whether it's a **positive** or **negative** difference.

Adding another layer to the process, we square the data, which again emphasizes the smallness of differences.

But remember, the differences could still be either positive or negative.

Further along the process, we take these squared differences and divide them by the expected value on both sides.


$$\chi^2 = \frac{(28-25)^2}{25} + \frac{(22-25)^2}{25}$$

Visualizing this process adds another layer of clarity.

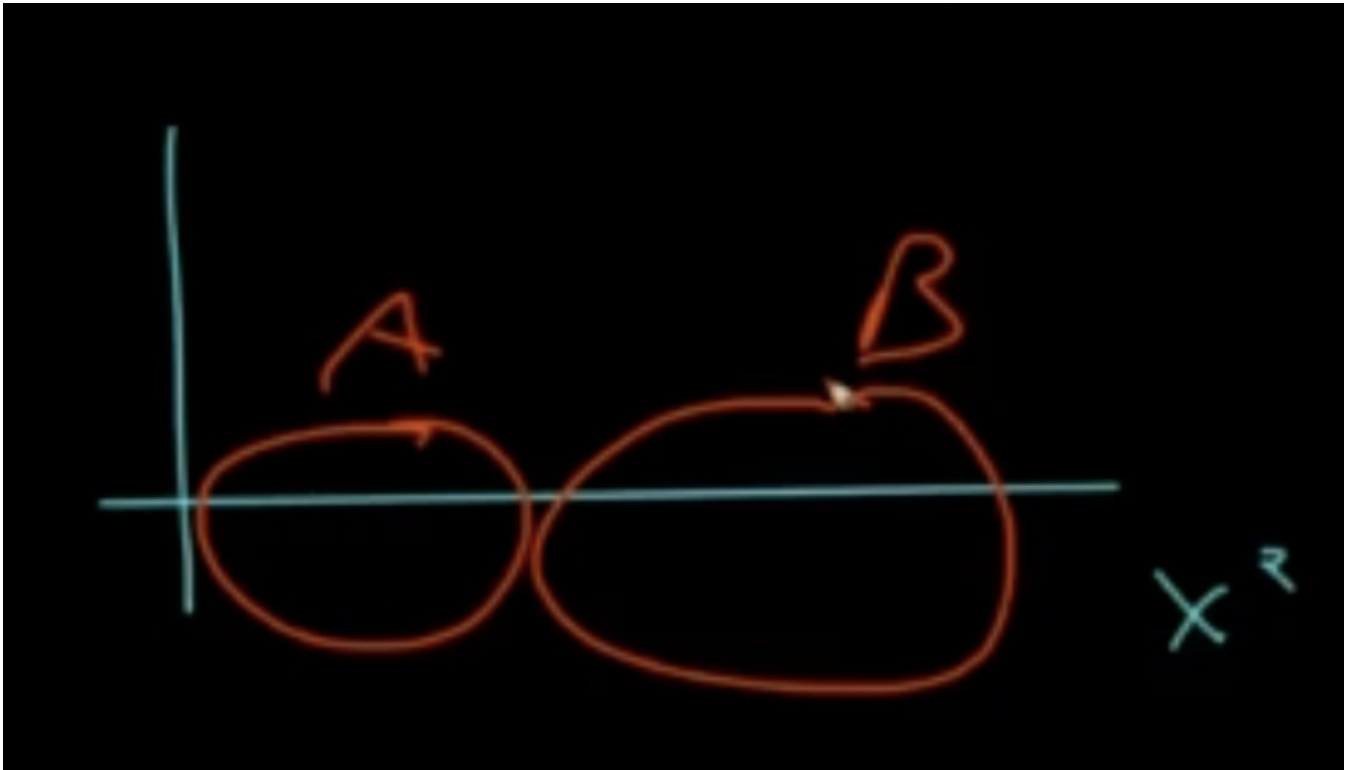
When we attempt to plot these values, we can position the values corresponding to the null hypothesis and alternative hypothesis in specific locations.

Let's walk through it:

- In the null hypothesis, where small differences are anticipated between actual and expected values,
 - The data points will tend to cluster around a certain point on the plot.
- On the other hand, for the alternative hypothesis, which suggests a substantial difference between actual and expected values,
 - The data points will likely be dispersed further from that central point.

Now, consider that this series of experiments is repeated multiple times (let's say 1000 times).

- For each set of experiments, we create a histogram to showcase the distribution of values obtained.



In the image provided, two distinct regions emerge.

When dealing with a fair coin, the chi-square values tend to cluster more in **Region A**.

- This is because the majority of the time, the differences between the observed and expected values are relatively small, aligning with the characteristics of fairness.

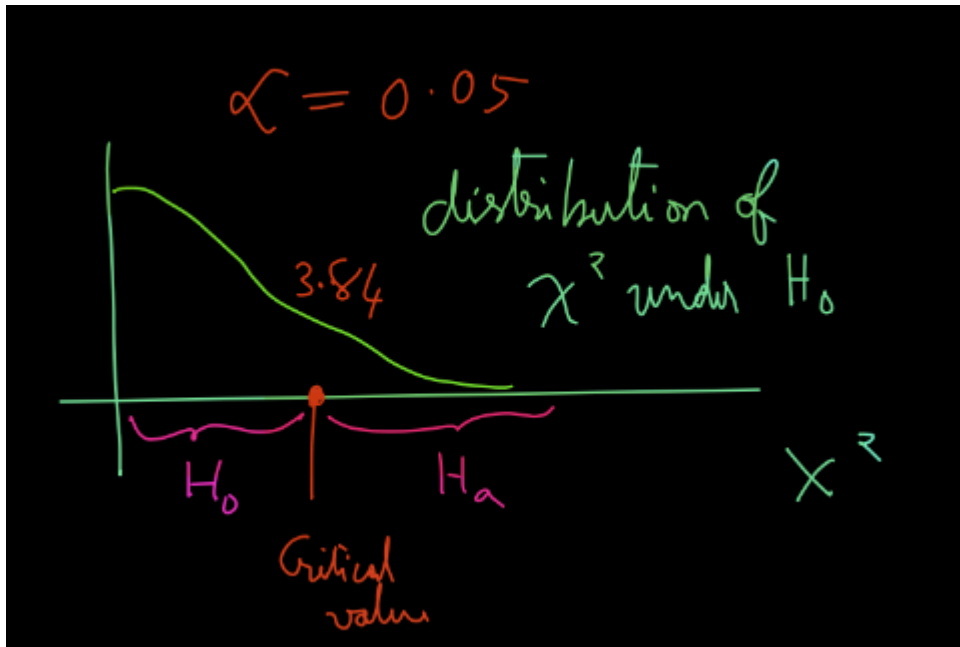
In this context, when the chi-square value is exceptionally small, it generally indicates a **fair coin**.

Conversely, if the chi-square value becomes exceptionally large, it suggests a **biased coin**.

Now, we draw the threshold using the critical value or the alpha value.

This threshold is established using either the critical value or the alpha value.

- It helps us draw a line between what we consider statistically significant (indicating a biased coin) and what we consider within the range of randomness (indicating a fair coin).
- This threshold aids us in making confident judgments about the fairness or bias of the coin based on the observed chi-square values.



This is known as the **Chi-Statistic** that yields **Chi-square Distribution**

And this modified framework is called **Chi square Test**.

How does dof plays a role in the chi-squared testing to accept/reject H_0 ?

Critical Values:

- As degrees of freedom increase, the chi-squared distribution changes shape.
- Higher degrees of freedom lead to higher critical values as per the chi-squared distribution tables, requiring a larger test statistic to reject the null hypothesis at a given significance level (α).

P-Value:

- With increasing degrees of freedom, the distribution becomes less skewed, resulting in smaller tail areas.
- Smaller tail areas mean smaller p-values for the same observed test statistic.
- More degrees of freedom lead to a higher likelihood of obtaining a smaller p-value for the same data.

Intuitive Explanation:

- Degrees of freedom represent flexibility in your analysis.
- More degrees of freedom mean you need a larger difference between observed and expected values to confidently reject the null hypothesis.

- Critical values rise, and p-values decrease with more degrees of freedom.
- Fewer degrees of freedom make it easier to reject the null hypothesis with smaller differences.

✓ **Example 1(Coin Toss):**

```
import numpy as np
import pandas as pd
import seaborn as sns
import matplotlib.pyplot as plt

from scipy.stats import chisquare # Statistical test (chistat, pvalue)
from scipy.stats import chi2
```

Question:

To assess whether a coin is fair or not, we need to compare the expected outcomes with the observed outcomes of tossing the coin.

The expected outcome for a fair coin toss is 50% heads and 50% tails, which corresponds to 25 heads and 25 tails in 50 tosses.

The observed outcomes from 50 coin tosses are 28 heads and 22 tails.

To determine whether the coin is fair, we perform a chi-square test to check if the observed results significantly deviate from the expected results. If the deviation is statistically significant, it may indicate that the coin is not fair.

STEP 1:

What should be the null and alternate hypothesis?

- H_0 : The Coin is Fair.
- H_1 : The Coin is Biased.

STEP 2:

What is the distribution?

- Chi-square distribution.

✓ **STEP 3:**

We perform chi-square test and calculate the P-Value

```
chi_stat, p_value = chisquare(
    [28, 22], # Observed or actual
    [25, 25], # Expected
)
print("p_value:", p_value)
print("chi_stat:", chi_stat)

p_value: 0.3961439091520741
chi_stat: 0.72
```

✓ **STEP 4:**

We defined $\alpha = 0.05$ for confidence level 95%

```
alpha = 0.05

if p_value < alpha:
    print("Reject H0")
    print("Coin is biased")
else:
    print("Fail to reject H0")
    print("Coin is fair")

Fail to reject H0
Coin is fair
```

Solving the same using the formula.

```
#Using the formula
(28 - 25)**2/25 + (22 - 25)**2 / 25 # chi2stat

0.72
```

```
1 - chi2.cdf(0.72, df=1)

0.3961439091520741
```

```
chi2.ppf(0.95, df=1) # If the chi-squared value is greater than 3.84 we reject th

3.841458820694124
```

✓ **Example 2(Coin Toss):**

Question:

To assess whether a coin is fair or not, we need to compare the expected outcomes with the observed outcomes of tossing the coin.

The expected outcome for a fair coin toss is 50% heads and 50% tails, which corresponds to 25 heads and 25 tails in 50 tosses.

The observed outcomes from 50 coin tosses are 45 heads and 5 tails.

To determine whether the coin is fair, we perform a chi-square test to check if the observed results significantly deviate from the expected results. If the deviation is statistically significant, it may indicate that the coin is not fair.

STEP 1:

What should be the null and alternate hypothesis?

- H_0 : The Coin is Fair.
- H_1 : The Coin is Biased.

STEP 2:

What is the distribution?

- Chi-square distribution.

✓ STEP 3:

We perform chi-square test and calculate the P-Value

```
chi_stat, p_value = chisquare(
    [45, 5], # Observed or actual
    [25, 25], # Expected
)
print("p_value:", p_value)
print("chi_stat:", chi_stat)

p_value: 1.5417257900280013e-08
chi_stat: 32.0
```

✓ STEP 4:

We defined $\alpha = 0.05$ for confidence level 95%

```
alpha = 0.05
```

```
if p_value < alpha:
    print("Reject H0")
    print("Coin is biased")
else:
    print("Fail to reject H0")
    print("Coin is fair")

    Reject H0
    Coin is biased
```

Solving the same using the formula.

```
(45 - 25)**2/25 + (5 - 25)**2 /25 # chi2stat

32.0
```

```
1 - chi2.cdf(32, df=1)

1.5417257914762672e-08
```

```
1 - chi2.cdf(3.84, df=1)

0.05004352124870515
```

✓ Chi-square Test for Independence

The Chi-squared test for independence is used when you have **two categorical variables**, and you want to see if they are **related or independent** of each other.

Let's look into it.

Imagine you are running a **Marketing Campaign** where we will ask our existing customers which method they choose to purchase offline or online.

- So, we want to run the campaign to increase the number of online purchases.

Let's say the Marketing Strategist makes a claim:

- That marketing campaign should focus on women
- Females have a higher chance of purchasing online than males

So, the assumption is that there is a dependency between purchases and the gender of the customer.

- This is another hypothetical scenario

- We want to **test and evaluate this assumption**

We want to find out whether there is really or NOT a Dependency between these two variables

- Gender and Purchases

How do we check our assumptions and validate them?

- We need to analyze Purchase preferences (Online or Offline) for both Genders (Male and Female)
- To decide whether or NOT we should focus our Marketing Campaign only on 1 gender
- We need a way to test the dependency of one variable on another
- Before we start the campaign

This is where **Test of Independence** comes into the picture

- Test whether our assumptions/beliefs about that dependency between Gender and Purchase preferences are even True or NOT
- This becomes our motivation to discuss the **Test of Independence**

✓ Gender Vs Offline and Online

Testing independence with `chi2_contingency`

Now, let us say we are conducting a survey on whether gender impacts offline and online purchases.

In the survey, we got the following data:

Gender impacts offline/online purchase				
Observed value				
	M	W		
Offline	527	72	599	66%
Online	206	102	308	34%
	733	174	907	

Expected value		
	M	W
Off	484.	115.
On	249.	59.

H_0 : Gender and preference are independent

H_1 : Gender and preference are not independent

In this scenario, under the assumption of null hypothesis do we have the expected value?

Firstly, we observe that 66% of the respondents prefer offline and 34% of the respondents prefer online.

- Now, if gender has no impact then among 733 men, how many are expected to prefer offline? => 66% of 734 = 484
- Now, if gender has no impact then among 172 women, how many are expected to prefer offline? => 66% of 174 = 115
- Now, if gender has no impact then among 733 men, how many are expected to prefer online? => 34% of 734 = 249
- Now, if gender has no impact then among 172 women, how many are expected to prefer online? => 34% of 174 = 59

All the expected values are calculated using the observed values.

Therefore, we can calculate the chi-squared value as:

$$\chi^2 = \frac{(527-484)^2}{484} + \frac{(72-115)^2}{115} + \frac{(206-249)^2}{249} + \frac{(102-59)^2}{59}$$

```
from scipy.stats import chi2_contingency
```

STEP 1:

What should be the null and alternate hypothesis?

- H_0 : Gender and preference are independent
- H_1 : Gender and preference are not independent

STEP 2:

What is the distribution?

- Chi-square distribution.

✓ STEP 3:

We perform chi-square test and calculate the P-Value

```

observed = [
    [527, 72],
    [206, 102],]

chi_stat, p_value, df, exp_freq = chi2_contingency(observed) # chi_stat, p_value,
print("chi_stat:",chi_stat)
print("p_value:",p_value)
print("df:",df)
print("exp_freq:",exp_freq)

chi_stat: 57.04098674049609
p_value: 4.268230756875865e-14
df: 1
exp_freq: [[484.08710033 114.91289967]
 [248.91289967  59.08710033]]

```

✓ STEP 4:

We defined $\alpha = 0.05$ for confidence level 95%

```

alpha = 0.05

if p_value < alpha:
    print("Reject H0")
    print("Gender and preference are not independent")
else:
    print("Fail to reject H0")
    print("Gender and preference are independent")

Reject H0
Gender and preference are not independent

```

The difference between this example and the above coin toss example.

- In coin toss -> we just fit the expected distribution
- Preference vs Gender -> Here we are testing for independence

✓ Aerofit Example

```

import pandas as pd

!wget --no-check-certificate https://drive.google.com/uc?id=12muE0rUvEtKAPVhKr4rS
--2024-01-18 09:35:19-- https://drive.google.com/uc?id=12muE0rUvEtKAPVhKr4rS
Resolving drive.google.com (drive.google.com)... 142.250.97.102, 142.250.97.1
Connecting to drive.google.com (drive.google.com)|142.250.97.102|:443... conn
HTTP request sent, awaiting response... 303 See Other
Location: https://drive.usercontent.google.com/download?id=12muE0rUvEtKAPVhKr4rS
--2024-01-18 09:35:19-- https://drive.usercontent.google.com/download?id=12muE0rUvEtKAPVhKr4rS
Resolving drive.usercontent.google.com (drive.usercontent.google.com)... 74.1

```

```
16/02/2024, 15:45                                05.ipynb - Colaboratory
Connecting to drive.usercontent.google.com (drive.usercontent.google.com)|74.
HTTP request sent, awaiting response... 200 OK
Length: 7461 (7.3K) [application/octet-stream]
Saving to: 'aerofit.csv'

aerofit.csv      100%[=====>]    7.29K  --.-KB/s    in 0s

2024-01-18 09:35:19 (63.5 MB/s) - 'aerofit.csv' saved [7461/7461]
```

```
df_aerofit = pd.read_csv('/content/aerofit.csv')

df_aerofit.head()
```

	Product	Age	Gender	Education	MaritalStatus	Usage	Fitness	Income	Miles
0	KP281	18	Male	14	Single	3	4	29562	112
1	KP281	19	Male	15	Single	2	3	31836	75
2	KP281	19	Female	14	Partnered	4	3	30699	66
3	KP281	19	Male	12	Single	3	3	32973	85
4	KP281	20	Male	13	Partnered	4	2	35247	47

▼ **STEP 1:**

What should be the null and alternate hypothesis?

- H_0 : Gender does not impact the buying of product
- H_1 : Gender impacts the buying of product

```
gender_product = pd.crosstab(index=df_aerofit['Gender'],columns=df_aerofit['Product'])
gender_product # This will give the count of each gender for each product
```

Product	KP281	KP481	KP781
Gender			
Female	40	29	7
Male	40	31	33

STEP 2:

What is the distribution?

- Chi-square distribution.

▼ STEP 3:

We perform chi-square test and calculate the P-Value

```
chi_stat, p_value, df, exp_freq = chi2_contingency(gender_product) # chi_stat, p_

print("chi_stat:",chi_stat)
print("p_value:",p_value)
print("df:",df)
print("exp_freq:",exp_freq)

chi_stat: 12.923836032388664
p_value: 0.0015617972833158714
df: 2
exp_freq: [[33.77777778 25.33333333 16.88888889]
 [46.22222222 34.66666667 23.11111111]]
```

▼ STEP 4:

We defined $\alpha = 0.05$ for confidence level 95%

```
alpha = 0.05

if p_value < alpha:
    print("Reject H0")
    print("Gender impacts product")
else:
    print("Fail to reject H0")
    print("Gender does not impact product")

Reject H0
Gender impacts product
```

Let's look at another example

▼ STEP 1:

What should be the null and alternate hypothesis?

- H_0 : Gender does not impact the buying of product
- H_1 : Gender impacts the buying of product

```
gender_product = pd.crosstab(index=df_aerofit['Gender'],columns=df_aerofit['Produ
gender_product # This will give the count of each gender for each product
```

Product	KP281	KP481	KP781
Gender			
Female	40	29	7

If KP781 data is not there

STEP 2:

What is the distribution?

- Chi-square distribution.

STEP 3:

We perform chi-square test and calculate the P-Value

```
chi_stat, p_value, df, exp_freq = chi2_contingency([[40, 29], [40, 31]])
```

```
print("chi_stat:",chi_stat)
print("p_value:",p_value)
print("df:",df)
print("exp_freq:",exp_freq)
```

```
chi_stat: 0.0005953595971967067
p_value: 0.9805335549105975
df: 1
exp_freq: [[39.42857143 29.57142857]
 [40.57142857 30.42857143]]
```

STEP 4:

We defined $\alpha = 0.05$ for confidence level 95%

```
alpha = 0.05
```

```
if p_value < alpha:
    print("Reject H0")
    print("Gender impacts product")
else:
    print("Fail to reject H0")
    print("Gender does not impact product")
```

```
Fail to reject H0
Gender does not impact product
```