In [1]:
```
# ignore warnings
import warnings
warnings.filterwarnings("ignore")
```

In [2]:
```
# import libries
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt
```

Topics Covered

1. Chi Square Distribution
2. Chi Square Tests
3. Goodness of Fit Test
4. Test for Independence

1. Chi Square Distribution

- The Chi-Square distribution, also written as χ2 distribution, is a continuous

probability distribution that is widely used in statistical hypothesis testing, particularly in the context of goodness-of-fit tests and tests for independence in contingency tabl

- It arises when the sum of the squares of independent

standard normal random variables follows this distribution

- The Chi-Square distribution has a single parameter, the degrees of freedom (df),

which influences the shape and spread of the distribution

- The degrees o

freedom are typically associated with the number of independent variables or constraints in a statistical probl

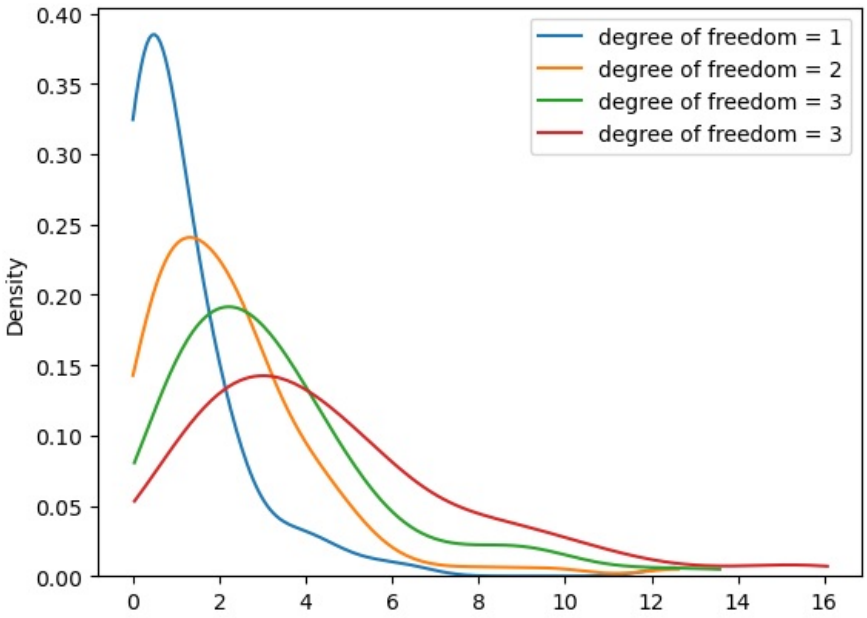
em..es.

In [3]:
```
# Generate  samples of 100 numbers from a standard normal distribution
sample1 = np.random.normal(loc=0, scale=1, size=100)
sample2 = np.random.normal(loc=0, scale=1, size=100)
sample3 = np.random.normal(loc=0, scale=1, size=100)
sample4 = np.random.normal(loc=0, scale=1, size=100)
```

In [4]:
```
x = sample1**2 # square of starndard normal variable (df=1)
y = sample1**2 + sample2**2 # sum of the squares of standard normal variable (df=2)
z = sample1**2 + sample2**2 + sample3**2 # sum of the squares of standard normal variable (df=3)
u = sample1**2 + sample2**2 + sample3**2 + sample4**2 # sum of the squares of standard normal variable (df=4)
```

In [5]:
```
sns.kdeplot(x, clip=(x.min(),x.max()),label='degree of freedom = 1')
sns.kdeplot(y, clip=(y.min(),y.max()),label='degree of freedom = 2')
sns.kdeplot(z, clip=(z.min(),z.max()),label='degree of freedom = 3')
sns.kdeplot(u, clip=(u.min(),u.max()),label='degree of freedom = 3')
plt.legend()
```

Out[5]:
```
<matplotlib.legend.Legend at 0x251c012fe10>
```



Some key properties of the Chi-Square distribution are:

- A. It is a continuous distribution, defined for non-negative values.

It is positively skewed, with the degree of skewness decreasing as the degrees of freedom increase.

- B.The mean of the Chi-Square distribution is equal to its degrees of

freedom, and its variance is equal to twice the degrees of freedom.

- C. As the degrees of freedom increase, the Chi-Square distribution

approaches the normal distribution in shape.

- D.The Chi-Square distribution is used in various statistical tests, such as the Chi-

Square goodness-of-fit test, which evaluates whether an observed frequency distribution fits an expected theoretical distribution, and the Chi-Square test for independence, which checks the association between categorical variables in a contingency table.

In [ ]:

2. Chi Square Tests

- The Chi-Square test is a statistical hypothesis test used to determine if there is a significant

association between categorical variables or if an observed distribution of categorical data differs from an expected theoretical distributio

- It is based on the Chi-Square ($\chi 2$) distribution,

and it is commonly applied in two main scenari

- A. Chi-Square Goodness-of-Fit Test: This test is used to determine if the observed

distribution of a single categorical variable matches an expected theoretical distribution. It is often applied to check if the data follows a specific probability distribution, such as the uniform or binomial distributi

- B. Chi-Square Test for Independence (Chi-Square Test for Association): This test is used to

determine whether there is a significant association between two categorical variables in a sample.on. o a sample.n.

In [ ]:

3. Chi-Square Goodness of Fit Test

The Chi-Square Goodness-of-Fit test is a statistical hypothesis test used to determine if the observed distribution of a single categorical variable matches an expected theoretical distribution. It helps to evaluate whether the data follows a specific probability distribution, such as uniform, binomial, or Poisson distribution, among others. This test is particularly useful when you want to assess if the sample data is consistent with an assumed distribution or if there are significant deviations from the expected pattern.

Assumptions

1. Independence: The observations in the sample must be independent

of each other. This means that the outcome of one observation should not influence the outcome of another observatio

2. Categorical data: The variable being analysed must be categorical,

not continuous or ordinal. The data should be divided into mutually exclusive and exhaustive categorie

3. Expected frequency: Each category should have an expected

frequency of at least

4. Fixed distribution: The theoretical distribution being compared to

the observed data should be specified before the test is conducted. It is essential to avoid choosing a distribution based on the observed data, as doing so can lead to biased results.5.s.n.

The Chi-Square Goodness-of-Fit test involves the following steps:

1. Define the null hypothesis (H0) and the alternative hypothesis (H1):

   H0: The observed data follows the expected theoretical distribution.

   H1: The observed data does not follow the expected theoretical distribution.

2. Calculate the expected frequencies for each category based on the

theoretical distribution and the sample size. 3. Compute the Chi-Square test statistic ($\chi 2$) by comparing the observed and expected frequencies. The test statistic is calculated as

4. Determine the degrees of freedom (df), which is typically the

number of categories minus one (df = k - 1), where k is the number of categorie

5. Calculate the p-value for the test statistic using the Chi-Square

distribution with the calculated degrees of freedom

6. Compare the test statistic to the critical value or the p-value.s.:

Example 1

Suppose we have a six-sided fair die, and we want to test if the die is indeed fair. We roll the die 60 times and record the number of times each side comes up. We'll use the Chi-Square Goodness-of-Fit test to determine if the observed frequencies are consistent with a fair die (i.e., a uniform distribution of the side

Observed frequencies: {'side1':12,'side2':8,'side3':11,'side4':9,'side5':10,'side6':10} s).

Step.1 Define null hypothesis and alternative hypothesis

In [6]:
```
# Observed frequencies for each side of the die (example data)
observed_freq = np.array([10, 12, 8, 15, 9, 6])

# Step.1 Define null hypothesis and alternative hypothesis
H0 = 'The Die Is Fair'       # the observed Data Follow Uniform Distribution
H1 = 'The Die Is Not Fair'   # the observed data not follow uniform distribution

# Step2. Calculate the expected frequencies for each category based on the theoretical distribution and the sam
# Expected frequencies for a fair die (each side has an equal probability of 1/6)
total_rolls = 60
total_categories = 6
probability = 1/6
expected_freq = np.array([probability * total_rolls] * total_categories)
print(observed_freq)

# Step3. Compute the Chi-Square test statistic (χ2) by comparing the observed and expected frequencies.
from scipy.stats import chisquare
chi2_stat, p_value = chisquare(f_obs=observed_freq, f_exp=expected_freq)
print(f"Chi-Square Statistic: {chi2_stat}")


# Step4. Determine the degrees of freedom (df), which is typically the number of categories minus one (df = k -
df = total_categories - 1
print(f"degree of freedom :{df}")

# Step5. Calculate the p-value for the test statistic using the Chi-Square distribution with the calculated deg
import scipy.stats as stats

test_statistic = chi2_stat  # Chi-Square test statistic
degrees_of_freedom = df  # Degrees of freedom

p_value = stats.chi2.sf(test_statistic, degrees_of_freedom)
print("P-value:", p_value)

# Step6.  Interpret the result
alpha = 0.05
print('alpha :',alpha)

flag = 1
if p_value<alpha:
    print('Reject The Null Hypothesis')

else:
    print('Can Not The Null Hypothesis')
    flag = 0

if flag==1:
    print('Result :',H1)
else:
    print('Result :',H0)
```

```
[10 12  8 15  9  6]
Chi-Square Statistic: 5.0
degree of freedom :5
P-value: 0.4158801869955079
alpha : 0.05
Can Not The Null Hypothesis
Result : The Die Is Fair
```

Example 2

Suppose a marketing team at a retail company wants to understand the distribution of visits to their website by day of the week. They have a hypothesis that visits are uniformly distributed across all days of the week, meaning they expect an equal number of visits on each day. They collected data on website visits for four weeks and want to test if the observed distribution matches the expected uniform distribut Monday: 420 • Tuesday: 380 • Wednesday: 410 • Thursday: 400 • Friday: 410 • Saturday: 430 • Sunday: 390ion.

In [7]:
```
# Observed frequencies (number of website visits per day of the week for four weeks):
observed_freq1 = np.array([420,380,410,400,410,430,390])

# Step.1 Define null hypothesis and alternative hypothesis
H0 = 'number of visit on website are equal on each day'       # the observed Data Follow Uniform Distribution
H1 = 'number of visit on website are not equal each day'   # the observed data not follow uniform distribution

# Step2. Calculate the expected frequencies
# since equal number of visit on each day
total_visit = np.sum(observed_freq1)
total_categories = observed_freq1.shape[0]
expected_freq1 = np.array([total_visit/total_categories]*total_categories)
print(expected_freq1)
```

```
# Step3. Compute the Chi-Square test statistic (χ2)
from scipy.stats import chisquare
chi2_stat, p_value = chisquare(f_obs=observed_freq1, f_exp=expected_freq1)
print(f"Chi-Square Statistic: {chi2_stat}")

# Step4. Determine the degrees of freedom (df)
df = total_categories - 1
print(f"degree of freedom :{df}")

# Step5. Calculate the p-value
import scipy.stats as stats

test_statistic = chi2_stat  # Chi-Square test statistic
degrees_of_freedom = df  # Degrees of freedom

p_value = stats.chi2.sf(test_statistic, degrees_of_freedom)
print("P-value:", p_value)

# Step6.  Interpret the result
alpha = 0.05
print('alpha :',alpha)

flag = 1
if p_value<alpha:
    print('Reject The Null Hypothesis')

else:
    print('Can Not Reject The Null Hypothesis')
    flag = 0

if flag==1:
    print('Result :',H1)
else:
    print('Result :',H0)
```

```
[405.71428571 405.71428571 405.71428571 405.71428571 405.71428571
 405.71428571 405.71428571]
Chi-Square Statistic: 4.366197183098591
degree of freedom :6
P-value: 0.6272491491065606
alpha : 0.05
Can Not Reject The Null Hypothesis
Result : number of visit on website are equal on each day
```

In [ ]:

Chi Square Test On Titanic Dataset

In [8]:
```
# import Dataset
import pandas as pd
import numpy as np
from scipy.stats import chisquare

# Load the Titanic dataset
url = "https://raw.githubusercontent.com/datasciencedojo/datasets/master/titanic.csv"
data = pd.read_csv(url)

data.head()
```

Out[8]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

In [9]:
```
# Step.1 Define null hypothesis and alternative hypothesis
H0 = 'Each Class Has equal number of Passengers'       # The distribution of passengers among the classes is un
H1 = 'Each Class Has not equal number of Passengers'   # The distribution of passengers among the classes is no


# Count passengers in each class (observed_freq)
class_counts = data['Pclass'].value_counts().sort_index()
observed_freq3 = class_counts.to_list()
```

```
# Step2. Calculate the expected frequencies
# since each class has eqaul number of passenger
total_categories = len(observed_freq3)
total_passengers = len(data)
expected_counts = total_passengers / total_categories
expected_freq3 = [expected_counts] * total_categories

# Step3. Compute the Chi-Square test statistic (χ2)
from scipy.stats import chisquare
chi2_stat, p_value = chisquare(f_obs=observed_freq3, f_exp=expected_freq3)
print(f"Chi-Square Statistic: {chi2_stat}")

# Step4. Determine the degrees of freedom (df)
df = total_categories - 1
print(f"degree of freedom :{df}")

# Step5. Calculate the p-value
import scipy.stats as stats

test_statistic = chi2_stat  # Chi-Square test statistic
degrees_of_freedom = total_categories - 1  # Degrees of freedom

p_value = stats.chi2.sf(test_statistic, degrees_of_freedom)
print("P-value:", p_value)

# Step6.  Interpret the result
alpha = 0.05
print('alpha :',alpha)

flag = 1
if p_value<alpha:
    print('Reject The Null Hypothesis')

else:
    print('Can Not Reject The Null Hypothesis')
    flag = 0

if flag==1:
    print('Result :',H1)
else:
    print('Result :',H0)
```

```
Chi-Square Statistic: 191.8047138047138
degree of freedom :2
P-value: 2.2394202231028854e-42
alpha : 0.05
Reject The Null Hypothesis
Result : Each Class Has not equal number of Passengers
```

In [ ]:

4. Test For Independence

The Chi-Square test for independence, also known as the Chi-Square test for association, is a statistical test used to determine whether there is a significant association between two categorical variables in a sample. It helps to identify if the occurrence of one variable is dependent on the occurrence of the other variable, or if they are independent of each oth

The test is based on comparing the observed frequencies in a contingency table (a table that displays the frequency distribution of the variables) with the frequencies that would be expected under the assumption of independence between the two variables.er.

Steps

- 1.State the null hypothesis (H0) and alternative hypothesis (H1):
  -

H0: There is no association between the two categorical variables (they are
independent) *
○ H1: There is an association between the two categorical variables (they are depend

- 2.Create a contingency table with the observed frequencies for each combination of the

categories of the two variables

- 3.Calculate the expected frequencies for each cell in the contingency table assuming that

the null hypothesis is true (i.e., the variables are independent)

- 4.Compute the Chi-Square test statistic
- 5.Determine the degrees of freedom: df = (number of rows - 1) * (number of columns - 1)
- 6.Obtain the critical value or p-value using the Chi-Square distribution table or a statistical

software/calculator with the given degrees of freedom and significance level (commonly α = 0.05

- 7.Compare the test statistic to the critical value or the p-value to the significance level to

decide whether to reject or fail to reject the null hypothesis. If the test statistic is greater than the critical value, or if the p-value is less than the significance level, we reject the null hypothesis and conclude that there is a significant association between the two variables.)...ent).

Assumptions

1. Independence of observations: The observations in the sample should be independent of

each other. This means that the occurrence of one observation should not affect the occurrence of another observation. In practice, this usually implies that the data should be collected using a simple random sampling method.

2. Categorical variables: Both variables being tested must be categorical, either ordinal or

nominal. The Chi-Square test for independence is not appropriate for continuous variables.

3. Adequate sample size: The sample size should be large enough to ensure that the

expected frequency for each cell in the contingency table is sufficient. A common rule of thumb is that the expected frequency for each cell should be at least 5. If some cells have expected frequencies less than 5, the test may not be valid, and other methods like Fisher's exact test may be more appropriate.

4. Fixed marginal totals: The marginal totals (the row and column sums of the contingency

table) should be fixed before the data is collected. This is because the Chi-Square test for independence assesses the association between the two variables under the assumption that the marginal totals are fixed and not influenced by the relationship between the variables.

We will use the Chi-Square test for independence to see if the survival rate of passengers is independent of the passenger class.

In [10]:
```
# Data
data.head()
```

Out[10]:

	PassengerId	Survived	Pclass	Name	Sex	Age	SibSp	Parch	Ticket	Fare	Cabin	Embarked
0	1	0	3	Braund, Mr. Owen Harris	male	22.0	1	0	A/5 21171	7.2500	NaN	S
1	2	1	1	Cumings, Mrs. John Bradley (Florence Briggs Th...	female	38.0	1	0	PC 17599	71.2833	C85	C
2	3	1	3	Heikkinen, Miss. Laina	female	26.0	0	0	STON/O2. 3101282	7.9250	NaN	S
3	4	1	1	Futrelle, Mrs. Jacques Heath (Lily May Peel)	female	35.0	1	0	113803	53.1000	C123	S
4	5	0	3	Allen, Mr. William Henry	male	35.0	0	0	373450	8.0500	NaN	S

Set Hypothesis

In [11]:
```
H0 = 'There Is No Association Between Survival and Pclass' # There is no association between the two categorica
H1 = 'There Is  Association Between Survival and Pclass' # There is  association between the two categorical va
```

Create a contingency table with the observed frequencies

In [21]:
```
contingency_table = pd.crosstab(columns=data["Pclass"],index=data['Survived'])
contingency_table
```

Out[21]:

Pclass	1	2	3
Survived			
0	80	97	372
1	136	87	119

Calculate the expected frequencies

In [23]:
```
from scipy.stats import chi2_contingency
chi2,p_value,dof,expected = chi2_contingency(contingency_table)
print('Expected Frequencies : \n',expected)
```

```
Expected Frequencies :
 [[133.09090909 113.37373737 302.53535354]
 [ 82.90909091  70.62626263 188.46464646]]
```

In [ ]:
```
# Perform the Chi-Square test
chi2, p_value, dof, expected = chi2_contingency(contingency_table)

# Print the results
print("\nChi-Square Statistic: {:.2f}".format(chi2))
print("P-value: {:.4f}".format(p_value))
print("Degrees of Freedom: {}".format(dof))
print("Expected Frequencies: \n{}".format(expected))
```

Calculate T-statistics

In [24]:
```
print('Chi-Square Statistic: \n',chi2)
```

```
Chi-Square Statistic:
 102.88898875696056
```

Determine the degrees of freedom

In [25]:
```
print('Degree of Freedom :\n',dof)
```

```
Degree of Freedom :
 2
```

P_value

In [26]:
```
print('P_value :',p_value)
```

```
P_value : 4.549251711298793e-23
```

Result Of Test

In [27]:
```
alpha = 0.5
flag = 1
if p_value<0.5:
    print('Reject The Null Hypothesis')
else:
    print("Can't Reject The Null Hypothesis")
    flag = 0
if flag==1:
    print('Result : ',H1)
else:
    print('Result : ',H0)
```

```
Reject The Null Hypothesis
Result :  There Is  Association Between Survival and Pclass
```

In [ ]:

In [ ]:

In [ ]: