**hi 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 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:**

*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.*

*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.*

*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.

**Mathematical formulation of the Chi-Square distribution**

The mathematical formulation of the Chi-Square distribution in terms of standard normal random variables can be represented as follows:

Let Z1, Z2, Z3, …, Zn be n independent standard normal random variables. The Chi-Square random variable (χ²) with k degrees of freedom can be defined as the sum of the squares of these independent standard normal random variables:

χ² = Z1² + Z2² + Z3² + … + Zk²

where k is the number of degrees of freedom.

In other words, the Chi-Square distribution with k degrees of freedom is obtained by summing the squares of k independent standard normal random variables.

This formulation establishes the relationship between the Chi-Square distribution and the sum of squares of independent standard normal variables, which is the basis for many statistical tests and applications of the Chi-Square distribution in hypothesis testing and contingency table analysis.

import numpy as np  
import seaborn as sns  
  
# 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)  
  
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  
  
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()))

**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 (χ2) distribution,and it is commonly applied in two main scenarios:

**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 distribution.

**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.

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

*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.*

*Calculate the expected frequencies for each category based on the  
theoretical distribution and the sample size.  
  
Compute the Chi-Square test statistic (χ2) by comparing the observed  
and expected frequencies. The test statistic is calculated as:*

*χ² = Σ((O — E)² / E)*

*Where:*

*Σ represents the summation across all categories.*

*O is the observed frequency for each category.*

*E is the expected frequency for each category, which is determined based on the assumed theoretical distribution and the sample size.*

*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*

import scipy.stats as stats  
#if test statics and df is known  
  
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)

**Chi Square test on real datasets(goodness of fit):-**

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()  
  
# Count passengers in each class  
class\_counts = data['Pclass'].value\_counts().sort\_index()  
class\_counts  
  
# 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)  
  
# 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))  
  
# 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.")

**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 expectedfrequency of at least 5. This guideline helps ensure that the Chi-Square distribution is a reasonable approximation for thedistribution of the test statistic. Having small expected frequencies can lead to an inaccurate estimation of the Chi-Square distribution.*

***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.*

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.

Question:-  
  
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 are consistent with a fair die (i.e., a uniform distribution of the sides), we can perform a Chi-Square Goodness-of-Fit test. Here's how to solve this problem:  
  
Step 1: Define the null hypothesis (H0) and the alternative hypothesis (H1):  
H0: The observed frequencies follow a fair die (i.e., a uniform distribution).  
H1: The observed frequencies do not follow a fair die.  
  
Step 2: Set the significance level (α) for the test. Let's assume α = 0.05.  
  
Step 3: Calculate the expected frequency for each side assuming a fair die. Since we have 6 sides and rolled the die 60 times, the expected frequency for each side would be 60/6 = 10.  
  
Expected frequencies:  
- Side 1: 10 times  
- Side 2: 10 times  
- Side 3: 10 times  
- Side 4: 10 times  
- Side 5: 10 times  
- Side 6: 10 times  
  
Step 4: Compute the Chi-Square test statistic (χ^2) by comparing the observed and expected frequencies using the formula:  
  
χ^2 = Σ((O - E)^2 / E)  
  
where Σ represents the summation across all sides, O is the observed frequency, and E is the expected frequency.  
  
Using the given observed and expected frequencies, we can calculate the Chi-Square test statistic as follows:  
  
χ^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)  
  
Step 5: Calculate the degrees of freedom (df) for the test. In this case, df = number of categories - 1 = 6 - 1 = 5.  
  
Step 6: Determine the critical value from the Chi-Square distribution table or using statistical software at the given significance level (α) and degrees of freedom (df). For α = 0.05 and df = 5, the critical value is approximately 11.07.  
  
Step 7: Compare the calculated test statistic (χ^2) with the critical value. If χ^2 > critical value, reject the null hypothesis; otherwise, fail to reject the null hypothesis.  
  
In this case, if the calculated test statistic χ^2 is greater than 11.07, we would reject the null hypothesis and conclude that the observed frequencies are not consistent with a fair die. However, if χ^2 is less than or equal to 11.07, we would fail to reject the null hypothesis and conclude that the observed frequencies are consistent with a fair die.  
  
You can calculate the χ^2 value and compare it with the critical value to determine the result of the test.

**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**

*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).*

*Create a contingency table with the observed frequencies for each combination of the  
categories of the two variables.  
  
Calculate the expected frequencies for each cell in the contingency table assuming that the null hypothesis is true (i.e., the variables are independent).  
  
Compute the Chi-Square test statistic:  
χ2 = Σ [(O\_ij — E\_ij)2 / E\_ij]  
where O\_ij is the observed frequency in each cell and E\_ij is the expected frequency.*

*Determine the degrees of freedom: df = (number of rows — 1) \* (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 (commonlyα = 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 ofeach 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.

**Application in ML:-**

**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.

**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.

**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.

**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.

**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