VIT Vidyalankar Institute of Technology
Accredited A+ by NAAC

Prof. Sampat Mali
Department of Mathematics

χ^2-Distribution

Suppose we are given a die and we want to know whether it biased or unbiased or suppose in a cholera
epidemic, we inoculated a group and we want to know whether inoculation is effective in preventing the attack of cholera.
In such situation chi-square test is used to test the hypothesis the die is based or the inoculation is effective

The test statistic χ^2 is defined by

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(O_i - E_i)^2}{E_i}\right]$$

O_i = Observed frequency,
E_i = expected frequency $\chi\chi$

If the calculated value of χ^2 is greater than the tabulated value, then we conclude that the difference between the observed values and the expected frequencies is significant and the hypothesis is rejected, If on the other hand the calculated value of χ^2 is less than tabulated value, we conclude that the difference between the observed values and the expected frequencies is not significant and the hypothesis is accepted.

Use of χ^2-test:

1) To determine association between two or more attributes

χ^2- Test is used to determine whether there is association between
1) The color of mother's eye and daughter's eye. 2) Between inoculation and prevention of disease. In such cases null hypothesis is,

H_0: There is no association between attributes

Or **H_0: Two given attributes are independent**

2) To test the goodness of fit:

χ^2- Test is used to judge whether a given sample may be regarded as a simple sample from a certain hypothetical population.
χ^2- Test enables us to ascertain how the theoretical distribution such as binomial, Poisson, or normal fit the observed frequencies

H_0: The theory support the observations or fit is good

3) The test the difference between observed frequencies and expected frequencies:

χ^2- Test can also be used to ascertain where the difference between observed frequencies and the expected frequencies is purely due to inadequancy in the theory applied

4) To test equality of several proportions:

χ^2- Test can also be used to test whether the populations p_1, p_2, p_3, p_4 in different populations are equal i.e. χ^2- test can also be used to test hypothesis that $p_1 = p_2 = p_3 = p_4$

5) To test the hypothesis about σ^2 :

χ^2- Test is also used to test the population variance

Condition forχ^2-test

1) The number of observation, N must be sufficiently large i.e. $N \geq 50$
2) **Frequency of every cell must be greater than 10**, if any frequency is less than 10, it combine with neighboring frequency so that the combined frequency is greater than 10 and the degree of freedom reduced accordingly
3) The number of class, n must not be to small nor to large preferable $4 \leq n \leq 16$

Prof. Sampat Mali
Department of Mathematics

Yates correction: In a 2 ×2 table degree of freedom is (2-1) × (2-1) = 1. If any of the cell frequency is less than **5** we have to use pooling method. But this will result in χ^2 with zero degree of freedom. This is meaningless in this case Yates suggested to used

$$\chi^2 = \sum_{i=1}^{n}\left[\frac{(|O_i - E_i| - 0.5)^2}{E_i}\right]$$

Note: Even if Yates correction is not made we would have arrive at the same conclusion

χ^2- Test of independency of attributes

If the population is known to have two attributes. A and B, then A can be divided in to m-categories $A_1, A_2, A_3, A_4, A_5, A_6, A_7, \ldots\ldots, A_m$ and B can be divide into n-categories $B_1, B_2, B_3, B_4, B_5, B_6, B_7 \ldots B_n$ accordingly the members of the populations and hence those of the samples can be divided in to mn classes. In this case the sample data may be presented in the form of a matrix containing m- rows and n-columns and hence mn cell shows the observed frequencies, O_{ij} in the various cells where i=1,2,3,....,m & j=1,2,3,.....,n

O_{ij}-be the observed frequencies possessing the attributes A_i and B_j

A↓\ B→	B_1	B_2	B_3	.....	B_n	Row total
A_1	O_{11}	O_{12}	O_{13}	.....	O_{1n}	O_{1*}
A_2	**$O_{21/E21=}$**	O_{22}	O_{23}	.....	O_{1n}	O_{2*}
A_3	O_{31}	**O_{32}**	O_{33}	.....	O_{3n}	O_{3*}
⋮	⋮	⋮	⋮	⋮	⋮	⋮
A_m	O_{m1}	O_{m2}	O_{m3}	.....	O_{mn}	O_{m*}
Column total	O_{*1}	O_{*2}	O_{*3}	.....	O_{*n}	N

Null Hypothesis, H_0: Two attributes A and B are independent

or There is no association between attributes A and B

We compute the expected frequencies E_{ij} for varies cell using the formula

$E_{ij} = \frac{O_{i*} \times O_{*j}}{N}$, i=1,2,3,.....,m & j=1,2,3,.....,n

We compute

$$\chi^2 = \sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}}\right]$$

The number of degree if freedom for thisχ^2- computed from the m×n contingency table is

ν= (m-1) × (n-1)

If $\chi^2 < \chi^2_{\nu}(\alpha)$, H_0 is accepted at α% LOS, ∴ A & B are independent.

If$\chi^2 > \chi^2_{\nu}(\alpha)$, H_0 is rejected at α% LOS, ∴ A & B are dependent.

Prof. Sampat Mali
Department of Mathematics

Note;

A↓ \ B→	B1	B_2	Total
A_1	a	b	a+ b
A_2	c	d	c+ d
Total	a+ c	b+ d	N=a+ b+ c+ d

$$\chi^2 = \frac{N\times(ad-bc)^2}{(a+b)(a+c)(d+c)(d+b)}$$

In this case Yates correction is

$$\chi^2 = \frac{N\times(|ad-bc|-N/2)^2}{(a+b)(a+c)(d+c)(d+b)}$$

Prof. Sampat Mali
Department of Mathematics

Example on uniformity:-

1) The following table gives the number of accident in district during a week. Test whether the **accident are uniformly distributed over the week**.

Day	Sun	Mon	Tue	Wed	Thu	Fri	Sat
No. of accident	13	12	11	10	14	10	14

Solution:-

H_0: *Accident are uniformally distributed over a week*

Day	No. of Accident (O_i)	Expected No. of accident	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
Sun	13	12	1	1/12
Mon	12	12	0	0
Tue	11	12	1	1/12
Wed	10	12	4	4/12
Thur	14	12	4	4/12
Fri	10	12	4	4/12
Sat	14	12	4	4/12
Total=84				$\chi^2 = \sum_{i=1}^{n}\left[\frac{(O_i-E_i)^2}{E_i}\right] = \frac{18}{12} = 1.5$

$v = degree\ of\ freedom = n - 1 = 7 - 1 = 6$

Thus $\chi^2_{cal} = 1.5$

$\chi^2_v(\alpha) = \chi^2_6(5\%) = 12.592$

Since $\chi^2_{cal} < \chi^2_v(\alpha)$

H_0 is accepted

We can concluded that

Accident are uniformally distributed.

Prof. Sampat Mali
Department of Mathematics

2) A die was thrown 132 times and the following frequencies were obtained:

Number obtained	1	2	3	4	5	6
Frequency	15	20	25	15	29	28

Test whether the die is unbiased.

Solution:-

$H_0: Die\ is\ unbiased$

Number Obtained	Frequency (O_i)	Expected frequency(E_i)	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
1	15	22	49	49/22
2	20	22	4	4/22
3	25	22	9	9/22
4	15	22	49	49/22
5	29	22	49	49/22
6	28	22	36	36/22
Total frequency=132				$\chi^2 = \sum_{i=1}^{n}\left[\frac{(O_i-E_i)^2}{E_i}\right]$=**8.9091**

$v = degree\ of\ freedom = n - 1 = 6 - 1 = 5$

Thus $\chi^2_{cal} = 8.9091$

$\chi^2_v(\alpha) = \chi^2_5(5\%)$=11.07

Since $\chi^2_{cal} < \chi^2_v(\alpha)$

H_0 is accepted

We can concluded that die is unbiased die

Prof. Sampat Mali
Department of Mathematics

3) 300 digits were chosen at random from a table of random numbers. The frequency of digits is as follows:

digit	0	1	2	3	4	5	6	7	8	9
Frequency	28	29	33	31	26	35	32	30	31	25

Using chi-square test examination the hypothesis that the digit were distribution in equal in the table

Solution:-

$H_0: Digits\ are\ uniformely\ distributed$

Digits	Observed frequency(O_i)	Expected frequency(E_i)	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
0	28	30	4	4/30
1	29	30	1	1/30
2	33	30	9	9/30
3	31	30	1	1/30
4	26	30	16	16/30
5	35	30	25	25/30
6	32	30	4	4/30
7	30	30	0	0/30
8	31	30	1	1/30
9	25	30	25	25/30
Total frequency=300				$\chi^2 = \sum_{i=1}^{n}\left[\frac{(O_i - E_i)^2}{E_i}\right]$=2.866667

$v = degree\ of\ freedom = n - 1 = 10 - 1 = 9$

Thus $\chi^2_{cal} = 2.866667$

$\chi^2_v(\alpha) = \chi^2_9(5\%)$=16.919

Since $\chi^2_{cal} < \chi^2_v(\alpha)$

H_0 is accepted

We can concluded that digits are uniformally distributed.

VIT Vidyalankar Institute of Technology
Accredited A+ by NAAC

Prof. Sampat Mali
Department of Mathematics

Fitting of binomial distribution

1) In a study designed to determine patient acceptance of a new pain reliever, 100 physicians each selected a sample of 25 patients, to participate in the study. Each patient, after trying the new pain reliever for a special period of time was asked whether it was preferable to the pain reliever used regularly in the past. The results of the study are as shown below.

Number of patient out of 25,referring new pain reliever	0	1	2	3	4	5	6	7	8	9	10>
Number of doctors reporting the patients	5	6	8	10	10	15	17	10	10	9	0

We are interested in determining whether or not these data are compatible with the hypothesis that they were drawn from a population that follows a binomial distribution. Again we employ a Chi- square test for good of fit.

Solution:-df=n-k, k=No. of constraints

H_0: Fitting of Binomial distribution is good for the given data.

x_i	f_i	$x_i f_i$	$P(X=x)=\boldsymbol{n_{c_x}p^x\times q^{n-x}}$ $=25_{c_x}0.2^x\times 0.8^{25-x}$	$E_i=N\times P(X=x)$ $=100\times P(X=x)$ $=100\times 25_{c_x}0.2^x\times 0.8^{25-x}$
0	5	0	0.003778	0.3778≈ 0
1	6	6	0.023612	2.3612≈ 2
2	8	16	0.070835	7.0835≈7
3	10	30	0.135768	13.5768≈14
4	10	40	0.186681	18.6681≈19
5	15	75	0.196015	19.6015≈20
6	17	102	0.163346	16.3346≈16
7	10	70	0.110842	11.0842≈11
8	10	80	0.062349	6.2349≈6
9	9	81	0.029442	2.9442≈ 3
>10	0	0	0.017332	1.1778≈2
	N=$\sum(f_i)=100$	$\sum(x_i\times f_i)=500$		

Mean=$\frac{1}{N}\times\sum(x_i\times f_i)=\frac{1}{100}\times 500=5$

By given for every doctor there are 25 patients i.e. n=25?

But in Binomial distribution mean = n p

5 = 25×p $\therefore p=\frac{5}{25}=\frac{1}{5}=0.2$,$\therefore q=1-p=1-0.2=0.8$

Prof. Sampat Mali
Department of Mathematics

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
29	23	6	36	1.565217
10	19	-9	81	4.263158
15	20	-5	25	1.25
17	16	1	1	0.0625
10	11	-1	1	0.090909
19	11	8	64	5.818182
				$\chi^2 = \sum_{i=1}^{n}\left[\frac{(O_i-E_i)^2}{E_i}\right]$=13.04997

Thus χ^2_{cal}=13.049966

$v = degree\ of\ freedom = n - k = 6 - 2 = 4$

$\chi^2_v(\alpha)$ = $\chi^2_4(5\%)$ = 9.4877

Since $\chi^2_{cal} > \chi^2_v(\alpha)$

$H_0\ is\ rejected$

Therefore fitting of Binomial distribution is not good for the given data.

i.e. Binomial distribution is not suitable for the given data.

Prof. Sampat Mali
Department of Mathematics

2) The face sheet of patients record maintained in a local health department contains 10 entries. A sample of 100 records revealed the following distribution of erroneous entries

Number of erroneous entries out of 10	0	1	2	3	4	5 or more	Total
Number of records	8	25	32	24	10	01	100

Test the goodness of fit of these data to the binomial distribution with P=.20

Solution:- By given P=.20, q=0.8, n=10

H_0: Fitting of Binomial distribution is good for the given data.

x_i	f_i	$P(X=x)=n_{c_x}p^x \times q^{n-x}$ $=10_{c_x}0.2^x \times 0.8^{10-x}$	$E_i = N \times P(X=x)$ $= 100 \times P(X=x)$ $= 100 \times 10_{c_x}0.2^x \times 0.8^{10-x}$
0	8	0.107374	10.7374≈11
1	25	0.268435	26.8435≈27
2	32	0.301990	30.1990≈30
3	24	0.201327	20.1327≈20
4	10	0.088080	8.8080≈9
5≥ 0	1	0.032794	2.6424≈3
	N=$\sum(f_i) = 100$		

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
33	38	-5	25	0.657895
32	30	2	4	0.133333
24	20	4	16	0.8
11	12	-1	1	0.083333
				$\chi^2 = \sum_{i=1}^{n}\left[\frac{(O_i-E_i)^2}{E_i}\right]$=1.674561

Thus χ^2_{cal}=1.674561

$v = degree\ of\ freedom = n - k = 4 - 1 = 3$

$\chi^2_v(\alpha) = \chi^2_3(5\%) = 7.8147$

Since $\chi^2_{cal} < \chi^2_v(\alpha)$

H_0 *is accepted*

Therefore fitting of Binomial distribution is good for the given data.

i.e. Binomial distribution is suitable for the given data.

VIT Vidyalankar Institute of Technology
Accredited A+ by NAAC

Prof. Sampat Mali
Department of Mathematics

FITTING OF POISSON DISTRIBUTION

1) A hospital administrator wishes to test the null hypothesis that emergency admission follows Poisson distribution with λ=3, Suppose that over a period of 90 days the number of emergency admissions were as shown in the following table

Number of emergency admission in a day	0	1	2	3	4	5	6	7	8	9	10& above	Total
Number of days this number of emergency admission occurred	5	14	15	23	16	9	3	3	1	1	0	90

Test the goodness of these data to the Poisson distribution.
Df=n-k, n is number of observation, k=no. of constraints

Solution: $H_0: Fitting\ of\ Poisson\ distribution\ is\ good\ for\ the\ given\ data$

By given mean of Poisson distribution $\lambda = 3$ Total frequency N = 90

x_i	f_i	$P(X=x)=\frac{e^{-\lambda}\times\lambda^x}{x!}$ $=\frac{e^{-3}\times 3^x}{x!}$	$E_i = N \times P(X=x)$ $= 90 \times P(X=$ $= 90 \times \frac{e^{-3}\times 3^x}{x!}$
0	5	0.049787	4.48083≈5
1	14	0.149361	13.44249≈13
2	15	0.224042	20.16378≈20
3	23	0.224042	20.16378≈20
4	16	0.168031	15.12279≈15
5	9	0.100819	9.07371≈9
6	3	0.050409	4.53681≈5
7	3	0.021604	1.94436≈2
8	1	0.008102	0.72918≈1
9	1	0.002701	0.24309≈0
≥ 10	0	0.001102	0.09918≈0

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
19	18	1	1	0.055556
15	20	-5	25	1.25
23	20	3	9	0.45
16	15	1	1	0.066667
17	17	0	0	0
				$\chi^2 = \sum_{i=1}^{n}\left[\frac{(O_i-E_i)^2}{E_i}\right]$=1.822223

Thus χ^2_{cal}=1.822223

$v = degree\ of\ freedom = n - k = 5 - 1 = 4$

$\chi^2_v(\alpha)$ = $\chi^2_4(5\%)$ = 9.4877

Since $\chi^2_{cal} < \chi^2_v(\alpha)$

$H_0\ is\ accepted$

Therefore fitting of Poisson distribution is good for the given data.

i.e. Poisson distribution is suitable for the given data.

Prof. Sampat Mali
Department of Mathematics

2) The following are the number of a particular organism found in 100 samples of water from a pound

Frequency of organisms per sample	0	1	2	3	4	5	6	7	Total
Frequency	15	30	25	20	5	4	1	0	100

Test the null hypothesis that data were drawn from a Poisson distribution.

Solution: Fitting of Poisson distribution is good for the given data

By given Total frequency N = 100

x_i	f_i	$x_i \times f_i$	$P(X=x)=\frac{e^{-\lambda}\times\lambda^x}{x!}=\frac{e^{-1.86}\times 1.86^x}{x!}$	$E_i = N\times P(X=x) = 100\times P(X=x) = 100\times\frac{e^{-1.86}\times 1.86^x}{x!}$
0	15	0	0.155673	15.5673≈ 15
1	30	30	0.289551	28.9551≈29
2	25	50	0.269283	26.9283≈27
3	20	60	0.166955	16.6955≈17
4	5	20	0.077634	7.7634≈8
5	4	20	0.028880	2.888≈3
6	1	6	0.008953	0.8953≈1
7	0	0	0.002379	0.2379≈0
Total	N=$\sum(f_i)$ =100	$\sum(x_i\times f_i)$ =186		

Mean = $\frac{1}{N}\times\sum(x_i\times f_i)=\frac{1}{100}\times 186 = 1.86$

O_i	E_i	$O_i - E_i$	$(O_i - E_i)^2$	$\frac{(O_i - E_i)^2}{E_i}$
15	15	0	0	0
30	29	1	1	0.034483
25	27	-2	4	0.148148
20	17	3	9	0.529412
10	12	-2	4	0.333333
				$\chi^2=\sum_{i=1}^{n}\left[\frac{(O_i-E_i)^2}{E_i}\right]$=1.0635

Thus χ^2_{cal}=1.045376

$v = degree\ of\ freedom = n - k = 5 - 2 = 3$

$\chi^2_v(\alpha)$ = $\chi^2_3(5\%)$ = 7.8147

Since $\chi^2_{cal} < \chi^2_v(\alpha)$

H_0 *is accepted*

Therefore fitting of Poisson distribution is good for the given data.

i.e. Poisson distribution is suitable for the given data.

Prof. Sampat Mali
Department of Mathematics

Test of independency

1) A sample of 500 college students participated in a study designed to evaluate the level of college student's knowledge of a certain group of common disease. The following tables shows the students classified by major field of study and level of knowledge of the group of disease

Major	Knowledge of disease		Total
	Good	Poor	
Premedical	31	91	122
Other	19	359	378
Total	50	450	500

Do these data suggest that there is a relationship between **knowledge of the group of disease** and **major field of study** the college students from which the present sample was drawn?

Solution:-There is no association between two attributes or two attribites are independent

H_0: There is no association between knowledge of group of disease and major field of study

Major	Knowledge of disease		Total
	Good	Poor	
Premedical	31=a	91=b	122=a+b
Other	19=c	359=d	378=c+d
Total	50=a+c	450=b+d	500=a+b+c+d

∵ every cell frequency is greater than 5

$\therefore \quad \chi^2_{cal} = \frac{N\times(ad-bc)^2}{(a+b)(a+c)(d+c)(d+b)} = \frac{500\times[(31\times359)-(19\times91)]^2}{(122)\times(50)\times(378)\times(450)} = 42.578618$

Degree of freedom $v=(m-1)\times(n-1)=1\times1=1$

$\chi^2_v(\alpha) = \chi^2_1(5\%) = 3.84146$

Since $\chi^2_{cal} > \chi^2_v(\alpha)$

H_0 *is rejected*

Therefore we can conclude that there is some association between knowledge of group of disease and major field of study

Prof. Sampat Mali
Department of Mathematics

2) Investigate the association between the darkness of eye color in father and son from the following data

	Dark	Non Dark
Dark	44	90
Not Dark	80	786

Solution: -

H_0: There is no association between darkness of eye color in father and son.

	Dark	Non Dark	Total
Dark	44=a	90=b	**134=a+b**
Not Dark	80=c	786=d	**866=c+d**
Total	**124=a+c**	**876=b+d**	**1000=N=a+b+c+d**

∵ every cell frequency is greater than 5

$\therefore \quad \chi^2_{cal} = \frac{N\times(ad-bc)^2}{(a+b)(a+c)(d+c)(d+b)} = \frac{1000\times[(44\times786)-(80\times90)]^2}{(134)\times(124)\times(866)\times(876)}$ **=59.490180**

Degree of freedom $v=(m-1)\times(n-1)=1\times1=1$

$\chi^2_v(\alpha) = \chi^2_1(5\%) = 3.84146$

Since $\chi^2_{cal} > \chi^2_v(\alpha)$

H_0 *is rejected*

∴ ***we can conclude that there*** **is some association between darkness of eye color in father and son.**

Homework

3) The following data is collected on two characters. Based on this, can you say that there is no relation between smoking and literacy:

	Smokers	Non-smokers
Literates	83	57
Illiterates	45	68

Prof. Sampat Mali
Department of Mathematics

01/08/2020

3) Two batches each of 12 animals are taken for the test of inoculation, one batch was inoculated and the other batch was not inoculated. The numbers of dead and surviving animals are given in the following table for both cases. Can the inoculation be regarded as effective against the disease?

	Dead	survived	Total
Inoculated	2	10	12
Non-inoculated	8	4	12
Total	10	14	24

Solution: -

H_0: There is no association between inoculation and survival.

OR We can-not say that inoculation be regarded as effective against the disease

	Dead	survived	Total
Inoculated	2=a	10=b	12=a+b
Non-inoculated	8=c	4=d	12=c+d
Total	10=a+c	14=b+d	24=N=a+b+c+d

∵ every cell frequency is not greater than 5

$\therefore \quad \chi^2_{cal} = \frac{N\times(|ad-bc|-N/2)^2}{(a+b)(a+c)(d+c)(d+b)} = \frac{24\times(|(2\times4)-(8\times10)|-12)^2}{(12)(10)(12)(14)} =$ **4.285714**

Degree of freedom $v=(m-1)\times(n-1)=1\times1=1$

$\chi^2_v(\alpha) = \chi^2_1(5\%) = 3.84146$

Since $\chi^2_{cal} > \chi^2_v(\alpha)$

H_0 *is rejected*

We can conclude that there is some association between inoculation and survival.

Prof. Sampat Mali
Department of Mathematics

4) The following information was obtained in a sample of 50 small general shops

	Shops in urban area	Shops in rural area	Total
Owned by men	17	18	35
Owned by women	3	12	15
Total	20	30	50

Can it be said that there are more women owners in rural areas than urban areas

Solution: -

H_0: There is no association between area and gender.

	Shops in urban area	**Shops in rural area**	**Total**
Owned by men	17=a	18=b	**35=a+b=35**
Owned by women	3=c	12=d	**15=c+d=15**
Total	**20=a+c**	**30=b+d**	**50=N=a+b+c+d**

∵ every cell frequency is not greater than 5

$\therefore \quad \boldsymbol{\chi^2}_{cal} = \frac{N\times(|ad-bc|-N/2)^2}{(a+b)(a+c)(d+c)(d+b)} = \frac{50\times(|(17\times12)-(18\times3)|-25)^2}{(35)(20)(15)(30)} = \mathbf{2.480159}$

Degree of freedom $v=(m-1)\times(n-1)=1\times1=1$

$\boldsymbol{\chi^2}_v(\alpha) = \boldsymbol{\chi^2}_1(5\%) = 3.84146$

Since $\boldsymbol{\chi^2}_{cal} < \boldsymbol{\chi^2}_v(\alpha)$

H_0 *is accepted*

We can conclude that there is no association between area and gender.

i.e. We cannot conclude that there are more women owners in rural areas than urban areas.

5) The purpose of a study by Vermund et al. was to investigate the hypothesis that HIV-infected women who are also infected with human papillomavirus (HPV), detected by molecular hybridization are more likely to have cervical cytologic abnormalities than are women with only one or neither virous. The data shown in the following table were reported the investigators. **We wish to know if we may conclude that there is a relationship between HPV status and stage of HIV infection**

HPV ↓ \ HIV→	Seropositive, symptomatic	Seropositive, asymptomatic	Seronegative	Total
Positive	23	04	10	37
Negative	10	14	35	59
Total	33	18	45	96

Solution:

H_0= There is a no relationship between HPV status and stage of HIV infection

Take $\boldsymbol{E_{ij}} = \frac{\boldsymbol{O_{i*}} \times \boldsymbol{O_{*j}}}{\boldsymbol{N}}$

HPV ↓ \ HIV→	**Seropositive, symptomatic**		**Seropositive, asymptomatic**		**Seronegative**		**Total**
Positive	O_{11} =23	$\boldsymbol{E_{11} = 12.72 \approx 13}$	O_{12} =04	$\boldsymbol{E_{12} = 6.94 \approx 7}$	O_{13} = 10	$\boldsymbol{E_{13} = 17.34 \approx 17}$	**37=$\boldsymbol{O_{1*}}$**
Negative	O_{21} =10	$\boldsymbol{E_{21} = 20.28 \approx 20}$	O_{22} =14	$\boldsymbol{E_{22} = 11.06 \approx 11}$	O_{23} = 35	$\boldsymbol{E_{23} = 27.66 \approx 28}$	**59=$\boldsymbol{O_{2*}}$**
Total	**33= $\boldsymbol{O_{*1}}$**		**18=$\boldsymbol{O_{*2}}$**		**45=$\boldsymbol{O_{*3}}$**		**96=N**

$$\boldsymbol{\chi^2_{cal}} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\right] = \frac{(23-13)^2}{13} + \frac{(4-7)^2}{7} + \frac{(10-17)^2}{17} + \frac{(10-20)^2}{20} + \frac{(14-11)^2}{11} + \frac{(35-28)^2}{28} = 19.42856$$

Thus $\boldsymbol{\chi^2_{cal}}$ =**19.42856**

Degree of freedom $v=(m-1)\times(n-1) = 1\times 2 = 2$

$\boldsymbol{\chi^2}_v(\alpha)$ = $\boldsymbol{\chi^2}_2(5\%)$ = 5.991

Since $\boldsymbol{\chi^2_{cal}} > \boldsymbol{\chi^2}_v(\alpha)$

H_0 *is rejected*

We can conclude that there is a relationship between HPV status and stage of HIV infection.

Prof. Sampat Mali
Department of Mathematics

6) The sharing of injection equipment among drug users was investigated by Klee et al. As part of their research they collected the following information regarding use of needle exchange of injection drug users who were located either through treatment agency files or through outreach work designed to involve those not receiving counseling treatment.

Users↓	Use of needle exchange				Total
	Regular	Occasional	never	Not known	
Agency	56	15	20	24	115
Non-agency	19	06	16	53	94
Total	75	21	36	77	209

May we conclude from these data that use of needle exchange and agency status is related?

Solution:

H_0: There is no relation between use of needle exchange and agency status.

Take $\boldsymbol{E_{ij} = \frac{O_{i*} \times O_{*j}}{N}}$

Users↓	**Use of needle exchange**								**Total**
	Regular		**Occasional**		**never**		**Not known**		
Agency	O_{11} =56	$E_{11} =$ **41.27** $\approx$ **41**	O_{12} =15	$E_{12} =$ **11.56** $\approx$ **12**	O_{13} =20	$E_{13} =$ **19.81** $\approx$ **20**	O_{14} =24	$E_{14} =$ **42.37** $\approx$ **42**	**115=O_{1*}**
Non-agency	O_{21} =19	$E_{21} =$ **33.73** $\approx$ **34**	O_{22} =06	$E_{22} =$ **9.44** $\approx$ **9**	O_{23} =16	$E_{23} =$ **16.19** $\approx$ **16**	O_{24} =53	$E_{24} =$ **34.63** $\approx$ **35**	**94=O_{2*}**
Total	**75=O_{*1}**		**21=O_{*2}**		**36=O_{*3}**		**77=O_{*4}**		**209**

$$\boldsymbol{\chi^2_{cal} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{(O_{ij} - E_{ij})^2}{E_{ij}}\right]}$$

$$\boldsymbol{\chi^2_{cal} = \frac{(56-41)^2}{41} + \frac{(15-12)^2}{12} + \frac{(20-20)^2}{20} + \frac{(24-42)^2}{42} + \frac{(19-34)^2}{34} + \frac{(6-9)^2}{9} + \frac{(16-16)^2}{16} + \frac{(53-35)^2}{35} = 30.82688}$$

Thus $\boldsymbol{\chi^2_{cal}}$**=30.82688**

Degree of freedom $v = (m-1) \times (n-1) = 1 \times 3 = 3$

$\boldsymbol{\chi^2_v}(\alpha) = \boldsymbol{\chi^2_3}(5\%) = 7.815$

Since $\boldsymbol{\chi^2_{cal}} > \boldsymbol{\chi^2_v}(\alpha)$

H_0 *is rejected*

We can conclude that there is some relation between use of needle exchange and agency status.

i.e. we can conclude from these data that use of needle exchange and agency status is related

7) Concern about acquired immunodeficiency syndrome (AID) was the motivation for a survey conducted by Professor Patty J. Hale of the University of Virginia. She used a mailed questionnaire to survey business. Among the information she collected were size of business and whether or not the employer had provided AIDS education for employees. The following results we reported.

Number of employees ↓ \ AIDS education provided→	YES	NO	Total
0-10	**2**	**20**	22
50-500	**5**	**11**	16
More than 500	**11**	**5**	16
Total	18	36	54

May we concluded on the basis of these data that whether or not a business provides AIDS education is independent of the size of the business? Let $\propto = 0.5$

Solution:-

H_0: *AIDS education is independent of the size of the business*

Since table is of size 3× 2

Since every cell frequency is not greater than 5

So we combine first and second row

Number of employees/AIDS education provided	C_1	C_2	**Total**
R_1	7=a	31=b	38= a+b
R_2	11=c	5=d	16= c+d
Total	**18= a+c**	**36=b+d**	**54=N=a+b+c+d**

∵ every cell frequency is greater than 5

$\therefore \quad \chi^2_{cal} = \frac{N\times(ad-bc)^2}{(a+b)(a+c)(d+c)(d+b)} = \frac{54\times[(7\times5)-(11\times31)]^2}{(38)\times(18)\times(16)\times(36)} = 12.833882$

Degree of freedom $v=(m-1)\times(n-1) = 1\times 1 = 1$

$\chi^2_v(\alpha) = \chi^2_1(5\%) = 3.84146$

Since $\chi^2_{cal} > \chi^2_v(\alpha)$

H_0 *is rejected*

AIDS education is not independent of the size of the business

Prof. Sampat Mali
Department of Mathematics

Chi-square test of homogeneity

Null hypothesis is

H_0:- Are the samples drawn from populations that **are homogeneous** with respect to some criterion of classification

1)Kodama et al. studied the relationship between age and several prognostic factors in squamous cell carcinoma of the cervix among the data collected were the frequencies of histologic cell type in four age groups. The results are shown in the following table

Age group	Cell type			Total Number
	Large cell non keratinizing cell type	keratinizing cell type	Small cell non keratinizing cell type	
30-39	18	7	9	34
40-49	56	29	12	97
50-59	83	38	23	144
60-69	62	25	18	105
Total	219	99	62	380

We wish to know if **we may conclude that the populations represented by the four age-group samples are not homogenous with respect to cell type**

Solution:

H_0: *Four age groups are homogeneous with respect to cell type*

Age group	Cell type						**Total Number**
	Large cell non keratinizing cell type		keratinizing cell type		Small cell non keratinizing cell type		
30-39	O_{11} =18	E_{11} **=19.59 ≈ 19**	O_{12} =7	E_{12} **=8.86 ≈ 9**	O_{13} =9	E_{13} **=5.54 ≈ 6**	**34**
40-49	O_{21} =56	E_{21} **=55.90 ≈ 56**	O_{22} =29	E_{22} **=25.27 ≈ 25**	O_{23} = 12	E_{23} **=15.83 ≈16**	**97**
50-59	O_{31} =83	E_{31} **=82.99 ≈83**	O_{32} =38	E_{32} **=37.52 ≈38**	O_{33} =23	E_{33} **=23.49 ≈23**	**144**
60-69	O_{41} =62	E_{41} **=60.51 ≈ 61**	O_{42} =25	E_{42} **=27.36 ≈27**	O_{43} =18	E_{43} **=17.13 ≈ 17**	**105**
Total	**219**		**99**		**62**		**380**

$$\chi^2_{cal} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\right]$$

$$\chi^2_{cal} = \frac{(18-19)^2}{19} + \frac{(7-9)^2}{9} + \frac{(9-6)^2}{6} + \frac{(56-56)^2}{56} + \frac{(29-25)^2}{25} + \frac{(12-16)^2}{16} +$$

$$+ \frac{(83-83)^2}{83} + \frac{(38-38)^2}{35} + \frac{(23-23)^2}{23} + \frac{(62-61)^2}{61} + \frac{(25-27)^2}{27} + \frac{(18-17)^2}{17} = 3.860441$$

Thus χ^2_{cal} = 3.860441

Degree of freedom $v=(m-1)\times(n-1) = 3\times 2 = 6$

$\chi^2_v(\alpha) = \chi^2_6(5\%) = 12.592$

Since $\chi^2_{cal} < \chi^2_v(\alpha)$

H_0 *is accepted*

We can conclude that Four age groups are homogeneous with respect to cell type

2) In a telephone survey conducted by professor Bikram Garcha(A-9) responds were asked to indicate their level of agreement with the statement ''Cigarette smoking should be banned in public places" The results were as follows

Gender	Level of agreement					Total
	Strongly agree	Agree	neutral	Disagree	Strongly disagree	
Female	40	38	16	37	5	136
Male	16	25	11	25	10	87
Total	56	63	27	62	15	223

Can we conclude on the basis of these data that males and females differ with respect to their level of agreement on the banning of cigarette smoking in public place?

Solution:-

H_0: – Male and female are homogeneous with respect to their level of agreement on the banning of cigarette smoking in public place.

Gender	**Level of agreement**										**Total**
	Strongly agree		**Agree**		**neutral**		**Disagree**		**Strongly disagree**		
Female	O_{11} =40	E_{11} =34.15 ≈ 34	O_{12} =38	E_{12} =38.42 ≈ 38	O_{13} =16	E_{13} =16.47 ≈17	O_{14} =37	E_{14} =37.81 ≈38	O_{15} =5	E_{15} =9.14 ≈ 9	**136**
Male	O_{21} =16	E_{21} =21.85 ≈ 22	O_{22} =25	E_{22} =24.58 ≈ 25	O_{23} =11	E_{23} =10.53 ≈10	O_{24} =25	E_{24} =24.19 ≈ 24	O_{25} =10	E_{25} =5.85 ≈6	**87**
Total	**56**		**63**		**27**		**62**		**15**		**223**

$$\chi^2_{cal} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\right]$$

$$\chi^2_{cal} = \frac{(40-34)^2}{34} + \frac{(38-38)^2}{38} + \frac{(16-17)^2}{17} + \frac{(37-38)^2}{38} + \frac{(5-9)^2}{9} +$$

$$+\frac{(16-22)^2}{22} + \frac{(25-25)^2}{25} + \frac{(11-10)^2}{10} + \frac{(25-24)^2}{24} + \frac{(10-6)^2}{6} = 7.366438$$

Thus χ^2_{cal} = 7.366438

Degree of freedom $v=(m-1)\times(n-1) = 1\times 4 = 4$

$\chi^2_v(\alpha) = \chi^2_4(5\%) = 9.488$

Since $\chi^2_{cal} < \chi^2_v(\alpha)$

H_0 *is accepted*

We can conclude that male and female are homogeneous with respect to their level of agreement on the banning of cigarette smoking in public place.

Prof. Sampat Mali
Department of Mathematics

3) In an air pollution study, a sample of 200 house holds was selected from each of two communities. A respondent in each household was asked whether or not anyone in the household was bothered by air pollution. The response were as follows

Community	Any member of household bothered by air pollution		Total
	Yes	No	
A	43	157	200
B	81	119	200
Total	124	276	400

If there is communities differ with respect to variable of interest?

Solution:-

H_0: *Communities are homogeneous with respect to variable of interest*

Sine every cell frequencies is greater than 5

Given table is of order 2× 2

Therefore

Community	**Any member of household bothered by air pollution**		**Total**
	Yes	**No**	
A	43=a	157=b	**200=a+b**
B	81=c	119=d	**200=c+d**
Total	**124=a+c**	**276=b+d**	**400=N=a+b+c+d**

$\therefore \quad \chi^2_{cal} = \frac{N\times(ad-bc)^2}{(a+b)(a+c)(d+c)(d+b)} = \frac{400\times[(43\times119)-(81\times157)]^2}{(200)\times(124)\times(200)\times(276)} = \mathbf{16.877045}$

Degree of freedom $v=(m-1)\times(n-1)=1\times1=1$

$\chi^2_v(\alpha) = \chi^2_1(5\%) = 3.841$

Since $\chi^2_{cal} > \chi^2_v(\alpha)$

H_0 *is rejected*

We can conclude that

communities are not homogeneous with respect to variable of interest

i.e. there is communities differ with respect to variable of interest

4) In a simple random sample **of 250 industrial workers with cancer researchers found that** 102 had worked at jobs classified "high exposure" with respect to suspected cancer agents. **Of the remainder, 84 had worked at "Moderate exposure" jobs** , In an independent simple random sample of **250 industrial workers from the same area who had no history of cancer,** 31 worked in "high exposure" job 60 worked in "moderate exposure" jobs and 159 worked in job involving no known exposure to suspected cancer causing agents. Does it appear from these data that person working in jobs that expose them to suspect cancer causing agents have an increased risk of contracting cancer?

Solution:-

H_0:Person working in jobs that not expose them to suspect cancer causing agents have an increased risk of contracting cancer

Sample	**high exposure**		**Moderate exposure**		**no known exposure**		**Total**
Sample-1	O_{11} =102	E_{11} **=66.5**	O_{12} =84	E_{12} **=72**	O_{13} =64	E_{13} **=111.5**	**250**
Sample-2	O_{21} =31	E_{21} **=66.5**	O_{22} =60	E_{22} **=72**	O_{23} =159	E_{23} **=111.5**	**250**
Total	**133**		**144**		**223**		**500**

$$\chi^2_{cal} = \sum_{i=1}^{m}\sum_{j=1}^{n}\left[\frac{(O_{ij}-E_{ij})^2}{E_{ij}}\right]$$

$$\chi^2_{cal} = \frac{(102-66.5)^2}{66.5} + \frac{(84-72)^2}{72} + \frac{(64-111.5)^2}{111.5} + \frac{(31-66.5)^2}{66.5} + \frac{(60-72)^2}{72} + \frac{(159-111.5)^2}{111.5} = 20.235$$

Thus χ^2_{cal} =20.235

Degree of freedom $v=(m-1)\times(n-1)=1\times 2=2$

$\chi^2_v(\alpha) = \chi^2_2(5\%) = 5.991$

Since $\chi^2_{cal} > \chi^2_v(\alpha)$

H_0 *is arejected*

We can conclude that Person working in jobs that expose them to suspect cancer causing agents have an increased risk of contracting cancer**.**

Prof. Sampat Mali
Department of Mathematics

5) The objective of a study by Sutker et al. describe the long-term psychological and psychiatric sequence of prisoner of war confinement against the backdrop of psychiatric evaluation of Korean conflict repatriates more than 35 years in the past. Subjects were 22 (POW) and 22 combat veteran survivors (CVS) of the Korean conflict. They were compared on measures of problem solving. Personality characteristics mood states and psychiatric clinical diagnoses. **Nineteen of the POWs reported problem with depression**. The number of combat veterans reporting **problem with depression was 9.** Do these data provide sufficient evidence for us to conclude that the two populations are not homogeneous with respect to the incidence of problem of depression?

Solution:-

H_0: *Two population are homogeneous with respect to the incidence of problem of depression*

	Problem with depression	problem without depression	Total
POW	19=a	3=b	22=a+b
CVS	9=c	13=d	22=c+d
Total	28=a+c	16=b+d	44=N=a+b+c+d

Since given contingency table is of order 2× 2 and one of the cell frequency is less than 5
So we use the formula

$$\chi^2_{cal} = \frac{N\times(|ad-bc|-N/2)^2}{(a+b)(a+c)(d+c)(d+b)} = \frac{44\times(|(19\times13)-(9\times3)|-22)^2}{(22)(28)(22)(16)} = 7.955357$$

Degree of freedom $v=(m-1)\times(n-1)=1\times1=1$

$\chi^2_v(\alpha) = \chi^2_1(5\%) = 3.84146$

Since $\chi^2_{cal} > \chi^2_v(\alpha)$

H_0 *is rejected*

$\therefore$ ***we can conclude that***

two populations are not homogeneous with respect to the incidence of problem of depression.