Chi-Square Test (χ^2-test):

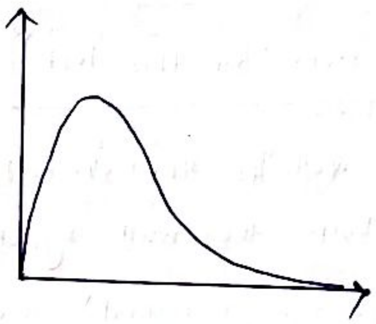
Before going into the deep of this test, let the chi-square distribution be explained. This test is based on χ^2-distribution.

Chi-square Distribution:

Chi-square distribution arises when we do the sum of squares of independent standard normal random variables. and chi-square distribution are getting the shape and spread as Normal distribution when we are going to increase the number of independent standard normal random variables.

Also chi-square distribution is a special case of gamma distribution. It is a continuous distribution —

having pdf as — $f_X(x) = \begin{cases} \dfrac{(1/2)^{n/2}}{\Gamma(n/2)} e^{-x/2} x^{n/2-1} & 0 < x < \infty \\ 0 & \text{otherwise} \end{cases}$



- Chi-square distribution has only one paremeter, that is degree of freedom, which is equals to number of independent standard normal random variables.

- Mean $= n$ (degree of freedom) and Variance $= 2n$

- This chi-square test is used in various statistical test, one of them is chi-square test.

Let us explain about chi-square test briefly —

⑪ chi-square test is one of the non-parametric statistical test by which we can find out the dependency of two categorical columns or we can ~~state~~ draw conclusion about the equality of theoritical distribution and practical distribution / observed distribution with respect to any categorical column.

With based on those things, chi-square Test can be classified into two types —

(a) Goodness of Fit (b) Test for Independence.

Ⓐ <u>Goodness of fit :</u>

This type of chi-square test helps us to find out the relation between observed distribution and theoretical distribution. for any categorical column.

In one word, it helps us to find out that the theoretical distribution and observed distribution, both are same or not.

⑪ How can it be conducted (steps for this test)

① Create the null hypothesis as well as the alternative hypothesis.

⑪ For this test by default the null hypothesis is that both distribution (observed and theoretical) are same.

Alternative will be that they are not same, they are different.

② Find out the expected value corresponding the categories in that column. (according to the theoretical distribution).

③ Find out the χ^2- statistics with help of the following formula —

$$\chi^2 = \sum \frac{(\text{observed Value} - \text{Expected Value})^2}{\text{Expected Value}}$$

Means, we need to find out (observed Value − Expected Value) / Expected Value for each and every categories in that categorical column. Then their sum would be our χ^2-statistic.

(4) Based on the ~~P-Value~~ χ^2-statistic and degree of freedom we easily find out the p-Value. (also can be done by the method of Region of rejection).

(5) Compare the Value with significance level (α). Based on this we can reach our target.

Numerical 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 sides).

Observed frequecies:

(1) Side 1 : 12 times	(2) Side 2: 8 times	(3) 11 times
(4) Side 4 : 9 times	(5) Side 5: 10 times	(6) Side 6 : 10 times.

Answer:

⇒ Here our Null Hypothesis (H_0) : Observed frequencies are consistent with a fair die.

Alternative hypothesis (H_a) ; Observed frequencies are consistent with a baised die.

Now we need to find out the expected value according to our theoretical distribution (Here uniform distribution).

So, the expected frequencies should be $(1/6 \times 60) = 10$ times for each side.

(ii) Now we have to find out the χ^2-statistics —

$$\chi^2 = \frac{(12-10)^2}{10} + \frac{(8-10)^2}{10} + \frac{(11-10)^2}{10} + \frac{(9-10)^2}{10} + 2\times\frac{(10-10)^2}{10}$$

$$= 4/10 + \frac{4}{10} + \frac{1}{10} + \frac{1}{10}$$

$$= 1$$

So, we have χ^2 statistic $= 1$

degree of freedom $= (6-1) = 5$

level of significance $= 0.05$

Now, from the table we got the value as 11.070 with respect to ~~χ^2 statistic~~ and degree of freedom (5).
level of significance
(0.05)

and we have —

$$1 < 11.070 .$$

So, we cannot reject our Null Hypothesis.

(Note that we didn't use the P-Value).

(iii) For finding P-Value we can use the statistical tools in any environment (python, excel etc)

So, P-Value corresponding to χ^2-statistic and $df = 5$ is 0.962.

So, P-Value > level of significance $(\alpha) = 0.05$

So, we can't reject our Null Hypothesis.

So, For this example, observed frequencies are consistant with a fair die. (Proved)

(~~All though the strength of evidence is not so strong~~).

Example -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
Family	32	178	290	~~8~~236	64

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

(i) Null-hypothesis : (H_0) : Equally probable

Alternative hypothesis (H_a): Not equally probable.

(ii) We have to find the expected value of families —

so, our data is following the binomial distribution. And according to theoretical assumption $p\,(\text{male}) = p\,(\text{female}) = 1/2$.

so, for. girl = 4 , boys = 0.

the number of families would be $= \left\{ {}^4c_0 \left(\frac{1}{2}\right)^0 \left(\frac{1}{2}\right)^4 \right\} \times 800$

$= 50$

For, girl = 3 boys = 1

$= \left\{ {}^4c_1 \left(\frac{1}{2}\right)^1 \left(\frac{1}{2}\right)^3 \right\} \times 800$

$= 200$

For girls = 2 boys = 2 $= \left\{ {}^4c_2 \left(\frac{1}{2}\right)^4 \right\} \times 800 = 300$

For girls = 1 boys = 3 $= \left\{ {}^4c_3 \left(\frac{1}{2}\right)^4 \right\} \times 800 = 200$

For girls = 0 boys = 4 $= 50$ (similarly)

(iii) Now we have to find our χ^2- statistic —

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

$$= 6.48 + 2.42 + 0.33 + 6.48 + 3.92$$

$$= 19.63.$$

and we have degree of freedom $= 4$.

Now, With respect to χ^2-statistic and df, we got 9.488 with level of significance $= 0.05$

Now, as, χ^2-statistics $(19.63) > 9.488$

So, We reject the Null hypothesis.

- With respect to p-Value it is also shown that we should reject our null hypothesis —

as we got p-Value as ~~0.047~~ 0.00059

which is less than level of significance $(\alpha) = 0.05$.

So, our alternative hypothesis is true.

So, Male and Female birth are not equally probable. (Proved)

- Test of Independence :

With this type of chi-square distribution, we find out the dependency relationship between two columns.

In other word, it tests if any two ~~categorical~~ columns are independent or not.

- Steps for this test :

1. Create the null hypothesis as well as alternative hypothesis.
2. Create the contingency table with the observed frequences for each combination of the categories of two variables.

(3) Calculate the expected frequencies of each cell in the contingency table assuming that the null hypothesis is true.

(4) Compute the χ^2- statistic and degree of freedom.

(5) Based on that we draw conclusion about our hypothesis.

- Two condition to be proved our null hypothesis —

  (A) (statistic value) < (level of significance's area)

  (B) P value > level of significance

  We can use any of these.

- Example :

A ~~rech~~ researcher wants to investigate, if there is an association between the level of education (categorical variable) and the preferance for a particular type of excercise (categorical variable) among a group of 150 individuals. The researcher collects data and create the following contingency table.

	Exercise type			
Education	Yoga	Running	swimming	Total
High school	15	20	10	45
B.Sc	20	30	15	65
M.sc or Phd	5	15	20	40
Total	40	65	45	150

- Null hypothesis (Ho): There is no association between Education and the preferance for a particular type of excercise.

Alternative hypothesis (Ha): There is .....

(i) Now I have to construct a contingency table based on that the Null hypothesis is true.

Now, the first cell should be the number of student, whose education level is high school, choose to do yoga.

Now, as the education level and Exercise type are independent so, the probability of such people should be —

(probability (of education level = high school) × probability (yoga)

$$= \left(\frac{45}{130} \times \frac{40}{150} \right)$$

Now, the frequency of such student would be —

$$\left(\frac{45}{150} \times \frac{40}{150} \times 150 \right) = 12$$

Similarly we can find out the expected frequencies for all the cell. and the cell looks like —

Education	Yoga	Running	swimming	Total
High school	12	20	13	45
B.sc	17	28	20	65
Msc. or Phd	11	17	12	40
Total	40	65	45	150

(Exercise type)

(ii) We have to find out χ^2-statistic —

$$\chi^2\text{-statistic} = \frac{(15-12)^2}{12} + \frac{(20-20)^2}{20} + \frac{(10-13)^2}{13} + \frac{(20-17)^2}{17} + \frac{(30-28)^2}{28}$$

$$+ \frac{(15-20)^2}{20} + \frac{(5-11)^2}{11} + \frac{(15-17)^2}{17} + \frac{(20-12)^2}{12}$$

$$= 0.75 + 0.69 + 0.53 + 0.14 + 1.25 + 3.27$$

$$+ 0.23 + 5.33$$

$$= 12.19$$

(ii) Degree of freedom = $\left(\text{categories}\Big|_{\text{Education}} - 1\right)\left(\text{categories}\Big|_{\text{Exercise}} - 1\right)$

$= (3-1)(3-1) = 4.$

(iii) The value with respect to level of significance (0.05) and degree of freedom (4) is — 9.488.

clearly it is shown that χ^2-statistic $\nless$ 9.488

so, we reject the null hypothesis.

(iii) If we calculate the p-value with respect to χ^2-statistic and df we got — 0.0159. Which is less than α (level of significance).

so, we reject our null hypothesis.

so, there is association between Education ~~and~~ level and preferance of excercise types. (Proved)

CHI-SQURE TEST :

Visualization of the chi-square distribution with independent standard normal variables :

In [5]:

```
import numpy as np
```

In [10]:

```
sample1=np.random.normal(0,1,100)
sample2=np.random.normal(0,1,100)
sample3=np.random.normal(0,1,100)
sample4=np.random.normal(0,1,100)
sample5=np.random.normal(0,1,100)
sample6=np.random.normal(0,1,100)
```

In [24]:

```
legand=["for x","for y","for z","for p","for q","for r"]
```

In [23]:

```
x=sample1**2
y=x+sample2**2
z=y+sample3**2
p=z+sample4**2
q=p+sample5**2
r=q+sample6**2
```

In [20]:

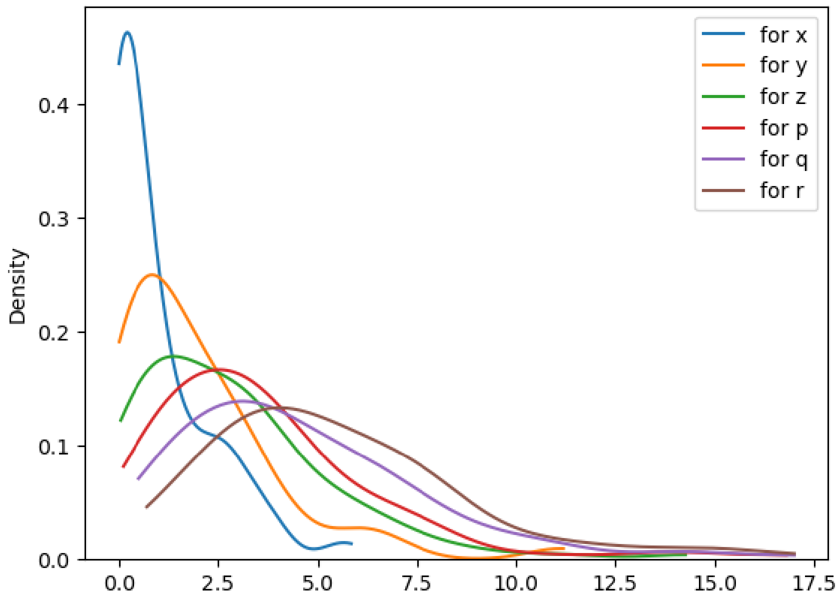
```
import seaborn as sns
import matplotlib.pyplot as plt
```

In [27]:

```
arr=[x,y,z,p,q,r]
for i in arr:
    sns.kdeplot(i,clip=(i.min(),i.max()))
plt.legend(legand)
```

Out[27]:

```
<matplotlib.legend.Legend at 0x1d2c4b40dc0>
```



How to find the p-value with respect to chi-sqr statistic and degree of freedom :

In [28]:

```
import scipy.stats as stat
chi_stat=12.19
df=4
p_value=stat.chi2.sf(chi_stat,df)
print(f"p-value : {p_value}")
```

```
p-value : 0.015992911448370114
```

Case study of chi-sqr test:

In [29]:

```
import pandas as pd
```

In [30]:

```
df=pd.read_csv(r"C:\Users\DELL\Downloads\titanic(train).csv")
```

In [31]:

```
df.head()
```

Out[31]:

	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

Goodness of Fit :

Now we are going to check if the distribution of passengers among the classes is uniform or not in titanic.

so lets construct the null hypothesis and alternative hypothesis.

null hypothesis(Ho):The distribution of passengers among the classes is uniform in titanic.

alternative hypothesis(Ha):The distribution of passengers among the classes is not uniform in titanic.

observed value :

In [42]:

```
obderved_value=df["Pclass"].value_counts().sort_index()
obderved_value
```

Out[42]:

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

now we have to calculate the expected value of three class

Expected value :

as they follow the uniform distribution then the number of passengers in each class should be equal.

expected value for each class is

In [43]:

```
expected_value=[len(df)/3]*3
expected_value
```

Out[43]:

```
[297.0, 297.0, 297.0]
```

In [44]:

```
import scipy.stats as stat
from scipy.stats import chisquare
```

In [46]:

```
chi_2,p_value=chisquare(obderved_value,expected_value)
```

In [47]:

```
print(f"the value of the chi-squre is {chi_2}")
print(f"\n p-value is {p_value}")
```

```
the value of the chi-squre is 191.8047138047138

 p-value is 2.2394202231028854e-42
```

In [50]:

```
alpha=0.05
if p_value>alpha:
    print("we cannot reject the null hypothesis.\nso The distribution of passengers among the classes is uniform in titanic
else:
    print("we reject our null hypothesis.\nso The distribution of passengers among the classes is not uniform in titanic."
```

```
we reject our null hypothesis.
so The distribution of passengers among the classes is not uniform in titanic.
```

Test for independance:

We will use the Chi-Square test for independence to see if the survival rate of passengers is independent of the passenger class or not

null hypothesis(Ho):the survival rate of passengers is independent of the passenger class.

alternative hypothesis(Ha):the survival rate of passengers is not independent of the passenger class.

1. we have to contruct the contingency table based on the observations.

In [52]:

```
contingency_table=pd.crosstab(df["Survived"],df["Pclass"])
contingency_table
```

Out[52]:

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

In [53]:

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

In [54]:

```
chisqr_val,p_value,dof,expected_table=chi2_contingency(contingency_table)
```

In [60]:

```
print("chi-2 statistic :",chisqr_val)
print("P-value :",p_value)
print("expected contigency table :\n",expected_table)
```

```
chi-2 statistic : 102.88898875696056
P-value : 4.549251711298793e-23
expected contigency table :
 [[133.09090909 113.37373737 302.53535354]
 [ 82.90909091  70.62626263 188.46464646]]
```

In [61]:

```
alpha=0.05
if p_value>alpha:
    print("we cannot reject the null hypothesis.\nso the survival rate of passengers is independent of the passenger class
else:
    print("we reject our null hypothesis.\nso the survival rate of passengers is not independent of the passenger class.")
```

```
we reject our null hypothesis.
so the survival rate of passengers is not independent of the passenger class.
```

__THE END________________________________