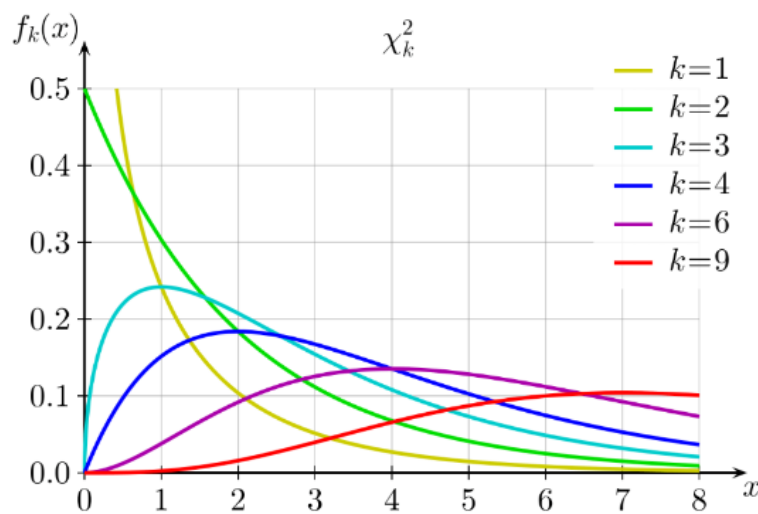


What is Chi-Square Distribution ?

The Chi-Square distribution, also written as χ^2 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 tables. 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 of freedom are typically associated with the number of independent variables or constraints in a statistical problem.

Some key properties of the Chi-Square distribution are:

1. It is a **continuous distribution**, defined for non-negative values.
2. It is **positively skewed**, with the degree of skewness decreasing as the degrees of freedom increase.
3. 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**.
4. **As the degrees of freedom increase**, the Chi-Square distribution approaches the normal distribution in shape.

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.

chi-square



$$\chi^2 = \sum_{i=1}^k Z_i^2 \rightarrow \text{degree of freedom}$$

$\rightarrow df=1$

$$\chi^2 = Z_1^2 + Z_2^2 \rightarrow df=2$$

$$\chi^2 = Z_1^2 + Z_2^2 + Z_3^2 \rightarrow df=3$$

$$\chi^2 = \sum_{i=1}^k Z_i^2 \quad \boxed{df=k}$$

$df \uparrow$

In [5]:

```
# Pratical

import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# Generate a sample 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)
sample5 = np.random.normal(loc=0, scale=1, size=100)
```

In [6]:

sample1

Out[6]:

```
array([-0.70886873, -0.61518884, -1.82762272, -0.02466896, -0.13970317,
       -0.65562498,  0.78241776, -0.61811061, -0.29254929,  0.19864709,
       -0.06085364,  1.68998882,  0.38019819,  1.15502404,  1.17440723,
        3.14242201,  0.35347626, -0.64925827, -1.08396822,  1.07132816,
       -0.73235931, -0.51349668, -2.13765619, -0.85204048, -0.79725994,
        0.28366404,  0.07670644,  0.18312004,  2.15258492, -2.22700436,
       -0.47191283,  0.74058807, -0.33389935,  0.04568418,  1.92693737,
        0.87359758, -0.14347806, -0.1426297 , -0.16660434, -1.05975011,
       -0.84174539,  2.1873914 , -0.55398832,  0.45891942,  1.53377717,
       -0.90459531,  1.40160351, -0.43007715, -0.76826726,  0.83009722,
       -0.26771675, -0.89126552, -0.16024367, -0.47255563,  0.96826852,
       -1.28736164,  0.46597113,  0.69937767, -1.6077451 ,  1.21813043,
       -0.83894196, -0.73985296,  0.52828529,  0.40400941,  0.33339018,
       -1.25530346,  0.87452445, -1.22594934,  0.01371629, -1.08454 ,
        0.8894766 ,  0.95523974, -2.0852497 ,  0.39471483, -0.26241187,
       -0.27463945, -0.9767013 ,  0.18944353, -1.38123716, -0.78430688,
       -0.07489508,  0.64214936, -0.14462964,  0.1362247 ,  1.64580603,
        3.58055032, -0.52016034,  0.66788309, -0.31745217,  0.47465947,
       -0.98294005,  0.01206155, -0.88338844,  0.50284482,  0.25761511,
        0.36316632, -0.95687219,  1.28071872,  1.67716119, -1.72695337])
```

In [7]:

```
# Square the samples

x = sample1**2
y = sample1**2 + sample2**2
z = sample1**2 + sample2**2 + sample3**2
u = sample1**2 + sample2**2 + sample3**2 + sample4**2
v = sample1**2 + sample2**2 + sample3**2 + sample4**2 + sample5**2
```

In [8]:

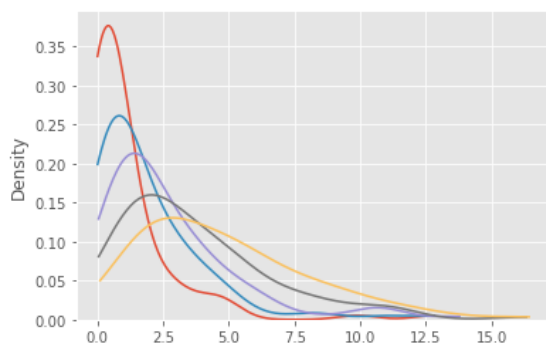
plt.style.use('ggplot')

In [9]:

```
# Plot KDE (Kernel Density Estimation) for each variable
sns.kdeplot(x, clip=(x.min(), x.max()))
sns.kdeplot(y, clip=(y.min(), y.max()))
sns.kdeplot(z, clip=(z.min(), z.max()))
sns.kdeplot(u, clip=(u.min(), u.max()))
sns.kdeplot(v, clip=(v.min(), v.max()))
```

Out[9]:

<AxesSubplot:ylabel='Density'>



Chi Square Test

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 distribution.

It is based on the Chi-Square (χ^2) distribution, and it is commonly applied in two main scenarios:

1. **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 distribution.
2. **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.

1. 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.

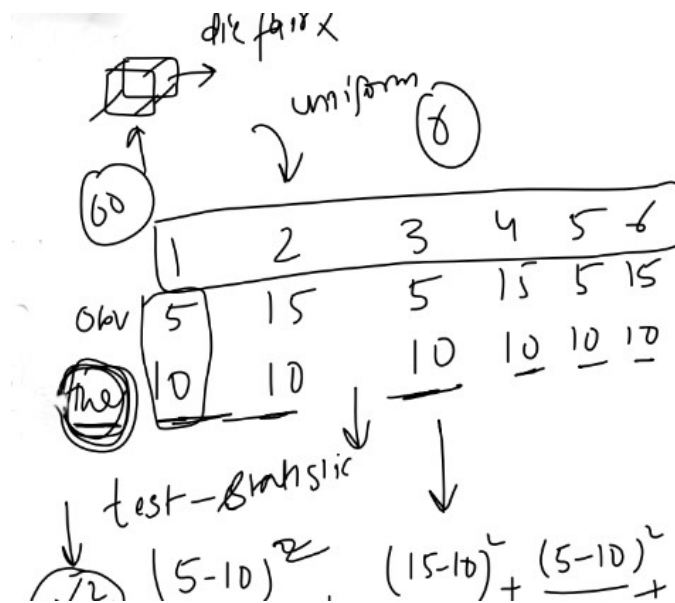
Steps:

- The Chi-Square Goodness-of-Fit test involves the following steps: Define the null hypothesis (H_0) and the alternative hypothesis (H_1):
- **H_0 : The observed data follows the expected theoretical distribution.**
- **H_1 : The observed data does not follow the expected theoretical distribution.**
- Calculate the expected frequencies for each category based on the theoretical distribution and the sample size.
- Compute the Chi-Square test statistic (χ^2) by comparing the observed and expected frequencies. The test statistic is calculated as:

$$\chi^2 = \sum \frac{(\text{Observed} - \text{Expected})^2}{\text{Expected}} = \sum \frac{(O - E)^2}{E}, df = k - 1$$

- where O_i is the observed frequency in category i , E_i is the expected frequency in category i , and the summation is taken over all categories.
- Determine the degrees of freedom (df), which is typically the number of categories minus one ($df = k - 1$), where k is the number of categories.
- Calculate the p-value for the test statistic using the Chi-Square distribution with the calculated degrees of freedom.
- Compare the test statistic to the critical value or the p-value

Explanation:



Assumptions

- 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 observation.
- Categorical data:** The variable being analysed must be categorical, not continuous or ordinal. The data should be divided into mutually exclusive and exhaustive categories.
- Expected frequency:** Each category should have an expected frequency of at least 5. This guideline helps ensure that the Chi-Square distribution is a reasonable approximation for the distribution of the test statistic. Having small expected frequencies can lead to an inaccurate estimation of the Chi-Square distribution, potentially increasing the likelihood of a Type I error (incorrectly rejecting the null hypothesis) or a Type II error (incorrectly failing to reject the null hypothesis).
- 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.

Q: Is the chi-square test a parametric or nonparametric test?

A: The Chi-Square Goodness-of-Fit test is a **non-parametric test**. Non-parametric tests do not assume that the data comes from a specific probability distribution or make any assumptions about population parameters like the mean or standard deviation.

- In the Chi-Square Goodness-of-Fit test, we compare the observed frequencies of the categorical data to the expected frequencies based on a hypothesized distribution. The test doesn't rely on any assumptions about the underlying distribution's parameters. Instead, it focuses on comparing observed counts to expected counts, making it a non-parametric test.

Examples:

- 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 sides).

Observed frequencies:

- Side 1: 12 times
- Side 2: 8 times
- Side 3: 11 times
- Side 4: 9 times
- Side 5: 10 times
- Side 6: 10 times

Solution:

To test if the observed frequencies of rolling a fair six-sided die are consistent with a uniform distribution, we can perform a chi-square goodness-of-fit test.

Let's define the null hypothesis (H_0) and alternative hypothesis (H_a):

- H_0 : The die is fair and the observed frequencies follow a uniform distribution.**

- **H_a: The die is not fair and the observed frequencies do not follow a uniform distribution.**

We have the observed frequencies for each side of the die: Side 1: 12 times Side 2: 8 times Side 3: 11 times Side 4: 9 times Side 5: 10 times Side 6: 10 times

First, we need to calculate the expected frequencies under the assumption of a fair die. Since we rolled the die 60 times, each side would be expected to come up $1/6$ of the time ($60 / 6 = 10$) in a fair scenario.

Next, we can calculate the chi-square test statistic:

$$\chi^2 = \sum ((\text{Observed} - \text{Expected})^2 / \text{Expected})$$

Calculating the chi-square test statistic:

$$**\chi^2 = [(12-10)^2/10] + [(8-10)^2/10] + [(11-10)^2/10] + [(9-10)^2/10] + [(10-10)^2/10] + [(10-10)^2/10]**$$

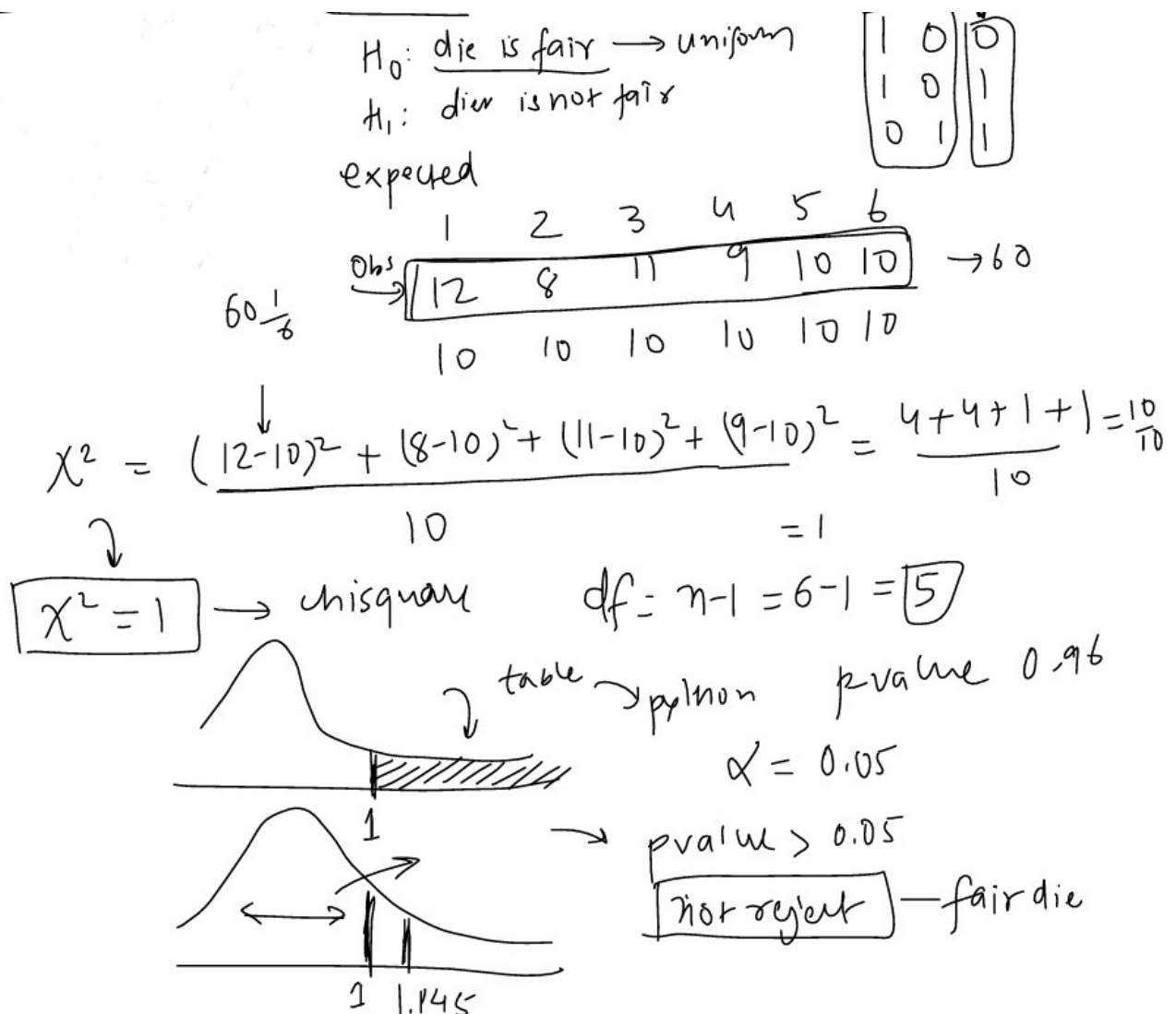
After performing the calculations, we obtain a chi-square test statistic value.

To determine if the observed frequencies are consistent with a fair die, we compare this chi-square test statistic value to the critical chi-square value at a given significance level (e.g., $\alpha = 0.05$) with degrees of freedom equal to the number of categories minus 1 (in this case, $6 - 1 = 5$).

If the calculated chi-square test statistic is greater than the critical chi-square value, we reject the null hypothesis and conclude that the die is not fair.

Conversely, if the calculated chi-square test statistic is less than or equal to the critical chi-square value, we fail to reject the null hypothesis and conclude that the observed frequencies are consistent with a fair die.

Performing the calculations will provide the final result



In [18]:

```
import scipy.stats as stats
```

```
test_statistic = 1 # Chi-Square test statistic from the previous example
degrees_of_freedom = 5 # Degrees of freedom from the previous example
```

```
# Calculate the p-value using the chi2 survival function (sf)
p_value = stats.chi2.sf(test_statistic, degrees_of_freedom)
```

```
print("P-value:", p_value)
```

P-value: 0.9625657732472964

2. A survey of 800 families in a village with 4 children each revealed the following distribution:

```
girls 4 3 2 1 0
boys  0 1 2 3 4
families 32 178 290 236 64
```

Is this data consistent with the result that male and female births are equally probable?

survey → village → 800 families
 ↓
 4 children
 binomial

girls 4 3 2 1 0
 # boys 0 1 2 3 4
 # families 32 178 290 236 64

$p(s) = p(d) = \frac{1}{2}$
 $H_0: p(m) = p(f) = \frac{1}{2}$
 $H_a: p(m) \neq p(f)$
 $p, q = 1 - p$
 $n = 4, p = \frac{1}{2}$

	0	1	2	3	4
Obs	32	178	290	236	64
Th _{exp}	50	200	300	200	50

$$\chi^2 = \frac{(32-50)^2}{50} + \frac{(178-200)^2}{200} + \frac{(290-300)^2}{300} + \frac{(236-200)^2}{200} + \frac{(64-50)^2}{50}$$

$$= \frac{324}{50} + \frac{484}{200} + \frac{100}{300} + \frac{1296}{200} + \frac{196}{50}$$

$$= 6.2 + 2.3 + 0.33 + 6.2 + 3.9$$

$$\chi^2 = 18.93$$

$$df = 5 - 1 = 4$$

$$0.00081 < \alpha (0.05)$$

reject the Null hypothesis

Binomial distribution:
 $P(x) = {}^nC_x p^x (1-p)^{n-x}$
 $P(0) = {}^4C_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4 = \frac{1}{16} \times 800 = 50$
 $P(1) = {}^4C_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^3 = \frac{4!}{1!3!} \times \frac{1}{16} = \frac{4}{16} \times 800 = 200$
 $P(2) = {}^4C_2 \left(\frac{1}{2}\right)^2 \left(\frac{1}{2}\right)^2 = \frac{4!}{2!2!} \times \frac{1}{16} = \frac{6}{16} \times 800 = 300$
 $P(3) = {}^4C_3 \left(\frac{1}{2}\right)^3 \left(\frac{1}{2}\right)^1 = \frac{4!}{3!1!} \times \frac{1}{16} = \frac{4}{16} \times 800 = 200$
 $P(4) = {}^4C_4 \left(\frac{1}{2}\right)^4 \left(\frac{1}{2}\right)^0 = \frac{1}{16} \times 800 = 50$

Solution:

To determine if the data is consistent with the assumption that male and female births are equally probable, we can perform a chi-square goodness-of-fit test.

First, let's define the null hypothesis (H0) and alternative hypothesis (Ha):

- H0: Male and female births are equally probable.
- Ha: Male and female births are not equally probable.

We can calculate the expected frequencies under the assumption of equal probability for male and female births. Since each family has 4 children, the expected frequency for each gender can be calculated as

$$(800 \text{ families} / 2) * (1/2) = 200.$$

Now, we can perform the chi-square test using the observed and expected frequencies. The chi-square test statistic can be calculated as:

$$\chi^2 = \sum ((\text{Observed} - \text{Expected})^2 / \text{Expected})$$

Using the provided data, the observed and expected frequencies are as follows:

Girls: 4 3 2 1 0
Boys: 0 1 2 3 4
Families: 32 178 290 236 64
Expected: 200 200 200 200 200

Calculating the chi-square test statistic:

$$\chi^2 = [(32-200)^2/200] + [(178-200)^2/200] + [(290-200)^2/200] + [(236-200)^2/200] + [(64-200)^2/200]$$

After performing the calculations, we obtain a chi-square test statistic value. To assess whether the data supports or rejects the null hypothesis, we compare this value to the critical chi-square value at a given significance level (e.g., $\alpha = 0.05$) with degrees of freedom equal to the number of categories minus 1.

If the calculated chi-square test statistic is greater than the critical chi-square value, we reject the null hypothesis, indicating that the data is not consistent with the assumption of equal probability for male and female births. On the other hand, if the calculated chi-square test statistic is less than or equal to the critical chi-square value, we fail to reject the null hypothesis, suggesting that the data is consistent with the assumption of equal probability for male and female births.

In [19]:

```
import scipy.stats as stats

test_statistic = 18.93 # Chi-Square test statistic from the previous example
degrees_of_freedom = 4 # Degrees of freedom from the previous example

# Calculate the p-value using the chi2 survival function (sf)
p_value = stats.chi2.sf(test_statistic, degrees_of_freedom)

print("P-value:", p_value)
```

P-value: 0.0008112261867904042

Python Case Study

Question: What are the observed frequencies of passengers in each class, and how can the expected counts be calculated assuming a uniform distribution for the chi-square goodness-of-fit test??

- we can simply do value_counts on data so why we do this test?
- here we have sample data not population data
- we do Hypothesis testing on sample data to predict Population

In [20]:

```
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[20]:

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 [21]:

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

Out[21]:

```
1    216
2    184
3    491
Name: Pclass, dtype: int64
```

In [22]:

```
# Calculate the expected counts assuming a uniform distribution
total_passengers = len(data)
expected_counts = total_passengers / 3
expected = [expected_counts] * 3
print("Expected Passenger Counts (assuming uniform distribution):\n", expected)
```

```
Expected Passenger Counts (assuming uniform distribution):
[297.0, 297.0, 297.0]
```

In [23]:

```
# Perform the Chi-Square Goodness of Fit test
chi2, p_value = chisquare(class_counts, expected)

# Print the results
print("\nChi-Square Statistic: {:.2f}".format(chi2))
print("P-value: {:.4f}".format(p_value))
```

```
Chi-Square Statistic: 191.80
P-value: 0.0000
```

In [25]:

```
# Interpret the results
alpha = 0.05
if p_value < alpha:
    print("\nWe reject the null hypothesis. The distribution of passengers among the classes is not uniform.")
else:
    print("\nWe fail to reject the null hypothesis. The distribution of passengers among the classes is uniform.")
```

We reject the null hypothesis. The distribution of passengers among the classes is not uniform.

2. 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 other.

- 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.

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 dependent).**
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:

$$\chi^2 = \sum \frac{(O - E)^2}{E}$$

Where:

χ^2 = Chi Square obtained
 \sum = the sum of
 O = observed score
 E = expected score

- Determine the degrees of freedom: $df = (\text{number of rows} - 1) * (\text{number of columns} - 1)$
- 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 $\alpha = 0.05$).
- 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.

Assumptions

- 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.
- 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.
- 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.
- 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.

Examples

- A researcher wants to investigate if there is an association between the level of education (categorical variable) and the preference for a particular type of exercise (categorical variable) among a group of 150 individuals. The researcher collects data and creates the following contingency table

Education	Yoga	Running	Swimming	Total
High School	15	20	10	45
Bachelor's	20	35	15	65
Master's or PhD	5	15	20	40
Total	40	40	65	150

Solution

To determine if there is an association between the level of education and the preference for a particular type of exercise, we can perform a chi-square test of independence.

The null hypothesis (**H0**) assumes that there is no association between the variables,

while the alternative hypothesis (**Ha**) suggests that there is an association.

To calculate the expected counts for the chi-square test, we assume that there is no association, and the variables are independent. We can calculate the expected counts by multiplying the row total and column total for each cell and dividing it by the grand total.

Expected count = (row total * column total) / grand total

Using the provided contingency table, we can calculate the expected counts:

Expected count for High School and Yoga = $(45 * 40) / 150 = 12$

Expected count for High School and Running = $(45 * 40) / 150 = 12$

Expected count for High School and Swimming = $(45 * 65) / 150 = 19.5$

Expected count for Bachelor's and Yoga = $(65 * 40) / 150 = 17.33$

Expected count for Bachelor's and Running = $(65 * 40) / 150 = 17.33$

Expected count for Bachelor's and Swimming = $(65 * 65) / 150 = 28.33$

Expected count for Master's or PhD and Yoga = $(40 * 40) / 150 = 10.67$

Expected count for Master's or PhD and Running = $(40 * 40) / 150 = 10.67$

Expected count for Master's or PhD and Swimming = $(40 * 65) / 150 = 17.33$

Now, with the observed and expected counts, we can perform the chi-square test of independence to determine if there is a significant association between education level and exercise preference.

Observed ✓

Education	Yoga	Running	Swimming	Total
High School	15	20	10	45
Bachelor's	20	30	15	65
Master's or PhD	5	15	20	40
Total	40	65	45	150

edu ↔ exercise (independent)
 H_0 : they are independent X
 H_1 : they are associated
 need a confgen

expected

	Yoga	Run	swim
High	12	19	13.5
Bach	17	28	20
PhD	10	17	12

$$\frac{45 \times 40}{150} \quad \frac{65 \times 65}{150} \times \frac{45 \times 45}{150}$$

$$\frac{(15-12)^2}{12} + \frac{(20-19)^2}{20} + \frac{(10-13.5)^2}{10}$$

p-value 0.04 (α)
 reject null $df = (3-1)(3-1) = 4$
 $\chi^2 = 9.95$

In [30]:

```
import scipy.stats as stats

test_statistic = 9.95 # Chi-Square test statistic from the previous example
degrees_of_freedom = 4 # Degrees of freedom from the previous example

# Calculate the p-value using the chi2 survival function (sf)
p_value = stats.chi2.sf(test_statistic, degrees_of_freedom)

print("P-value:", p_value)
```

P-value: 0.04127840066324082

Python case study

Q: Does the survival rate of passengers depend on the passenger class? We will utilize the Chi-Square test for independence to investigate the association between the survival rate and the passenger class.

In [32]:

```
# We will use the Chi-Square test for independence to see if the survival rate of passengers is independent of the passenger class
import pandas as pd
import numpy as np
from scipy.stats import chi2_contingency

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

data.head()
```

Out[32]:

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 [33]:

```
# Calculate the contingency table
contingency_table = pd.crosstab(data['Survived'], data['Pclass'])

contingency_table
```

Out[33]:

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

In [34]:

```
# 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))
```

Chi-Square Statistic: 102.89
P-value: 0.0000
Degrees of Freedom: 2
Expected Frequencies:
[[133.09090909 113.37373737 302.53535354]
 [82.90909091 70.62626263 188.46464646]]

In [35]:

```
# Interpret the results
alpha = 0.05
if p_value < alpha:
    print("\nWe reject the null hypothesis. There is a significant association between passenger class and survival rate.")
else:
    print("\nWe fail to reject the null hypothesis. There is no significant association between passenger class and survival rate.")
```

We reject the null hypothesis. There is a significant association between passenger class and survival rate.

Applications in Machine Learning

- 1. **Feature selection:** Chi-Square test can be used as a filter-based feature selection method to rank and select the most relevant categorical features in a dataset. By measuring the association between each categorical feature and the target variable, you can eliminate irrelevant or redundant features, which can help improve the performance and efficiency of machine learning models.

2. **Evaluation of classification models:** For multi-class classification problems, the Chi-Square test can be used to compare the observed and expected class frequencies in the confusion matrix. This can help assess the goodness of fit of the classification model, indicating how well the model's predictions align with the actual class distributions.
3. **Analysing relationships between categorical features:** In exploratory data analysis, the Chi- Square test for independence can be applied to identify relationships between pairs of categorical features. Understanding these relationships can help inform feature engineering and provide insights into the underlying structure of the data.
4. **Discretization of continuous variables:** When converting continuous variables into categorical variables (binning), the Chi-Square test can be used to determine the optimal number of bins or intervals that best represent the relationship between the continuous variable and the target variable.
5. **Variable selection in decision trees:** Some decision tree algorithms, such as the CHAID (Chi- squared Automatic Interaction Detection) algorithm, use the Chi-Square test to determine the most significant splitting variables at each node in the tree. This helps construct more effective and interpretable decision trees.

In [38]:

```
from IPython.core.display import display, HTML
display(HTML("<style>.container { width:100% !important; }</style>"))
```

In []: