Chi-Square Distribution

Define χ^2 distribution.

If $Z_1, Z_2, Z_3, \ldots, Z_n$ are independent standardized normal variables with zero mean and unit variance. Then sum and square of these variables are called Chi-Square statistics

$\chi_c^2 = Z_1^2 + Z_2^2 + Z_3^2 + \ldots + Z_n^2 = \sum_{i=1}^{n}\left(\frac{X-\mu}{\sigma}\right)^2$ with "n" degree of freedom and having its

probability density function (pdf) $f(\chi^2) = \dfrac{(\chi^2)^{\frac{n}{2}-1} e^{-\chi^2/2}}{\Gamma\frac{n}{2}\, 2^{\frac{n}{2}}}$ $\quad 0 \le \chi^2 \le \infty$

Properties of chi-square distribution

Important properties are given below

i) It is continuous distribution.

ii) Total area under curve is unity

iii) Its ranges from $0 \; to \; +\infty$ and +vely skewed and single peaked curve

iv) It has only one parameter degree of freedom. It is denoted by $n \;\; or \;\; v$

v) Its $Mean = v = \deg ree \;\; of \;\; freedom$ and $Variance = 2v = 2(\deg ree \;\; of \;\; freedom)$

vi) Chi-square distribution approaches to normal distribution as "n" tends to infinity.

vii) The sum of two independent chi-random variable is also a chi-square variable.

viii) When n=1 its curve is J-Shaped and skewness is highest.

ix) Mode is "n-2"

x) Coefficient of skewness $\beta_1 = \frac{8}{n}$ And Coefficient of kurtosis $\beta = 3 + \frac{12}{n}$

xi) Partitioning property is $\chi_n^2 = \chi_{(n-1)}^2 + \chi_{(1)}^2$

xii) Moment generating function of Chi-Square distribution is $M(t) = (1-2t)^{-\frac{n}{2}}$

xiii) Characteristic function of Chi-Square distribution is $M(t) = (1-2it)^{-\frac{n}{2}}$

xiv) rth cummulant of Chi-Square distribution is $K_r = n2^{r-1}(r-1)!$

xv) If Z_1 and Z_2 are two independent chi-square variable with "n_1 and n_2" degree of freedom then the sum of Z_1+Z_2 is also chi-square variable with "n_1+n_2" degree of freedom

Uses of Chi-Square distribution

i) It is used to test the goodness of fit

ii) It is used to test the equality of proportion of multinomial distribution

iii) It is used to test the independence of attributes in contingency table

iv) It is used find confidence interval for population variance

v) It is used to test the variance of normal population

vi) It is used test the equality of "k" population variances

vii) It is used to find confidence interval for several population variances

viii) It is used to test the homogeneity of population correlation coefficient

ix) It is used to test the homogeneity of population correlation coefficient

What is condition to apply chi-square distribution?

Ans: The following condition to apply chi-square distribution are given below

i) The sample size or total no. of observations must be greater than 50.

ii) Observations or cell frequency must be independence.

iii) Each cell frequency must be greater than 5, otherwise it will be pooled with previous one.

Assumptions of chi-square tests

The sample is selected independently and randomly

The population from which sample is selected is normally distributed.

Test of homogeneity

Chi-Square test is used to test the homogeneity that two or more than two different samples come from the same population or that samples are homogeneous. In statistic it is often used to indicate "the same" or "equal" we call it a test of homogeneity

Why we use correction of continuity in contingency table?

In contingency table the data is discrete and we are fitting it chi-square distribution that is continuous. So, we need to transfer/convert a discrete variable into continuous variable a correction is needed is called continuity correction.

What is the appropriate goodness of fit criterion for discrete distribution?
Here we discuss in two ways

a) If it is mentioned in the question to fit the following distribution to check its goodness of fit then it will be very easy to follow the said distribution
b) But in many circumstances the name of the distribution is not mentioned for goodness of fit is asked to fit an appropriate distribution. Then which distribution is to be fitted to check the goodness of fit? There are simple criteria for fitting it. We calculate the mean and variance of the given frequency distribution then we compare their mean with variance.

i) If the mean is equal to approximately equal to the variance then we prefer Poisson distribution
ii) If the mean is greater than the variance then we prefer binomial distribution
iii) If the mean is greater less than the variance then we prefer negative binomial distribution

What is the purpose of the goodness of fit test?
A goodness of fit test is a hypothesis test that is concerned with the determination whether results of a sample conform to a hypothesized distribution which may be uniform, binomial, Poisson, normal or any other distribution. The test statistic to be used for goodness of fit is $\chi_c^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i}$ Which, when H_0 is true, has an approximately chi-square distribution with v=k-1-(no. of parameter estimated) degree of freedom. A small computed value of chi-square indicates a good fit and it leads the acceptance of the null hypothesis while a large computed value of chi-square indicates a poor fit and it leads to the rejection of the null hypothesis.

Differentiate between attribute and variable?
Ans: Variable: A characteristic that varies from one object to another object either in quality or quantity is called variable. Such as age, height, prices etc.
Attribute: A characteristic that varies from one object to another only in quality and cannot be measurable is called attributes. Such as beauty, intelligence, taste etc.

Explain the consistency of the data.
If no ultimate class frequency is negative then data is said to be consistent otherwise inconsistent.

When two attributes are said to be positively associated?
Ans: Two attributes "A and B" are said to be positively associated, if
$(AB) > \frac{(A)(B)}{n}$

When two attributes are said to be negatively associated?
Ans: Two attributes "A and B" are said to be negative associated, if
$(AB) < \frac{(A)(B)}{n}$

When two attributes are said to be associated/ dependent?
Ans: Two attributes "A and B" are said to be associated or dependent, if
$(AB) \neq \frac{(A)(B)}{n}$

Explain what is meant by independence of attributes?
Ans: Two attributes "A and B" are said to be independent, if the occurrence of one does not affect the other. i.e. $(AB) = \frac{(A)(B)}{n}$

What do you understand by association?
Ans: Two attributes "A and B" are said to be associated, if the occurrence of one effect the occurrence other. i.e. $(AB) \neq \frac{(A)(B)}{n}$

Explain the terms independence and association as applied to attributes.
Ans: Two attributes "A and B" are said to be independent, if the occurrence of one does not affect the other. i.e. $(AB) = \frac{(A)(B)}{n}$.
Two attributes "A and B" are said to be associated, if the occurrence of one affect the occurrence of other. i.e. $(AB) \neq \frac{(A)(B)}{n}$.

Define a contingency
Ans: A table showing the cross-tabulation or joint distribution of two variables is known as contingency.
Explain the positive and negative association
Ans: Positive association
Association between the attributes is called positive if their observed frequency is grater than expected frequency.
i.e. $(AB) > \frac{(A)(B)}{n}$
Negative association
Association between the attributes is called negative if their observed frequency is less than expected frequency.
i.e. $(AB) < \frac{(A)(B)}{n}$
What is meant by attributes?
Ans: Any characteristic that varies only in quality from one individual to other is called attributes. It cannot be measured only its presence or absence is studied. Such as kindness, taste, intelligence etc.
Interpret the meaning of coefficient of association "Q" when
Ans: Yule's' coefficient of association denote by "Q" is defined as

$$Q = \frac{(AB)(\alpha\beta) - (A\beta)(\alpha B)}{(AB)(\alpha\beta) + (A\beta)(\alpha B)}$$

If $Q = 0$ the attributes are independent
If $Q = +1$ the attributes are completely associated
If $Q = -1$ the attributes are completely disassociated
Explain the general procedure for test of independence between the attributes.
Ans: Independence between attributes in a contingency table is tested by χ^2. The procedure involves six steps. The procedure is given below.
Procedure:
Setp-i: We set up our null and alternative hypothesis
H_0 : *The attributes are independence*
H_1 : *The attributes are dependence*
Step-ii: Assumption: A sample is drawn randomly and independently from a approximately normal population
Step-iii: Level of significance
$\alpha = (Comonly \;\; uesd\ 5\% \ or \ 1\%)$
Step-iv: Test-statistic

$$\chi^2 = \sum \frac{(O_i - e_i)^2}{e_i} \quad \text{Without continuity correction}$$

$$\chi^2 = \sum \frac{\left(|O_i - e_i| - \frac{1}{2}\right)^2}{e_i} \quad \text{With continuity correction}$$

For
2×2 Contingency table is given as

	A_1	A_2	
B_1	A	B	A+b
B_2	C	D	C+d
Total	a+c	B+d	N

Under H_0; χ^2 it has –distribution with $v = (r-1)(c-1)$
Direct formula

$$\chi^2 = \frac{n(ad - bc)^2}{(a+b)(a+c)(c+d)(b+d)} \quad \text{Without continuity correction}$$

$$\chi^2 = \frac{n\left(|ad-bc| - \frac{n}{2}\right)^2}{(a+b)(a+c)(c+d)(b+d)}$$ With continuity correction

Under H_0; χ^2 it has –distribution with $v=(r-1)(c-1)$

Step-v: Critical region
It is naturally depend on alternative hypothesis
$\chi^2 > \chi_\alpha^{\ 2}(v)$
Step-vi: Calculation
In this step we calculate the value of " χ^2 " test statistic on the basis of sample data.
Step-vii: Conclusion
If our calculated value does not fall's in critical region then we accept H_0 other wise we reject it.

Write down the direct formula for calculating χ^2 in a 2*2 contingency table.

Ans: The direct formula for calculating χ^2 in a 2*2 contingency table are given as

$$\chi^2 = \frac{n(ad-bc)^2}{(a+b)(a+c)(c+d)(b+d)}$$ With $v=(r-1)(c-1)$ d.f

$$\chi^2 = \frac{n\left(|ad-bc| - \frac{n}{2}\right)^2}{(a+b)(a+c)(c+d)(b+d)}$$ With continuity correction

Explain the coefficient of association.

Ans: The measure of degree of relationship between the attributes is called coefficient of association.

Explain the term positive and negative attributes.

Positive attributes

The attributes having **Negative** the qualities of interest are called positive attributes. These are denoted by English letters as A, B, C, etc.
Or
The attributes A, B, C, AB, AC, etc.are called as positive attributes.

Attributes

The attributes does not having the qualities of interest are called positive attributes. These are denoted by Greek letters as $\alpha, \beta, \gamma, etc$.
The attributes $\alpha, \beta, \gamma, etc$ are called as negative attributes.

Define ultimate class frequencies?

Ans: The class frequencies of highest order are called ultimate classes. In case of two attributes (AB), $(A\beta)$, (αB), $(\alpha\beta)$ are called ultimate frequencies.

Explain coefficient of contingency?

Ans: A measure of degree of association between the attributes expressed a contingency table is known as coefficient of contingency. Pearson's mean square co-efficient of contingency denoted by "C" is given by

$$C = \sqrt{\frac{\chi^2}{n+\chi^2}} \qquad 0 \le k \le \sqrt{\frac{k-1}{k}}$$

Where "k" is number of rows or columns which is smaller and "n" is the sample size.
How will you define a class in the theory of attributes?
Ans: In the theory of attributes, the objects having the same attributes are called a class.

What is meant by consistency of frequency?

Ans: If the observed frequencies are recorded correctly and no ultimate class frequency is negative, they are said to be consistent.

What is meant by order of classes?

Ans: The number of attributes present in a class is called order of the class for example A,B,C, α, β, γ are the classes of order lone similarly $A\beta$ and αB are classes of order two and so on. The order of sample size "n" is zero because it contains no attributes.

Relationship between Chi-Square distribution and Gamma Distribution

Chi-square distribution is a special case of gamma distribution with $\alpha = \frac{n}{2}$ and $\beta = 2$

Where "n" is the degree of freedom of Chi-square distribution.

Explain the test of independence

The data presented in a contingency table can be used to test the hypothesis that the two variables of classification are independence. If this hypothesis is rejected the two variables of classification are not independence there is some association between them.

The test statistic for this testing of hypothesis is $\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij} - e_{ij})^2}{e_{ij}}$

Where r=no. of rows and c= no. of columns

Under H_0 if it is true of independence which is approximately chi square distribution with v=(r-1)(c-1) degree of freedom

Where

O_{ij}=observed frequencies

e_{ij}=expected frequencies calculated as $\frac{(A_i)(B_j)}{n}$

A large value of chi square indicates that the null hypothesis is false.

When Bartlett, Hartley and Cochran test used for homogeneity of variances are recombined?

a) Bartlett test is used for the several variances when the samples sizes are not equal

b) Hartley test is used for the several variances when the samples sizes are equal but less then or equal to 12

c) Cochran test is used for the several variances when the samples sizes are same but there is no restriction.

Q: 17.5(a): Explain how you determine a confidence interval estimate of σ^2 of a normal population.

Ans: The confidence interval estimate of the population variance σ^2 is based on the sampling distribution of S^2 , the sample variance and the sampling distribution of S^2 is the Chi-Square distribution. We therefore use the χ^2 distribution to obtain the confidence interval for σ^2.

Let $\bar{x} = \frac{\sum x}{n}$ and $S^2 = \frac{\sum_{i=1}^{n}(x-\bar{x})^2}{n} = \frac{\sum_{i=1}^{n}x_i^2 - n\bar{x}^2}{n}$ be the mean and variance of a random sample $X_1, X_2, X_3, \ldots, X_n$ of size "n" drawn from a normal with variance σ^2. Then the statistic $\chi^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} = \frac{nS^2}{\sigma^2}$ or $\chi^2 = \frac{(n-1)s^2}{\sigma^2}$ $is\ s^2 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}$ from the sample mean to the population variance, has a Chi-Square distribution with (v=n-1) degree of freedom to construct a two sided confidence interval for σ^2 we proceed as Then $(1-\alpha)\%$ confidence interval for population variance σ^2 in form of probability statement

$$P\left[\chi^2_{1-\frac{\alpha}{2}}(v) \le \chi_c^2 \le \chi^2_{\frac{\alpha}{2}}(v)\right] = 1-\alpha$$

Substituting the value of χ_c^2

$$P\left[\chi^2_{1-\frac{\alpha}{2}}(v) \le \frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{\sigma^2} \le \chi^2_{\frac{\alpha}{2}}(v)\right] = 1-\alpha$$

Dividing inside the bracket $\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$

$$P\left[\frac{\chi^2_{1-\frac{\alpha}{2}}(v)}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2} \le \frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{\sigma^2\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2} \le \frac{\chi^2_{\frac{\alpha}{2}}(v)}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\right] = 1-\alpha$$

$$P\left[\frac{\chi^2_{1-\frac{\alpha}{2}}(v)}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2} \le \frac{1}{\sigma^2} \le \frac{\chi^2_{\frac{\alpha}{2}}(v)}{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}\right] = 1-\alpha$$

Taking the inverse of R.H.S and sign of inequality will be changed

$$P\left[\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{\chi^2_{1-\frac{\alpha}{2}}(v)} \ge \sigma^2 \ge \frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{\chi^2_{\frac{\alpha}{2}}(v)}\right] = 1-\alpha$$

Or

$$P\left[\frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{\chi^2_{\frac{\alpha}{2}}(v)} \le \sigma^2 \le \frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{\chi^2_{1-\frac{\alpha}{2}}(v)}\right] = 1-\alpha$$

Or

$$P\left[\frac{nS^2}{\chi^2_{\frac{\alpha}{2}}(v)} \le \sigma^2 \le \frac{nS^2}{\chi^2_{1-\frac{\alpha}{2}}(v)}\right] = 1-\alpha$$

Hence required confidence interval for population variance

Confidence interval for several population variances

Solution: Suppose $n_1, n_2, n_3, \ldots, n_k$ random sample are drawn from "k" different population $N_1, N_2, N_3, \ldots, N_k$ having a common variances $\sigma_1^2 = \sigma_2^2 = \sigma_3^2 = \ldots = \sigma_k^2 = \sigma^2$ or same with population variance σ^2.

Let $S_1^2, S_2^2, S_3^2, \ldots, S_k^2$ be the sample variances computed from k random samples then S_p^2 denotes the polled unbiased estimate of population variance σ^2 where

$$S_p^2 = \frac{n_1 S_1^2 + n_2 S_2^2 + \ldots + n_k S_k^2}{n_1 + n_2 + \ldots + n_k} = \frac{\sum_{i=1}^{k} n_i S_i^2}{\sum_{i=1}^{k} n_i - k}$$

Now

$$\chi^2 = \frac{\sum_{i=1}^{k} n_i S_i^2}{\sigma^2} = \frac{(\sum_{i=1}^{k} n_i - k) S_p^2}{\sigma^2}$$ Would be chi-square distribution with $V = (\sum_{i=1}^{k} n_i - k)\, df$

Then $(1-\alpha)\%$ confidence interval for population variance σ^2 in form of probability statement

$$P\left[\chi^2_{1-\frac{\alpha}{2}}(v) \le \chi_c^2 \le \chi^2_{\frac{\alpha}{2}}(v)\right] = 1-\alpha$$

Substituting the value of χ_c^2

$$P\left[\chi^2_{1-\frac{\alpha}{2}}(v) \le \frac{\sum_{i=1}^{n}\left(X_i-\overline{X}\right)^2}{\sigma^2} \le \chi^2_{\frac{\alpha}{2}}(v)\right] = 1-\alpha$$

Dividing inside the bracket $\sum_{i=1}^{n}\left(X_i - \overline{X}\right)^2$

$$P\left[\frac{\chi^2_{1-\frac{\alpha}{2}}(v)}{\sum_{i=1}^{k} n_i S_i^2} \le \frac{\sum_{i=1}^{k} n_i S_i^2}{\sigma^2 \sum_{i=1}^{k} n_i S_i^2} \le \frac{\chi^2_{\frac{\alpha}{2}}(v)}{\sum_{i=1}^{k} n_i S_i^2}\right] = 1-\alpha$$

$$P\left[\frac{\chi^2_{1-\frac{\alpha}{2}}(v)}{\sum_{i=1}^{k} n_i S_i^2} \le \frac{1}{\sigma^2} \le \frac{\chi^2_{\frac{\alpha}{2}}(v)}{\sum_{i=1}^{k} n_i S_i^2}\right] = 1-\alpha$$

Taking the inverse of R.H.S and sign of inequality will be changed

$$P\left[\frac{\sum_{i=1}^{k} n_i S_i^2}{\chi^2_{1-\frac{\alpha}{2}}(v)} \ge \sigma^2 \ge \frac{\sum_{i=1}^{k} n_i S_i^2}{\chi^2_{\frac{\alpha}{2}}(v)}\right] = 1-\alpha$$

Or

$$P\left[\frac{\sum_{i=1}^{k} (n_i - k) S_p^2}{\chi^2_{\frac{\alpha}{2}}(v)} \le \sigma^2 \le \frac{\sum_{i=1}^{k} (n_i - k) S_p^2}{\chi^2_{1-\frac{\alpha}{2}}(v)}\right] = 1-\alpha$$

Q.17.5(b): Given that X is normally distributed and given the sample values $\overline{X} = 42$, S = 5 and n= 20.Find the 98 % Confidence Interval for σ^2.

Solution:

The $100(1-\alpha)\%$ confidence interval for population variance σ^2

$$\frac{nS^2}{\chi^2_{\frac{\alpha}{2}(v=n-1)}} < \sigma^2 < \frac{nS^2}{\chi^2_{1-\frac{\alpha}{2}(v=n-1)}}$$

Given that $n = 20$ $\overline{X} = 42$ $S = 5$ $\chi^2_{\frac{\alpha}{2}(v=n-1)} = \chi^2_{0.01(19)} = 36.19$ $\chi^2_{1-\frac{\alpha}{2}(v=n-1)} = \chi^2_{0.99(19)} = 7.63$

Then 98% C.I for σ^2

$$\frac{20(25)}{36.19} < \sigma^2 < \frac{20(25)}{7.63}$$

$$13.82 < \sigma^2 < 65.53$$

Hence we have 98% confidence that our parameter lying the interval (13.82,65.53)

Q: 17.6(a): The following are the volumes in deciliters , of 10 cans of peaches distributed by a certain company: 46.4, 46.1, 45.8, 47.0, 46.1, 45.9, 45.8, 46.9, 45.2 and 46.0.Find a 95 % confidence interval for the variance of all such cans of peaches distributed by this company, assuming volume to be a normally distributed variable.

Solution:

The $100(1-\alpha)\%$ confidence interval for population variance σ^2

$$\frac{nS^2}{\chi^2_{\frac{\alpha}{2}(v=n-1)}} < \sigma^2 < \frac{nS^2}{\chi^2_{1-\frac{\alpha}{2}(v=n-1)}}$$

Given that

X	$(X-\overline{X})$	$(X-\overline{X})^2$
46.4 46.1 45.8 47.0 46.1 45.9 45.8 46.9 45.2 46.0		
$\sum X = 461.2$	$\sum (X-\overline{X}) = 0$	$\sum (X-\overline{X})^2 = 2.576$

$$\overline{X} = \frac{\sum X}{n} = \frac{461.2}{10} = 46.12$$

$$S^2 = \frac{\sum(\mathrm{X}-\overline{\mathrm{X}})^2}{n} = \frac{2.576}{10} = 0.2576$$

$n = 10$ $\quad \chi^2_{\frac{\alpha}{2}(v=n-1)} = \chi^2_{0.025(9)} = 19.02 \quad \chi^2_{1-\frac{\alpha}{2}(v=n-1)} = \chi^2_{0.975(9)} = 2.70$

Then 95% C.I for σ^2

$$\frac{10(0.2576)}{19.02} < \sigma^2 < \frac{10(0.2576)}{2.70}$$

$0.14 < \sigma^2 < 0.95$

Hence we have 95% confidence that our parameter lying the interval (0.14,0.95)

Q.17.6(b): The contents of 10 similar containers of a commercial soap, are 10.2, 9.7, 10.1, 10.3, 10.1, 9.8, 9.9, 10.4, 10.3 and 9.8 litres.Find a 95 % confidence interval for the variance of all such containers , assuming an approximately normal distribution.

Solution: Same as Q.6(a)

Q: 17.7(a): Given the following sample values from a normal population, Find 96 % confidence limit for σ^2 based on combining these sample values properly. The sample variances are :$S_1^2=25$, $S_2^2=36$, $S_3^2 = 16$ with $n_1 = 5$, $n_2 = 5$, $n_3 = 10$.

Solution: Same as Q.8

Q.17.7(b): Assume that the random variable X is N(μ, σ^2).Three random samples of X provide the following information

Sample Size : 3, 5, 7

Sample Mean: 42, 45, 40

Sample Variance: 25, 16, 9

Obtain a 95 % Confidence Interval for σ^2.

Solution: Same as Q.8

Q: 17.8: Suppose 10 samples of 9 values each, have variances as follows:

23.5, 30.6, 29.3, 27.5, 27.5, 26.3, 29.8, 30.7, 22.3, 26.5.

Obtained a pooled estimate of σ^2 and use it to find 90 % confidence interval limits for σ^2.

Solution:

The $100(1-\alpha)\%$ confidence interval for population variance σ^2

$$\frac{\sum_{i=1}^{k} n_i S_i^2}{\chi^2_{\frac{\alpha}{2}(v=\sum_{i=1}^{k} n-k)}} < \sigma^2 < \frac{\sum_{i=1}^{k} n_i S_i^2}{\chi^2_{1-\frac{\alpha}{2}(v=\sum_{i=1}^{k} n-k)}}$$

n_i	S_i^2	$n_i S_i^2$
9	23.5	211.6
9	30.6	275.4
9	29.3	263.7
9	27.5	247.5
9	27.5	247.5
9	26.3	236.7
9	29.8	268.2
9	30.7	276.3
9	22.3	200.7
9	26.5	238.5
$\sum n_i = 90$		$\sum n_i S_i^2 = 2466.1$

$1-\alpha = 0.90$ *Than* $\alpha = 0.10$

$$\chi^2_{\frac{\alpha}{2}(v=n-1)} = \chi^2_{0.05(80)} = ?$$

To find the table value by using Fisher approximation

$$Z_\alpha = \left[\sqrt{2\chi^2_\alpha} - \sqrt{2n-1}\right]$$

$$\chi^2_\alpha = \frac{1}{2}\left[Z_\alpha + \sqrt{2n-1}\right]^2$$

$$\chi^2_{0.0.5(80)} = \frac{1}{2}\left[Z_{0.05} + \sqrt{2n-1}\right]^2 = \frac{1}{2}\left[1.645 + \sqrt{2(80)-1}\right]^2 = 101.60$$

$$\chi^2_{1-\frac{\alpha}{2}(v=n-1)} = \chi^2_{0.95(80)} = ?$$

$$\chi^2_{0.95(80)} = \frac{1}{2}\left[Z_\alpha + \sqrt{2n-1}\right]^2 = \frac{1}{2}\left[-1.645 + \sqrt{2(80)-1}\right]^2 = 60.11$$

Then 90% C.I for σ^2

$$\frac{2466.1}{101.60} < \sigma^2 < \frac{2466.1}{60.11}$$

$$24.27 < \sigma^2 < 41.03$$

Hence we have 90% confidence that our parameter lying the interval (24.27,41.03)

Q: 17.9(a): Explain how you would test the hypothesis about variance of a normal population.

Ans: Suppose we want to test the hypothesis that the population variance is equal to some specific value σ_0^2. Let a random sample of size "n" are drawn from a normal population with variance σ^2. Than $\chi_c^2 = \frac{nS^2}{\sigma_0^2} = \frac{\sum(X_i - \bar{X})^2}{\sigma_0^2}$

It has χ^2-distribution with V=n-1 degree of freedom

Procedure

i) We state our null and alternative hypothesis

a) $H_0 : \sigma^2 = \sigma_0^2$ b) $H_0 : \sigma^2 \geq \sigma_0^2$ c) $H_0 : \sigma^2 \leq \sigma_0^2$

$H_1 : \sigma^2 \neq \sigma_0^2$ $H_1 : \sigma^2 < \sigma_0^2$ $H_1 : \sigma^2 > \sigma_0^2$

ii) Assumption: Samples are drawn randomly and independently from a normal population having the variance σ^2

Level of significance:

$\alpha = (\textit{Commonly use } 5\% \text{ 0}r\ 1\%)$

Test-Statistic

$$\chi_c^2 = \frac{nS^2}{\sigma_0^2} = \frac{\sum(X_i - \bar{X})^2}{\sigma_0^2}$$

Under H_0, Which has χ^2-distribution with V=n-1 degree of freedom

v) Critical region

a) If $H_1 : \sigma^2 \neq \sigma_0^2$ than we use two sided test

$\chi_c^2 < \chi^2_{(1-\frac{\alpha}{2})}(n-1)$ or $\chi_c^2 > \chi^2_{\frac{\alpha}{2}}(n-1)$

b) If $H_1 : \sigma^2 < \sigma_0^2$ than we use one sided test

$\chi_c^2 < \chi^2_{1-\alpha}(n-1)$

c) If $H_1 : \sigma^2 > \sigma_0^2$ than we use one sided test

$\chi^2 > \chi^2_\alpha(n-1)$

vi) Calculation

In this step we calculate the value of test statistic on the basis of sample data

vii) Conclusion

If our calculated value does not fall in critical region, then we accept H_0 otherwise we reject it.

Q.17.9(b): A sample of 9 parts produced by a certain production process are measured as 5, 7, 2, 4, 8, 9, 8, 6, and 5 inches respectively. Test the hypothesis that the process has the variance equal to 4(inches)^2 at the 5 % level.

Solution:

i) We state our null and alternative hypothesis

$H_0 : \sigma^2 = 4$

$H_1 : \sigma^2 \neq 4$

ii) Assumption: Samples are drawn randomly and independently from a normal population having the variance σ^2

Level of significance:

$\alpha = 5\% = 0.05$
Test-Statistic

$$\chi_c^2 = \frac{nS^2}{\sigma_0^2} = \frac{\sum\left(X_i - \overline{X}\right)^2}{\sigma_0^2}$$

Under H_0, Which has χ^2-distribution with V=n-1 degree of freedom
v) Critical region
If $H_1 : \sigma^2 \neq \sigma_0^2$ than we use two sided test

$\chi_c^2 < \chi^2_{(1-\frac{\alpha}{2})}(n-1)$ Or $\chi_c^2 > \chi^2_{\frac{\alpha}{2}}(n-1)$

$\chi_c^2 < \chi^2_{(1-\frac{0.05}{2})}(9-1)$ Or $\chi_c^2 > \chi^2_{\frac{0.05}{2}}(9-1)$

$\chi_c^2 < \chi^2_{0.975}(8)$ Or $\chi_c^2 > \chi^2_{0.025}(8)$

$\chi_c^2 < 2.18$ Or $\chi_c^2 > 17.54$

Calculation

X	$(X_i - \overline{X})$	$(X_i - \overline{X})^2$
5	-1	1
7	1	1
2	-4	16
4	2	4
8	2	4
9	3	9
8	2	4
6	0	0
5	-1	1
$\sum X = 54$	$\sum (X_i - \overline{X}) = 0$	$\sum (X_i - \overline{X})^2 = 40$

$$\overline{X} = \frac{\sum X}{n} = \frac{54}{9} = 6.0$$

$$\chi_c^2 = \frac{\sum\left(X_i - \overline{X}\right)^2}{\sigma^2} = \frac{40}{4} = 10.0$$

vii) Conclusion
Since our calculated value does not fall in critical region, then we accept H_0 and conclude that population variance is 4.

Q: 17.10(a): A sample of 25 observations has $S^2 = 12.6$, would you accept or reject at the 5 % level of significance the hypothesis that $\sigma^2 = 20$? Also compute a 90 % confidence interval for σ^2.
Solution: Do yourself similarly Q.17.9 (b)

Q.17.10(b): A standard examination has been given for several years with $\mu = 70$ and $\sigma^2 = 9$.A school using this examination for the first time, gave it to the group of 25 students who obtained a mean = 71 and a variance of $S^2 = 12$.Is there reason to doubt that the score of all students in the school would have a variance of 9?
Solution: Do yourself similarly Q.17.9 (b)

Q: 17.11(a): A random sample of 15 has the following values:
10.21, 9.72, 10.13, 8.89, 10.20, 9.65, 10.02, 10.00, 9.45, 10.11, 8.97, 10.21, 9.36, 9.55, 10.23.Test the hypothesis that $\sigma^2 = 0.12$ against $\sigma^2 > 0.12$ at (i) at 5 % and (ii) 1 % level of significance.
Solution:
i) We state our null and alternative hypothesis
a) $H_0 : \sigma^2 = 0.12$

$H_1 : \sigma^2 > 0.12$

ii) Assumption: Samples are drawn randomly and independently from a normal population having the variance σ^2
iii) Level of significance:
$\alpha = 5\% = 0.05$ $\alpha = 1\% = 0.01$
iv) Test-Statistic

$$\chi_c^2 = \frac{\sum\left(X_i - \overline{X}\right)^2}{\sigma^2}$$

Under H_0, Which has χ^2-distribution with V=n-1 degree of freedom

v) Critical region

If $H_1 : \sigma^2 > 0.12$ at $\alpha = 5\% = 0.05$ than we use one sided test

$\chi_c^2 > \chi_\alpha^2(n-1)$

a) $\chi_c^2 > \chi_{0.05}^2(15-1)$

$\chi_c^2 > \chi_{0.05}^2(14)$

$\chi_c^2 > 23.68$

b) $\alpha = 1\% = 0.01$

$\chi_c^2 > \chi_\alpha^2(n-1)$

$\chi_c^2 > \chi_{0.01}^2(15-1)$

$\chi_c^2 > \chi_{0.01}^2(14)$

$\chi_c^2 > 29.14$

vi) Calculation

X	$(X_i - \overline{X})$	$(X_i - \overline{X})^2$
10.21 9.72 10.13 8.89 10.20 9.65 10.02 10.0 9.45 10.11 8.97 10.21 9.36 9.55 10.23		
$\sum X = 54$	$\sum(X_i - \overline{X}) = 0$	$\sum(X_i - \overline{X})^2 = 2.8935$

$$\overline{X} = \frac{\sum X}{n} = 9.78$$

$$\chi_c^2 = \frac{\sum\left(X_i - \overline{X}\right)^2}{\sigma^2} = \frac{2.8935}{0.12} = 24.1125$$

vii) Conclusion

i) Since our calculated value fall's in critical region at 5% so we reject $H_0 : \sigma^2 = 0.12$

ii) Since our calculated value does not fall in critical region at 1% so we accept $H_0 : \sigma^2 = 0.12$

Q.17.11(b): A height distribution has a variance σ^2 of $(2.792)^2$ .Do the following 10 values ,selected at random have a greater variance than what is expected?67.50, 70.75, 72.00, 63.25, 65.25, 68.75, 69.25, 68.50, 66.50, and 64.75.

Solution: Do yourself similarly Q.17.11 (a)

Q:17.12(a): The weight of a random sample of 10 boxes of a particular brand of cereal are 14.2, 13.7, 14.1, 14.3, 14.1, 13.8, 14.4, 14.8, 13.9, and 14.3.Test the hypothesis that $H_o : \sigma^2 = 0.02$ against the alternative $H_1 : \sigma^2 < 0.02$,using a 0.01 level of significance.

Solution:

i) We state our null and alternative hypothesis

$H_0 : \sigma^2 = 0.02$

$H_1 : \sigma^2 < 0.02$

ii) Assumption: Samples are drawn randomly and independently from a normal population having the variance σ^2

iii) Level of significance:
$\alpha = 0.01$
iv) Test-Statistic
$$\chi_c^2 = \frac{\sum (X_i - \overline{X})^2}{\sigma^2}$$
Under H_0, Which has χ^2-distribution with V=n-1 degree of freedom
v) Critical region
If $H_1 : \sigma^2 < 0.02$ than we use one sided test
$\chi_c^2 < \chi_{1-\alpha}^2 (n-1)$
$\chi_c^2 < \chi_{1-0.01}^2 (10-1)$
$\chi_c^2 < \chi_{0.99}^2 (9)$
$\chi_c^2 < 2.09$
vi) Calculation

X	$(X_i - \overline{X})$	$(X_i - \overline{X})^2$
14.2 13.7 14.1 14.3 14.1 13.8 14.4 14.8 13.9 14.3		
$\sum X = 141.6$	$\sum (X_i - \overline{X}) = 0$	$\sum (X_i - \overline{X})^2 = 0.924$

$$\overline{X} = \frac{\sum X}{n} = 14.16$$
$$\chi_c^2 = \frac{\sum (X_i - \overline{X})^2}{\sigma^2} = \frac{0.924}{0.02} = 46.2$$
vii) Conclusion
Since our calculated value does not fall's in critical region. So, we accept $H_0 : \sigma^2 = 0.02$ at 1%.

Q.17.12 (b): A manufacturer of car batteries claims that the life of his batteries have a standard deviation equal to 0.9 years. If a random sample of 10 of these batteries have a standard deviation of 1.2 years, do you think that $\sigma > 0.9$ years? Use a 5 % level of significance.
Solution:
i) We state our null and alternative hypothesis
a) $H_0 : \sigma = 0.9 \quad Or \quad \sigma^2 = (0.9)^2$
$H_1 : \sigma > 0.9 \quad Or \quad \sigma^2 > (0.9)^2$
ii) Assumption: Samples are drawn randomly and independently from a normal population having the variance σ^2
iii) Level of significance:
$\alpha = 5\% = 0.05$
iv) Test-Statistic
$$\chi_c^2 = \frac{nS^2}{\sigma^2}$$
Under H_0, Which has χ^2-distribution with V=n-1 degree of freedom
v) Critical region
If $H_1 : \sigma^2 > 0.12$ at $\alpha = 5\% = 0.05$ than we use one sided test
$\chi_c^2 > \chi_{\alpha}^2 (n-1)$
$\chi_c^2 > \chi_{0.05}^2 (10-1)$
$\chi_c^2 > \chi_{0.05}^2 (9)$

$\chi_c^2 > 16.92$
vi) Calculation
Given that S=1.2 $S^2 = 1.44$ $\sigma = 0.9$ $\sigma^2 = 0.81$ n=10

$$\chi_c^2 = \frac{nS^2}{\sigma^2} = \frac{1.44}{0.81} = 1.78$$

vii) Conclusion
Since our calculated value does not fall in critical region at 5% so we accept
$H_0 : \sigma^2 = 0.9$

Q: 17.13(a): Describe how you would test the equality of k(k>2) variances of normal populations.
Ans: This test is used to test the hypothesis about the equality of several population variances when sample sizes are unequal.

Procedure:
i) We state our null and alternative hypothesis
$H_0 : \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = ... = \sigma_k^2 = \sigma^2$ *Or Population variances are homogeneous*
$H_1 :$ *At least two population variance are different Or Population variances are not homogeneous*
ii) Assumptions:
The samples are drawn randomly and independently from “k” approximately normal population with sample variances $s_1^2, s_2^2, s_3^2, ..., s_k^2$ are unbiased estimates of σ^2. When we give combine variance like polled estimate which is also unbiased estimate of population polled variance. From sample of size $n_1 + n_2 + n_3 + ... + n_k = \sum n_i = n$ from the population size $N_1, N_2, N_3, ..., N_k$ with degree of freedom (k-1), where “k” number of samples are used.
iii) Level of significance
$\alpha = Commonly\,use\ (5\%\ or\ 1\%)$
iv) Test-Statistic

$$u = 2.3026\frac{q}{c}$$

Under H0; which has χ^2 distribution with V=k-1 degree of freedom
Where $q = (\sum n_i - k)\log S_p^2 - \sum_{i=1}^{k}(n_i - 1)\log s_i^2$

$$c = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{n-k}\right]$$

$$s_i^2 = \frac{\sum_{i=1}^{n}(x_i - \bar{x})^2}{n-1} \qquad \text{And} \qquad s_p^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum_{i=1}^{k}(n_i - 1)} = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{n-k}$$

v) Critical region
It is naturally based on alternative hypothesis
$\chi_c^2 \geq \chi_\alpha^2 (V = k-1)$
vi) Calculation:
In this set up we calculate the value of test statistic on the basis of sample data.
We proceed as

Sample	s_i^2	$n_i - 1$	$\frac{1}{n_i - 1}$	$(n_i - 1)s_i^2$	$\log s_i^2$	$(n_i - 1)\log s_i^2$

vii) Conclusion
If our calculated value does not fall in critical region then we accept H0 otherwise we reject it.
Q.17.13(b): Show that the estimates 3.8, 4.4, 8.1, 6.1 and 9.4 of the population variance, based on 5, 8, 6, 7 and 4 d.f respectively. may be regarded as homogenous.
Solution:.
i) We state our null and alternative hypothesis

$H_0: \sigma_1^2 = \sigma_2^2 = \sigma_3^2 = ... = \sigma_k^2 = \sigma^2$ Or Population variances are homogeneous

H_1: At least two population variance are different Or Population variances are not homogeneous

ii) Assumptions:

The samples are drawn randomly and independently from "k" approximately normal population with sample variances $s_1^2, s_2^2, s_3^2, ..., s_k^2$ are unbiased estimates of σ^2. When we give combine variance like polled estimate which is also unbiased estimate of population polled variance. From sample of size $n_1 + n_2 + n_3 + ... + n_k = \sum n_i = n$ from the population size $N_1, N_2, N_3, ..., N_k$ with degree of freedom (k-1), where "k" number of samples are used.

iii) Level of significance

$\alpha = 5\% = 0.05$

vi)Test-Statistic

$$u = 2.3026\frac{q}{c}$$

Under H$_0$; which has χ^2 distribution with V=k-1 degree of freedom

Where $q = (\sum n_i - k)\log S_p^2 - \sum_{i=1}^{k}(n_i - 1)\log s_i^2$

$$c = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{n-k}\right]$$

v) Critical region:

$\chi_c^2 \geq \chi_\alpha^2 (V = k-1)$

$\chi_c^2 \geq \chi_{0.05}^2 (4)$

$\chi_c^2 \geq 9.49$

vi) Calculation:

$$u = 2.3026\frac{q}{c}$$

Where $q = (\sum n_i - k)\log S_p^2 - \sum_{i=1}^{k}(n_i - 1)\log s_i^2$

$$c = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{n-k}\right]$$

d.f=n-1	n$_i$	s_i^2	$n_i - 1$	$\frac{1}{n_i - 1}$	$(n_i - 1)s_i^2$	$\log s_i^2$	$(n_i - 1)\log s_i^2$
5 8 6 7 4	6 9 7 8 5	3.8 4.4 8.1 6.1 9.4	5 8 6 7 4	0.2 0.125 0.166 0.143 0.25	19 35.2 48.1 42.7 37.6	0.5797 0.6434 0.9085 0.7853 0.9731	2.8989 5.1476 5.451 5.4973 3.8925
	35		30	0.884	182.6	3.89	22.8873

$$s_p^2 = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{\sum_{i=1}^{k}(n_i - 1)} = \frac{\sum_{i=1}^{k}(n_i - 1)s_i^2}{n-k} = \frac{182.6}{30} = 6.087$$

$$q = (\sum n_i - k)\log S_p^2 - \sum_{i=1}^{k}(n_i - 1)\log s_i^2 = 30\log(6.087) - 22.8873 = 0.6441$$

$$c = 1 + \frac{1}{3(k-1)}\left[\sum_{i=1}^{k}\frac{1}{n_i - 1} - \frac{1}{n-k}\right] = 1 + \frac{1}{3(4)}\left[0.884 - \frac{1}{30}\right] = 1.071$$

$$u = 2.3026\frac{q}{c} = 2.3026\frac{0.6441}{1.071} = 1.3849$$

vii) Conclusion

Since our calculated value does not fall in critical region so, we accept H$_0$ and we conclude that population variances are homogeneous at 5%.

Q: 17.14(a): Describe Bartlett's test for homogeneity of variances.
Ans: Already explain
Q.17.14 (b): Three independent samples gave the following results:

Size	Observations
5 9 3	34, 40, 47, 60, 84 40, 59, 60, 67, 86, 92, 95, 98, 108 46, 93, 100

Use Bartlett's test to test the hypothesis of equal variances. Let α= 0.05.
Solution: Do yourself similarly Q.17.13(b)
Q: 17.15(a): Six samples of size 5 each have the variances 10.4, 13.8, 11.7, 19.3, 16.4 and 15.8.Test the hypothesis of homogeneity of variances by Bartlett's test.
Q.17.15 (b): For the data given below:
Sample 1: 4, 7, 6, 6
Sample 2: 5, 1, 3, 5, 3, 4
Sample 3: 3, 8, 6, 8, 9, 5
Use Bartlett's test to test the hypothesis that the variances of three populations are equal. (α = 0.05)
Solution: Do yourself similarly Q.17.13(b)
Q17.16: A random selection of nine individuals was made at each of 10 out patients' clinics across the country. Pulse rates were recorded and variances of 10 samples were 24, 31, 29, 28, 28, 26, 30, 31, 22, and 26.Test the hypothesis that variances are equal.
Solution: Do yourself similarly Q.17.13(b)

Testing procedure for multinomial distribution

Procedure

i) We stat our null and alternative hypothesis

$H_0 : P_1 = P_{10}, P_2 = P_{20}, P_3 = P_{30}, ..., P_k = P_{k0}$

$H_1 : P_i \neq P_{i0}$ $At\,least\ one\ value\ one\,value\,of\ i = 1,2,3,...,k$

Where $P_{10}, P_{20}, P_{30}, ..., P_{k0}$ are specified values

ii) Assumption:
Samples are drawn randomly and independently from a multinomial distribution which is approximately normal.

iii) Level of significance
$\alpha = Commonly\ used\ (5\%\ or\ 1\%)$

iv) Test-Statistic

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Under H_0, which has approximately Chi-Square distribution with V=k-1 degree of freedom

v) Calculation:
In this setup we calculate the value of test statistic on the basis of sample values.

vi) Critical region:
it is naturally based on alternative hypothesis
$\chi_c^2 \geq \chi_\alpha^2 (v = k - 1)$

vii) Conclusion:
If our calculated value does not fall in critical region then we accept H_0 otherwise we reject it.

Q: 17.18(a): Describe three distinct uses of chi-square distribution. According to a genetic model, the proportion in three groups should be p^2 : 2pq : q^2 ,where p + q = 1.Are the data consistent with the sample 9, 51, 45 if p = 0.4?
Solution:

i) We stat our null and alternative hypothesis

$H_0 : P_1 = P^2 = 0.16, P_2 = 2pq = 0.48, P_3 = q^2 = 0.36$

$H_1 : P_i \neq P_{i0}$ $At\,least\ one\ value\ one\,value\,of\ i = 1,2,3$

ii) Assumption:
Samples are drawn randomly and independently from a multinomial distribution which is approximately normal.

iii) Level of significance

$\alpha = 5\% = 0.05$

iv) Test-Statistic

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Under H_0, which has approximately Chi-Square distribution with V=k-1 degree of freedom

v) Calculation:

n_i	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
9 51 45	0.16 0.48 0.36	16.8 50.4 37.8	3.621 0.007 1.371
$\sum n = n = 105$	1.0	105	$\sum \frac{(n_i - np_i)^2}{np_i} = 4.999 = 5.0$

$$\chi_c^2 = \sum \frac{(n_i - np_i)^2}{np_i} = 4.999 = 5.0$$

vi) Critical region:

$\chi_c^2 \geq \chi_\alpha^2 (v = k-1)$

$\chi_c^2 \geq \chi_{0.05}^2 (v = 3-1 = 2)$

$\chi_c^2 \geq 5.89$

vii) Conclusion:

Since our calculated value does not fall in critical region so, we accept H_0 at 5% level of significance.

Q.17.18(b): The proportion of individuals possessing the four blood types should be in the proportion q2 :p2 + 2pq :r2 + 2qr : 2pr, where p+q+r +1.Given the observed frequencies 180, 360, 132, 98, test for compatibility with p = 0.4,q = 0.4 and r = 0.2.

Solution:

i) We stat our null and alternative hypothesis

$H_0 : P_1 = q^2 = 0.16, P_2 = p^2 + 2pq = 0.48 =, P_3 = r^2 + 2qr = 0.20, P_4 = 2pr = 0.16$

$H_1 : P_i \neq P_{i0}$ *At least one value one value of* $i = 1,2,3,4$

ii) Assumption:

Samples are drawn randomly and independently from a multinomial distribution which is approximately normal.

iii) Level of significance

$\alpha = 5\% = 0.05$

iv) Test-Statistic

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Under H_0, which has approximately Chi-Square distribution with V=k-1 degree of freedom

v) Calculation:

n_i	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
180 360 132 98	0.16 0.48 0.20 0.16	123.2 369.6 154 123.2	26.2 0.2 3.1 5.2
$\sum n = n = 770$	1.0	105	$\sum \frac{(n_i - np_i)^2}{np_i} = 34.7$

$$\chi_c^2 = \sum \frac{(n_i - np_i)^2}{np_i} = 34.7$$

vi) Critical region:

$\chi_c^2 \geq \chi_\alpha^2 (v = k-1)$

$\chi_c^2 \geq \chi_{0.05}^2 (v = 4 - 1 = 3)$

$\chi_c^2 \geq 7.82$

vii) Conclusion:

Since our calculated value fall in critical region, So we reject H_0 at 5% level of significance.

Q:17.19(a):Genetic theory, states that children having one parent of blood type M and the other of blood type N will always be of one of the three types M, MN, N and that proportion of these types will on the average be 1 : 2 : 1.A report states, of 162 children having one M parent and one N parent, 28.4 % were found to be of type M.42 % of type MN and the remainder of the type N.The low value of chi-square distribution demonstrates the truth of the genetic theory. Calculate value of chi-square distribution make the appropriate test of significance and comment on the conclusion quoted.

Solution:

i) We stat our null and alternative hypothesis

$H_0 : P_1 = M = \frac{1}{4}, P_2 = MN = \frac{2}{4}, P_3 = N = \frac{1}{4}$ $\quad$ *Therefore* $(1+2+1=4)$

$H_1 : P_i \neq P_{i0}$ *At least one value one value of* $i = 1,2,3$

ii) Assumption:

Samples are drawn randomly and independently from a multinomial distribution which is approximately normal.

iii) Level of significance

$\alpha = 5\% = 0.05$

iv) Test-Statistic

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Under H_0, which has approximately Chi-Square distribution with V=k-1 degree of freedom

v) Calculation:

Here M=28.4% MN=42% N=(100-28.4-42)=70.6%

n_i	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
$0.284 \times 162 = 46$ $0.42 \times 162 = 68$ $0.296 \times 162 = 48$	0.25 0.50 0.25	40.5 81 40.5	0.747 2.086 1.389
$\sum n = n = 162$	1.0	162	$\sum \frac{(n_i - np_i)^2}{np_i} = 4.22$

$$\chi_c^2 = \sum \frac{(n_i - np_i)^2}{np_i} = 4.22$$

vi) Critical region:

$\chi_c^2 \geq \chi_\alpha^2 (v = k - 1)$

$\chi_c^2 \geq \chi_{0.05}^2 (v = 3 - 1 = 2)$

$\chi_c^2 \geq 5.99$

vii) Conclusion:

Since our calculated value does not fall in critical region then we accept H_0 at the 5% level of significance.

Q.17.19 (b): A machine is supposed to mix peanuts, hazelnuts, cashews, and pecans in the ratio 5 : 2 : 2: 1.A can containing 500 of these mixed nuts was found to have 269 peanuts,112 hazelnuts, 74 cashews and 45 pecans. At the 0.05 level of significance, test the hypothesis that the machine is mixing the nuts in the supposed ratio.

Solution:

i) We stat our null and alternative hypothesis

$H_0 : P_1 = \frac{5}{10}, P_2 = \frac{2}{10}, P_3 = \frac{2}{10}, P_4 = \frac{1}{10}$ $\quad$ Therefore (5+2+2+1=10)

$H_1 : P_i \neq P_{i0}$ *At least one value one value of* $i = 1,2,3,4$

ii) Assumption:
Samples are drawn randomly and independently from a multinomial distribution which is approximately normal.
iii) Level of significance
$\alpha = 5\% = 0.05$
iv) Test-Statistic
$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$
Under H_0, which has approximately Chi-Square distribution with V=k-1 degree of freedom
v) Calculation:

n_i	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
269 112 74 45	0.5 0.2 0.2 0.1	250 100 100 50	1.44 1.44 6.76 0.50
$\sum n = n = 500$	1.0	500	$\sum \frac{(n_i - np_i)^2}{np_i} = 10.14$

$$\chi_c^2 = \sum \frac{(n_i - np_i)^2}{np_i} = 10.14$$
vi) Critical region:
$\chi_c^2 \geq \chi_\alpha^2 (v = k-1)$
$\chi_c^2 \geq \chi_{0.05}^2 (v = 4-1=3)$
$\chi_c^2 \geq 7.82$
vii) Conclusion:
Since our calculated valu fall's in critical region,So we reject H_0 at % level of significance.
Q.17.20 (a): A thousand individual were classified according to the sex and according to whether or not they were color- blind as follows:

Classes	Male	Female
Normal Colour Blind	452 38	494 16

According to the genetic theory, the frequencies in four classes should be

45%	50%
4%	1%

Test the hypothesis that the data are consistent with theory.
Solution:
i) We stat our null and alternative hypothesis
$H_0 : P_1 = 45\%, P_2 = 50\%, P_3 = 4\%, P_4 = 1\%$
$H_1 : P_i \neq P_{i0}$ $At\, least\;\; one\;\; value\;\; one\, value\, of\;\; i = 1,2,3,4$
ii) Assumption:
Samples are drawn randomly and independently from a multinomial distribution which is approximately normal.
iii) Level of significance
$\alpha = 5\% = 0.05$
iv) Test-Statistic
$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$
Under H_0, which has approximately Chi-Square distribution with V=k-1 degree of freedom

v) Calculation:

n_i	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
452 494 38 16	0.45 0.50 0.04 0.01	450 500 40 10	0.009 0.072 0.100 3.600
$\sum n = n = 1000$	1.0	1000	$\sum \frac{(n_i - np_i)^2}{np_i} = 3.78$

$$\chi_c^2 = \sum \frac{(n_i - np_i)^2}{np_i} = 3.78$$

vi) Critical region:

$\chi_c^2 \geq \chi_\alpha^2 (v = k-1)$

$\chi_c^2 \geq \chi_{0.05}^2 (v = 4-1 = 3)$

$\chi_c^2 \geq 7.82$

vii) Conclusion:

Sine our calculated value does not fall in critical region ,So we accept H_0 at 5% level of significance.

Q: 17.21(a): In 200 tosses of a coin, 115 heads and 85 tails were observed. Test the hypothesis that the coin is fair, using a level of significance of 0.05.

Solution:

i) We stat our null and alternative hypothesis

$H_0 : P_1(head) = \frac{1}{2}$ $\qquad$ $P_2(tail) = \frac{1}{2}$

$H_1 : P_i \neq P_{i0}$ $\ At\, least\ one\ value\ one\, value\, of\ i = 1,2$

ii) Assumption:

Samples are drawn randomly and independently from a multinomial distribution which is approximately normal.

iii) Level of significance

$\alpha = 5\% = 0.05$

iv) Test-Statistic

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Under H_0, which has approximately Chi-Square distribution with V=k-1 degree of freedom

Calculation:

n_i	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
115 85	0.5 0.5	100 100	2.25 2.25
$\sum n = n = 200$	1.0	200	$\sum \frac{(n_i - np_i)^2}{np_i} = 4.50$

$$\chi_c^2 = \sum \frac{(n_i - np_i)^2}{np_i} = 4.50$$

vi) Critical region:

$\chi_c^2 \geq \chi_\alpha^2 (v = k-1)$

$\chi_c^2 \geq \chi_{0.05}^2 (v = 2-1 = 1)$

$\chi_c^2 \geq 3.84$

vii) Conclusion:

Since our calculated value does not fall in critical region, So we accept H_0 at 5% level of significance.

Q.17.21 (b): In a 360 tosses of a pair of dice, 74 sevens and 24 elevens are observed. Using a 5 % level of significance, test the hypothesis that the dice is fair.

Solution:

i) We stat our null and alternative hypothesis

$H_0: P_1(total\ of\ \ seven) = \frac{6}{36} = \frac{1}{6}, P_2(total\ of\ \ eleven) = \frac{2}{36} = \frac{1}{18}; P_3(remaining) = \frac{28}{36}$

$H_1: P_i \neq P_{i0}\ \ At\,least\ \ one\ \ value\ \ one\,value\,of\ \ i = 1,2,3$

ii) Assumption:

Samples are drawn randomly and independently from a multinomial distribution which is approximately normal.

iii) Level of significance

$\alpha = 5\% = 0.05$

iv) Test-Statistic

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Under H_0, which has approximately Chi-Square distribution with V=k-1 degree of freedom

v) Calculation:

n_i	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
74 24 262	1/6=0.167 1/18=0.056 0.778	360(1/6)=60 360(1/18)=20 280	3.267 0.80 1.157
$\sum n = n = 360$		360	$\sum \frac{(n_i - np_i)^2}{np_i} = 5.22$

$$\chi_c^2 = \sum \frac{(n_i - np_i)^2}{np_i} = 5.22$$

vi) Critical region:

$\chi_c^2 \geq \chi_\alpha^2 (v = k - 1)$

$\chi_c^2 \geq \chi_{0.05}^2 (v = 3 - 1 = 2)$

$\chi_c^2 \geq 5.99$

vii) Conclusion:

Sine our calculated value does not fall in critical region, So we accept H_0 at 5% level of significance and we conclude that the die is fair.

Q: 17.22(a): The sex distribution of 98 births reported in a newspaper was 52 boys and 46 girls. Is this consistent with an equal sex division in the population? Use the Chi-Square approximation and the normal approximation.

Solution:

i) We stat our null and alternative hypothesis

$H_0: P_1 = \frac{1}{2}, P_2 = \frac{1}{2} \qquad It\ \ is\ \ consistent\ \ equal\ \ sex\ \ division$

$H_1: P_i \neq P_{i0}\ \ At\,least\ \ one\ \ value\ \ one\,value\,of\ \ \ or\ \ \ It\ \ is\ \ not\ \ consistent\ \ equal\ \ sex\ \ division\ \ \ i = 1,2$

ii) Assumption:

Samples are drawn randomly and independently from a multinomial distribution which is approximately normal.

iii) Level of significance

$\alpha = 0.05$

iv) Test-Statistic

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Under H_0, which has approximately Chi-Square distribution with V=k-1 d.f

v)Calculation:

n_i	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
52 46	½ ½	49 49	0.184 0.189

$\sum n = n = 105$	1.0	105	$\sum \frac{(n_i - np_i)^2}{np_i} = 0.3637 = 0.40$

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i} = 0.3673 = 0.40$$

vi) Critical region:

$\chi_c^2 \geq \chi_\alpha^2 (v = k - 1)$

$\chi_c^2 \geq \chi_{0.05}^2 (v = 2 - 1 = 1)$

$\chi_c^2 \geq 3.84$

vii) Conclusion:

Since our calculated value does not fall in critical region. So, we accept H_0 and we conclude that there is equal sex division at 5%.

ii) By normal approximation

i) We stat our null and alternative hypothesis

$H_0 : P = \frac{1}{2}$ *There is equal sex division*

$H_1 : P \neq \frac{1}{2}$ *There is no equal sex division*

ii) Assumption:

Samples are drawn randomly and independently from a multinomial distribution which is approximately normal.

iii) Level of significance

$\alpha = 5\% = 0.05$

iv) Test-Statistic

$$Z_c = \frac{\hat{p} - P}{\sqrt{\frac{pq}{n}}}$$

Under H_0, it has approximately Z-distribution

v) Calculation:

$$Z_c = \frac{\hat{p} - P}{\sqrt{\frac{pq}{n}}}$$

n=98 X= no. of boys=52 $\hat{P} = \frac{x}{n} = \frac{52}{98} = 0.53$ $\sqrt{\frac{pq}{n}} = \sqrt{\frac{(0.50)(0.50)}{98}} = 0.051$

$$Z_c = \frac{\hat{p} - P}{\sqrt{\frac{pq}{n}}} = \frac{0.53 - 0.50}{0.051} = 0.588 = 0.60$$

vi) Critical region:

$|Z_c| \geq Z_{\frac{\alpha}{2}}$

$|Z_c| \geq 1.96$

vii) Conclusion:

Since our calculated value does not fall in critical region. So, we accept H_0 and we conclude that there is equal sex division at 5% level of significance.

Q.17.22 (b): In a certain disease with 40 % mortality, of 10 patients given a certain treatment only one dies. Is the treatment effective at 5% level of significance?

Solution:

i) We stat our null and alternative hypothesis

$H_0 : P_1(mortality) = 40\% = 0.40, P_2(alive) = 60\% = 0.60$

$H_1 : P_i \neq P_{i0}$ *At least one value one value of* $i = 1,2$

ii) Assumption:

Samples are drawn randomly and independently from a multinomial distribution which is approximately normal.

iii) Level of significance

$\alpha = 0.05$

vi) Test-Statistic

$$\chi_c^2 = \sum \frac{\left(|n_i - np_i| - \frac{1}{2}\right)^2}{np_i}$$

Under H_0, which has approximately Chi-Square distribution with V=k-1 degree of freedom

v) Calculation:

n_i	p_i	np_i	$\frac{\left(\lvert n_i - np_i\rvert - \frac{1}{2}\right)^2}{np_i}$
1 9	0.40 0.60	4.0 6.0	1.5625 11.0417
$\sum n = n = 10$	1.0	10	$\sum \frac{\left(\lvert n_i - np_i\rvert - \frac{1}{2}\right)^2}{np_i} = 2.60$

$$\chi_c^2 = \sum \frac{\left(|n_i - np_i| - \frac{1}{2}\right)^2}{np_i} = 2.60$$

vi) Critical region:

$\chi_c^2 \geq \chi_\alpha^2 (v = k - 1)$

$\chi_c^2 \geq \chi_{0.05}^2 (v = 2 - 1 = 1)$

$\chi_c^2 \geq 3.84$

vii) Conclusion:

Since our calculated value does not fall in critical region, So we accept H_0 at 5% level of significance.

Q:17.23(a):The following table records the observed number of births at a hospital in 4 consecutive quarterly periods.

Quarter	Jan-Mar,	Apr-Jun,	Jul-Sep	Oct-Dec
Number of births	110	57	53	80

It is hypothesized that twice as many babies are born during the Jan-Mar. quarter than are born in any of the other three quarters. At α = 0.10,test if these data strongly contradict the stated hypothesis.

Solution:

i) We stat our null and alternative hypothesis

$$H_0 : P_1 = 2P = \frac{2}{5}, P_2 = \frac{1}{5}, P_3 = \frac{1}{5}, P_4 = \frac{1}{5} \qquad Therefore \qquad (2:1:1:1)$$

$H_1 : P_i \neq P_{i0}$ *At least one value one value of* $i = 1,2,3,4$

ii) Assumption: Samples are drawn randomly and independently from a multinomial distribution which is approximately normal.

iii) Level of significance $\alpha = 10\% = 0.10$

iv) Test-Statistic $\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$

Under H_0, which has approximately Chi-Square distribution with V=k-1 d.f

Calculation:

n_i	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
110 57 53 80	2/5 1/5 1/5 1/5	120 60 60 60	0.833 0.150 0.8167 6.667
$\sum n = n = 300$	1.0	300	$\sum \frac{(n_i - np_i)^2}{np_i} = 8.5$

$$\chi_c^2 = \sum \frac{(n_i - np_i)^2}{np_i} = 8.5$$

vi) Critical region:

$\chi_c^2 \geq \chi_\alpha^2 (v = k-1)$

$\chi_c^2 \geq \chi_{0.10}^2 (v = 4-1=3)$

$\chi_c^2 \geq 6.25$

vii) Conclusion:

Since our calculated value fall's in critical region, so we reject H_0 at 10% level of significance.

Q.17.23(b): The grade in a statistics course were as follows:

Grade	A	B	C	D	E
F	14	18	32	20	16

Test the hypothesis, at the 0.05 level of significance, that the distribution of grade.

Solution:

i) We stat our null and alternative hypothesis

$H_0 : P_1 = \frac{1}{5}, P_2 = \frac{1}{5}, P_3 = \frac{1}{5}, ..., P_4 = \frac{1}{5}, P_5 = \frac{1}{5}$ *Therefore* (1:1:1:1:1)

$H_1 : P_i \neq P_{i0}$ *At least one value one value of* $i = 1,2,3,4,5$

ii) Assumption:

Samples are drawn randomly and independently from a multinomial distribution which is approximately normal.

iii) Level of significance

$\alpha = 5\% = 0.05$

iv) Test-Statistic

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Under H_0, which has approximately Chi-Square distribution with V=k-1 degree of freedom

v)Calculation:

n_i	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
14 18 32 20 16	1/5 1/5 1/5 1/5 1/5	20 20 20 20 20	1.8 0.2 7.2 0 0.8
$\sum n = n = 100$	1.0	100	$\sum \frac{(n_i - np_i)^2}{np_i} = 10.0$

$$\chi_c^2 = \sum \frac{(n_i - np_i)^2}{np_i} = 10.0$$

vi) Critical region:

$\chi_c^2 \geq \chi_\alpha^2 (v = k-1)$

$\chi_c^2 \geq \chi_{0.05}^2 (v = 5-1=4)$

$\chi_c^2 \geq 9.49$

vii) Conclusion:

Since our calculated value fall's in critical region,so we reject H_0 at 5% level of significance and conclude that the distribution of grade not uniform.

Q: 17.24(a): A random number table of 250 digits showed the following distribution of the digits 0, 1,…,9.

Digits:	0	1	2	3	4	5	6	7	8	9
Frequency:	17	31	29	18	14	20	35	30	20	36

Test the hypothesis, at 0.05 level of significance, that the digits were distributed in equal numbers in the table.

Solution:
i) We stat our null and alternative hypothesis

$H_0: P_1 = \frac{1}{10}, P_2 = \frac{1}{10}, P_3 = \frac{1}{10}, ..., P_{10} = \frac{1}{10}$ $\quad$ *Therefore* $\quad$ (1:1:1:1:1:1:1:1:1:1)

$H_1: P_i \neq P_{i0}$ *At least one value one value of* $i = 1,2,3,...,10$

ii) Assumption:
Samples are drawn randomly and independently from a multinomial distribution which is approximately normal.

iii) Level of significance

$\alpha = Commonly\ used\ (5\%\ or\ 1\%)$

iv) Test-Statistic

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Under H_0, which has approximately Chi-Square distribution with V=k-1 degree of freedom

v)Calculation:

n_i	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
17 31 29 18 14 20 35 30 20 36	1/10 1/10 1/10 1/10 1/10 1/10 1/10 1/10 1/10 1/10	25 25 25 25 25 25 25 25 25 25	2.56 1.44 0.64 1.96 4.84 1.0 4.0 1 1 4.84
$\sum n = n = 250$	1.0	250	$\sum \frac{(n_i - np_i)^2}{np_i} = 23.28$

$$\chi_c^2 = \sum \frac{(n_i - np_i)^2}{np_i} = 23.28$$

vi) Critical region:

$\chi_c^2 \geq \chi_\alpha^2 (v = k-1)$

$\chi_c^2 \geq \chi_{0.05}^2 (v = 10 - 1 = 9)$

$\chi_c^2 \geq 16.92$

vii) Conclusion:
Since our calculated value fall's in critical region, so we reject H_0 at 5% level of significance and conclude that the digits are not distributed with equal number.

Q.17.24 (b): The following distribution shows the number of deaths from overdoses of narcotics. Use Chi-Square statistic to test the hypothesis that equal number die in all age group.

Age	15-19	20-24	25-29	30-34	35-39	40-44	45-49
No. of deaths	40	35	32	10	13	13	4

i) We stat our null and alternative hypothesis

$H_0: P_1 = \frac{1}{7}, P_2 = \frac{1}{7}, P_3 = \frac{1}{7}, ..., P_7 = \frac{1}{7}$ $\quad$ *Therefore* $\quad$ (1:1:1:1:1:1:1)

$H_1: P_i \neq P_{i0}$ *At least one value one value of* $i = 1,2,3,...,7$

ii) Assumption:
Samples are drawn randomly and independently from a multinomial distribution which is approximately normal.

iii) Level of significance

$\alpha = 0.05$

iv) Test-Statistic

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Under H_0, which has approximately Chi-Square distribution with V=k-1 degree of freedom

v) Calculation:

n_i	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
40 35 32 10 13 13 4	1/7 1/7 1/7 1/7 1/7 1/7 1/7	21 21 21 21 21 21 21	17.19 7.84 5.76 5.76 3.05 3.05 13.76
$\sum n = n = 147$	1.0	105	$\sum \frac{(n_i - np_i)^2}{np_i} = 56.41$

$$\chi_c^2 = \sum \frac{(n_i - np_i)^2}{np_i} = 56.41$$

vi) Critical region:

$\chi_c^2 \geq \chi_\alpha^2 (v = k-1)$

$\chi_c^2 \geq \chi_{0.05}^2 (v = 7-1 = 6)$

$\chi_c^2 \geq 12.59$

vii) Conclusion:

Since our calculated value fall's in critical region, so we reject H_0 at 5% level of significance and conclude that the number of die not equal in each group..

Q: 17.25(a): A die is tossed 180 times with the following results:

X	1	2	3	4	5	6
F	20	36	46	35	21	22

Is this a balance die? Use α = 0.01.

Solution:

i) We stat our null and alternative hypothesis

$$H_0 : P_1 = \frac{1}{6}, P_2 = \frac{1}{6}, P_3 = \frac{1}{6}, \ldots, P_6 = \frac{1}{6}$$

$$H_1 : P_i \neq P_{i0} \ \ \text{At least one value one value of } i = 1,2,3,\ldots,6$$

Where $P_{10}, P_{20}, P_{30}, \ldots, P_{k0}$ are specified values

ii) Assumption:

Samples are drawn randomly and independently from a multinomial distribution which is approximately normal.

iii) Level of significance $\alpha = 0.01$

iv) Test-Statistic

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Under H_0, which has approximately Chi-Square distribution with V=k-1 degree of freedom

v) Calculation:

n_i	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
20 36 46 35 21 22	1/6 1/6 1/6 1/6 1/6 1/6	30 30 30 30 30 30	3.33 0.36 8.53 0.83 2.70 2.13
$\sum n = n = 180$	1.0	180	$\sum \frac{(n_i - np_i)^2}{np_i} = 17.88$

$$\chi_c^2 = \sum \frac{(n_i - np_i)^2}{np_i} = 17.88$$

vi) Critical region:

$\chi_c^2 \geq \chi_\alpha^2 (v = k-1)$

$\chi_c^2 \geq \chi_{0.01}^2 (v = 6-1=5)$

$\chi_c^2 \geq 15.09$

vii) Conclusion:

Since our calculated value fall's in critical region, so we reject H_0 at 1% level of significance and conclude that the die is not balanced.

Q.17.25 (b): The following figure shows the number of births in an area over a year by months of occurrence.

January	50759	May	51371	September	52162
February	46472	June	47388	October	50824
March	51419	July	49995	November	47768
April	49670	August	51043	December	51129

Use the Chi-Square test to discuss whether there is any seasonality in births revealed by these data.

Solution:

i) We stat our null and alternative hypothesis

H_0 : *There is no seasonality in the births revealed by these data*

H_1 : *There is seasonality in the births revealed by these data*

ii) Assumption:

Samples are drawn randomly and independently from a multinomial distribution which is approximately normal.

iii) Level of significance

$\alpha = 5\% = 0.05$

iv) Test-Statistic

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Under H_0, which has approximately Chi-Square distribution with V=k-1 degree of freedom

v)Calculation:

n_i	p_i	np_i	$\frac{(n_i - np_i)^2}{np_i}$
50759	31/365	50959	0.78
46472	28/365	46027	4.30
51419	31/365	50959	4.15
49670	30/365	49315	2.56
51371	31/365	50959	3.33
47388	30/365	49315	75.30
49995	31/365	50959	18.24
51043	31/365	50959	0.14
52162	30/365	49315	164.36
50824	31/365	50959	0.36
47768	30/365	4931	48.53
51129	31/365	50959	0.57
n=600000	1.0		$\sum \frac{(n_i - np_i)^2}{np_i} = 322.62$

$$\chi_c^2 = \sum \frac{(n_i - np_i)^2}{np_i} = 322.62$$

vi) Critical region:

$\chi_c^2 \geq \chi_\alpha^2 (v = k-1)$

$\chi_c^2 \geq \chi_{0.05}^2 (v = 12-1=11)$

$\chi_c^2 \geq 19.68$

vii) Conclusion:

Since our calculated value fall's in critical region, So we reject H_0 and we conclude that *There is seasonality in the births revealed by these data* at 5%.

Q: 17.26(a): Discuss the Chi-Square test of goodness of fit. What are the assumptions in the application of these tests to practical problems?

Procedure:

i) We stat our null and alternative hypothesis

H_0 :*The data fit as good*

H_1 :*The data fit not good*

ii) Assumption:

Samples are drawn randomly and independently from a approximately normal.

iii) Level of significance

$\alpha = Commonly\ used\ (5\%\ or\ 1\%)$

iv) Test-Statistic

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Under H_0, which has approximately Chi-Square distribution with V=k-1-M (no. of parameter estimated) degree of freedom

v) Calculation:

In this setup we calculate the value of test statistic on the basis of sample information

vi) Critical region:

it is naturally based on alternative hypothesis

$\chi_c^2 \geq \chi_\alpha^2 (v = k - 1 - M)$

vii) Conclusion:

If our calculated value does not fall in critical region. Than, we accept H_0 and we conclude that the data fit as good at the given level of significance.

Q.17.26 (b): Records taken of the number of male and female births is 800 families having four children, are as follows:

No.of male births	0	1	2	3	4
Families	32	178	290	236	64

Test whether the data is consistent with the hypothesis that the binomial law holds and that the chance of a male birth is equal to that of a female birth,that is p = q = ½.

Solution: do yourself similarly Q.17.27(b)

Q:17.27(b): Three six sided dice were thrown 648 times and the number of 5's or 6's noted at each throw.

No.of 5's or 6's	0	1	2	3
No. of throw	179	298	141	30

Test the hypothesis that the data conform to binomial distribution with p=1/3 and n=3.Let $\alpha = 0.05$.

Solution:

i) We stat our null and alternative hypothesis

H_0 :*The data fit as binomial distribution with* $n = 3$ *and* $P = \frac{1}{3}$

H_1 :*The data not fit as binomial distribution with* $n = 3$ *and* $P = \frac{2}{3}$

ii) Assumption:

Samples are drawn randomly and independently from a approximately normal.

iii) Level of significance

$\alpha = 5\% = 0.05$

iv) Test-Statistic

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Under H_0, which has approximately Chi-Square distribution with V=k-1-M (no. of parameter estimated) degree of freedom

v)Calculation:

X	f	$P(x)=n_{C_x}(1/3)^x(2/3)^{n-x}$	$e=nP$	$\frac{(O_i-e_u)^2}{e_i}$
0 1 2 3	179 298 141 30	8/27 12/27 6/27 1/27	192 288 144 24	0.880 0.3472 0.0625 1.5
	648			$\sum\frac{(O_i-e_u)^2}{e_i}=2.79$

$$\chi_c^2=\sum\frac{(O_i-e_u)^2}{e_i}=2.79$$

vi) Critical region:

$\chi_c^2 \geq \chi_\alpha^2(v=k-1-M)$

$\chi_c^2 \geq \chi_{0.05}^2(4-1-0=3)$

$\chi_c^2 \geq 7.82$

vii) Conclusion:

Since our calculated value does not fall in critical region. So, we accept H_0 and we conclude that the data fit as good follow binomial distribution at 5% level of significance.

Q:17.28Twelve dice were thrown 4096 times and a throw a six was reckoned as a success.The observed frequencies were as given below:

No.of success	0	1	2	3	4	5	6	7&over	total
Frequency	447	1145	1181	796	380	115	24	8	4096

Find the value of Chi-Square on the hypothesis that the dice were unbiased and hence show that the data are consistent with the hypothesis so far as the Chi-Square test is concerned.

Solution:

i) We stat our null and alternative hypothesis

$$H_0: The\ \ data\ \ fit\ \ as\, binomial\ distribution\ with\, n=12\ \ and\ P=\frac{1}{6}$$

$$H_1: The\ \ data\ \ not\ \ fit\ \ as\, binomial\ distribution\ with\, n=12\ \ and\ P=\frac{1}{6}$$

ii) Assumption:

Samples are drawn randomly and independently from a approximately normal.

iii) Level of significance

$\alpha=1\%=0.01$

iv) Test-Statistic

$$\chi_c^2=\sum_{i=1}^{k}\frac{(O_i-e_i)^2}{e_i}=\sum_{i=1}^{k}\frac{(n_i-np_i)^2}{np_i}$$

Under H_0, which has approximately Chi-Square distribution with V=k-1-M (no. of parameter estimated) degree of freedom

v)Calculation:

X	f	$P(x)=n_{C_x}(1/6)^x(5/6)^{n-x}$	$e=nP$	$\frac{(O_i-e_u)^2}{e_i}$
0 1 2 3 4 5 6 7 And Over	447 1145 1181 796 380 115 24 8	0.1122 0.2692 0.2961 0.1974 0.0888 0.0284 0.0066 0.0011	459.4 1102.5 1212.8 808.5 363.8 116.4 27.2 5.0	0.3347 1.6383 0.8338 0.1932 0.7214 0.0168 0.3765 1.8
	4096			$\sum\frac{(O_i-e_u)^2}{e_i}=5.91$

$$\chi_c^2=\sum\frac{(O_i-e_u)^2}{e_i}=5.91$$

vi) Critical region:

$\chi_c^2 \geq \chi_\alpha^2 (v = k - 1 - M)$

$\chi_c^2 \geq \chi_{0.01}^2 (8 - 1 - 0 = 7)$

$\chi_c^2 \geq 18.48$

vii) Conclusion:

Since our calculated value does not fall in critical region. So, we accept H_0 and we conclude that the data fit as good follow binomial distribution at 1% level of significance.

Q:17.29(a): Suppose that 6 coins are tossed simultaneously 640 times and the following frequencies distribution is observed:

No.of heads	0	1	2	3	4	5	6
Frequency	13	70	137	210	145	56	9

Test the null hypothesis that the coins are well-balanced. Use α = 0.01.

Solution:

i) We stat our null and alternative hypothesis

H_0 : *The data fit as binomial distribution with* $n = 6$ *and* $P = \frac{1}{2}$

H_1 : *The data not fit as binomial distribution with* $n = 6$ *and* $P = \frac{1}{2}$

ii) Assumption:

Samples are drawn randomly and independently from a approximately normal.

iii) Level of significance

$\alpha = 1\% = 0.01$

iv) Test-Statistic

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Under H_0, which has approximately Chi-Square distribution with V=k-1-M (no. of parameter estimated) degree of freedom

v)Calculation:

X	f	$P(x) = n_{C_x}(1/2)^x (1/2)^{n-x}$	$e = nP$	$\frac{(O_i - e_u)^2}{e_i}$
0	13	1/64	10	0.9
1	70	6/64	60	1.667
2	137	15/64	150	1.127
3	210	20/64	200	0.500
4	145	15/64	150	0.167
5	56	6/64	60	0.267
6	9	1/64	10	0.1
	640			$\sum \frac{(O_i - e_u)^2}{e_i} = 4.73$

$$\chi_c^2 = \sum \frac{(O_i - e_u)^2}{e_i} = 4.73$$

vi) Critical region:

$\chi_c^2 \geq \chi_\alpha^2 (v = k - 1 - M)$

$\chi_c^2 \geq \chi_{0.01}^2 (7 - 1 - 0 = 6)$

$\chi_c^2 \geq 16.81$

vii) Conclusion:

Since our calculated value does not fall in critical region. So, we accept H_0 and we conclude that the data fit as good follow binomial distribution at 1% level of significance.

Q.17.29 (b): When the first proof of a book containing 250 pages was read, the following distribution of printing mistakes were found:

No.of mistakes per page	0	1	2	3	4	5
Frequency	139	76	28	4	2	1

Fit an appropriate distribution to the data and test the goodness of fit.

Solution:

i) We stat our null and alternative hypothesis

H_0 : *The data fit as good*

H_1 : *The data not fit as good*

ii) Assumption:

Samples are drawn randomly and independently from a approximately normal.

iii) Level of significance $\alpha = 5\% = 0.05$

iv) Test-Statistic

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Under H_0, which has approximately Chi-Square distribution with V=k-1-M (no. of parameter estimated) degree of freedom

v)Calculation:

X	f	fx	fx^2
0	139	0	0
1	76	76	76
2	28	56	112
3	4	12	36
4	2	8	32
5	1	5	25
	250	157	281

$$\overline{X} = \frac{\sum fx}{\sum f} = \frac{157}{250} = 0.628 = 1.0$$

$$S^2 = \frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2 = \frac{281}{250} - \left(\frac{157}{250}\right)^2 = 0.729 = 1.0$$

From the above calculation mean and variance approximately equal. So, we use Poisson distribution fit as appropriate

X	f	$P(x) = \frac{e^{-0.628}(0.628)^x}{x!}$	$e = nP$	$\frac{(O_i - e_u)^2}{e_i}$
0	139	0.5336	133.4	0.2351
1	76	0.3351	83.78	0.7225
2	28	0.1052	26.3	0.1099
3 4 5	$\begin{bmatrix}4\\2\\1\end{bmatrix} = 7$	0.0220 0.0035 0.0004	$\begin{bmatrix}5.5\\0.88\\0.1\end{bmatrix} = 6.48$	0.0417
	250			$\sum \frac{(O_i - e_u)^2}{e_i} = 1.109$

$$\chi_c^2 = \sum \frac{(O_i - e_u)^2}{e_i} = 1.109$$

vi) Critical region:

$$\chi_c^2 \geq \chi_\alpha^2 (v = k - 1 - M)$$

$$\chi_c^2 \geq \chi_{0.05}^2 (v = 4 - 1 - 1) = 2$$

$$\chi_c^2 \geq 5.99$$

vii) Conclusion:

Since our calculated value does not fall in critical region. So, we accept H_0 and we conclude that the data fit as good and follow Poisson distribution at the 5% level of significance.

Q:17.30: Given the data

X	0	1	2	3	4	5
F	40	32	18	8	2	0

Fit a Poisson distribution and test the goodness of fit.

Solution: Do yourself similarly Q.17.31

Q:17.31: Test whether the data given below may be regarded as conforming to a Poisson distribution?

X	0	1	2	3	4	5	6	7
F	305	365	210	80	28	9	2	1

Solution:
i) We stat our null and alternative hypothesis

H_0 : *The data fit as good*

H_1 : *The data not fit as good*

ii) Assumption:
Samples are drawn randomly and independently from a approximately normal.

iii) Level of significance

$\alpha = 5\% = 0.05$

iv) Test-Statistic

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Under H_0, which has approximately Chi-Square distribution with V=k-1-M (no. of parameter estimated) degree of freedom

v)Calculation:

X	f	fx	$P(x) = \frac{e^{-\mu}\mu^x}{x!}$	$e = nP$	$\frac{(O_i - e_u)^2}{e_i}$
0	305	0	0.3012	301.2	0.0479
1	365	365	0.3614	361.4	0.0359
2	210	420	0.2169	216.9	0.2195
3	80	240	0.0867	86.7	0.5177
4	28	112	0.0260	26	0.1538
5	$\left[\begin{matrix}9\\2\\1\end{matrix}\right]12$	45	0.0062	$\left[\begin{matrix}6.2\\1.2\\0.2\end{matrix}\right]7.6$	2.5474
6		12	0.0012		
7		7	0.0002		
	1000	1201			$\sum \frac{(O_i - e_u)^2}{e_i} = 3.522$

$$\mu = \frac{\sum fx}{\sum f} = \frac{1201}{1000} = 1.201 = 1.2$$

$$\chi_c^2 = \sum \frac{(O_i - e_u)^2}{e_i} = 3.522$$

vi) Critical region:

$\chi_c^2 \geq \chi_\alpha^2 (v = k - 1 - M)$

$\chi_c^2 \geq \chi_{0.05}^2 (v = 6 - 1 - 1) = 4$

$\chi_c^2 \geq 9.49$

vii) Conclusion:
Since our calculated value does not fall in critical region. So, we accept H_0 and we conclude that the data fit as good and follow Poisson distribution at the 5% level of significance.

Q: 17.32: The wages of 1,000 employees range from Rs.4.50 to Rs.19.50.They are grouped in 15 classes with a common class interval of Re.1 and the class frequencies, from the lowest class to the highest, are 6, 17, 35, 48, 65, 90, 131, 173, 155, 117, 75, 52, 21, 9, 6.Fit a normal distribution and apply the Chi-Square goodness of fit test.

Solution:
i) We stat our null and alternative hypothesis

H_0 : *The data fit as good*

H_1 : *The data not fit as good*

ii) Assumption:
Samples are drawn randomly and independently from a approximately normal.

iii) Level of significance

$\alpha = 5\% = 0.05$

iv) Test-Statistic

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Under H_0, which has approximately Chi-Square distribution with V=k-1-M (no. of parameter estimated) degree of freedom

v)Calculation:
First we find mean and standard deviation

Classes	f	X	fx	fx^2
4.5-5.5	6			
5.5-6.5	17			
6.5-7.5	35			
7.5-8.5	48			
8.5-9.5	65			
9.5-10.5	90			
10.5-11.5	131			
11.5-12.5	173			
12.5-13.5	155			
13.5-14.5	117			
14.5-15.5	75			
15.5-16.5	52			
16.5-17.5	21			
17.5-18.5	9			
18.5-19.5	6			

$\overline{X} = \dfrac{\sum fx}{\sum f} = 12.0$ $\qquad S = \sqrt{\dfrac{\sum fx^2}{\sum f} - \left(\dfrac{\sum fx}{\sum f}\right)^2} = 2.6$

U.c.b	$f = O_i$	$Z = \dfrac{uc.b - \overline{X}}{S}$	$P(Z < z)$	$\hat{P}(de-cumulative)$	$e = n\hat{P}$	$\dfrac{(O_i - e_u)^2}{e_i}$
5.5	6	-2.50	0.0062	0.0062	6.2	0.006
6.5	17	-2.12	0.0170	0.0108	10.8	3.559
7.5	35	-1.73	0.0418	0.0248	24.8	4.195
8.5	48	-1.35	0.0885	0.0467	46.7	0.036
9.5	65	-0.96	0.1685	0.0800	80.0	2.813
10.5	90	-58	0.2810	0.1125	112.5	4.500
11.5	131	-0.19	0.4247	0.1437	143.7	1.122
12.5	173	0.19	0.5753	0.1490	149.0	3.866
13.5	155	0.58	0.7190	0.1453	145.3	0.648
14.5	117	0.96	0.8315	0.1125	112.5	0.180
15.5	75	1.35	0.9115	0.0800	80.0	0.313
16.5	52	1.73	0.9582	0.0467	46.7	0.601
17.5	21	2.12	0.9830	0.0248	24.8	0.582
18.5	9	2.50	0.9938	0.0108	10.8	0.300
19.5	6	∞	1.000	0.0062	6.2	0.006
	1000				1000	$\sum \dfrac{(O_i - e_u)^2}{e_i} = 22.73$

$$\chi_c^2 = \sum \frac{(O_i - e_u)^2}{e_i} = 22.73$$

vi) Critical region:
it is naturally based on alternative hypothesis

$\chi_c^2 \geq \chi_\alpha^2 (v = k - 1 - M)$

$\chi_c^2 \geq \chi_{0.05}^2 (v = 15 - 1 - 2 = 12)$

$\chi_c^2 \geq 21.03$

vii) Conclusion:
Since our calculated value fall's in critical region. So, we reject H_0 and we conclude that the data not fit as good to the normal distribution at the 5% level of significance.

Q: 17.33: The height of 200 employees are distributed as follows:

Heights	58-60	61-63	64-66	67-69	70-72	73-75	76-78
Frequency	9	20	45	55	43	17	11

Test whether the normal distribution given a satisfactory fit to the data at α = 0.05.
Solution:

i) We stat our null and alternative hypothesis

$H_0: The\ \ data\ \ fit\ \ as\ good$

$H_1: The\ \ data\ \ not\ fit\ \ as\ good$

ii) Assumption:

Samples are drawn randomly and independently from a approximately normal.

iii) Level of significance

$\alpha = 5\% = 0.05$

iv) Test-Statistic

$$\chi_c^2 = \sum_{i=1}^{k} \frac{(O_i - e_i)^2}{e_i} = \sum_{i=1}^{k} \frac{(n_i - np_i)^2}{np_i}$$

Under H_0, which has approximately Chi-Square distribution with V=k-1-M (no. of parameter estimated) degree of freedom

v) Calculation:

First we find mean and standard deviation

Height	f	c.b	x	fx	fx^2
58-60	9	57.5-60.5	59	531	31329
61-63	20	60.5-63.5	62	1240	76880
64-66	45	63.5-66.5	65	2925	190125
67-69	55	66.5-69.5	68	3740	254320
70-72	43	69.5-72.5	71	3053	216763
73-75	17	72.5-75.5	74	1258	93092
76-78	11	75.5-78.5	77	847	65219
	200			13594	927728

$$\overline{X} = \frac{\sum fx}{\sum f} = \frac{13594}{200} = 67.97$$

$$S = \sqrt{\frac{\sum fx^2}{\sum f} - \left(\frac{\sum fx}{\sum f}\right)^2} = \sqrt{\frac{927728}{200} - \left(\frac{13594}{200}\right)^2} = 4.33$$

u.c.b	$f = O_i$	$Z = \frac{uc.b - \overline{X}}{S}$	$P(Z < z)$	$\hat{P}(de-cumulative)$	$e = n\hat{P}$	$\frac{(O_i - e_u)^2}{e_i}$
60.5	9	-1.73	0.042	0.042	8.4	0.043
63.5	20	-1.03	0.152	0.11	22	0.182
66.5	45	-0.34	0.367	0.215	43	0.093
69.5	55	0.35	0.637	0.270	54	0.019
72.5	43	1.05	0.853	0.216	43.2	0.001
75.5	17	1.74	0.959	0.106	21.2	0.833
78.5	11	∞	1.000	0.041	8.2	0.956
	200				200	$\sum \frac{(O_i - e_u)^2}{e_i} = 2.12$

$$\chi_c^2 = \sum \frac{(O_i - e_u)^2}{e_i} = 2.12$$

vi) Critical region:

it is naturally based on alternative hypothesis

$\chi_c^2 \geq \chi_\alpha^2 (v = k - 1 - M)$

$\chi_c^2 \geq \chi_\alpha^2 (v = 7 - 1 - 2 = 4)$

$\chi_c^2 \geq 9.49$

vii) Conclusion:

Since our calculated value does not fall in critical region. So, we accept H_0 and we conclude that the data fit as good and consisting with normal distribution at 5% level of significance.

The 2×2 Contingency table is given as

	A	α	Total
B	(AB)	(αB)	$(B) = (AB) + (\alpha B)$
β	$(A\beta)$	$(\alpha\beta)$	$(\beta) = (A\beta) + (\alpha\beta)$
Total	$(A) = (AB) + (A\beta)$	$(\alpha) = (\alpha B) + (\alpha\beta)$	$n = (B) + (\beta) = (A) + (\alpha)$

Example: Compute all remaining possible frequencies from the following data

$(\alpha)=50$ $\quad(\beta)=70$ $\quad(A\beta)=20$ $\quad n=100$

Solution:

	A	α	Total
B	(AB)	(αB)	$70=(AB)+(\alpha B)$
β	$(A\beta)=20$	$(\alpha\beta)$	$(\beta)=(A\beta)+(\alpha\beta)$
Total	$(A)=(AB)+20$	$\alpha=50$	$100=(B)+(\beta)=(A)+(\alpha)$

Now we find out the required frequencies

$(A)=?$ $\quad(\beta)=?$ $\quad(AB)=?$ $\quad(\alpha\beta)=?$ $\quad(\alpha B)=?$

As we know that

$n=(A)+(\alpha)$

$100=(A)+50$

$(A)=100-50=50$

$(A)=(AB)+(A\beta)$

$50=(AB)+20$

$(AB)=50-20=30$

$(B)=(AB)+(\alpha B)$

$70=30+(\alpha B)$

$(\alpha B)=70-30=40$

$(\alpha)=(\alpha B)+(\alpha\beta)$

$50=40+(\alpha\beta)$

$(\alpha\beta)=50-40=10$

$(\beta)=(A\beta)+(\alpha\beta)$

$(\beta)=20+10=30$

Example: Given that

$(AB)=150$ $\quad(A\beta)=250$ $\quad(\alpha B)=260$ $\quad(\alpha\beta)=2340$

Find the other frequencies and value of n.

$n=?$ $\quad(\alpha)=?$ $\quad(\beta)=?$ $\quad(A)=?$ $\quad(B)=?$

Solution: Now we find out the required frequencies

$n=?$ $\quad(\alpha)=?$ $\quad(\beta)=?$ $\quad(A)=?$ $\quad(B)=?$

$(\alpha)=(\alpha B)+(\alpha\beta)$

$(\alpha)=260+2340=2600$

$(\beta)=250+2340=2590$

$(A)=(AB)+(A\beta)$

$(A)=150+250=400$

$n=(A)+(\alpha)=400+2600=3000$

$(B)=(AB)+(\alpha B)=150+260=410$

Example: Given that

$(A)=304$ $\quad(AB)=256$ $\quad(A\beta)=48$ $\quad(\alpha B)=768$ $\quad(\alpha\beta)=144$

Two attributes “A and B” are said to be independent, if

$$(AB)=\frac{(A)(B)}{n}$$

First we find (B) and n

$(B)=(AB)+(\alpha B)=256+768=1024$

$(\beta)=48+144=192$

$n=(B)+(\beta)=1024+192=1216$

Now

$$(AB)=\frac{(A)(B)}{n}$$

$$256=\frac{(304\times 1024)}{1216}=256$$

Hence the two attributes “A and B” are independent,

Example: Whether attributes “A and B” are negatively associated, positively associated or independent.

i) $(A)=\frac{340}{17}$ $\quad(B)=\frac{13}{50}$ $\quad(AB)=\frac{1}{50}$ $\quad n=250$

ii) $(A)=\frac{340}{17}$ $\quad(\beta)=88$ $\quad(AB)+(A\beta)=35$ $\quad(A\beta)=20$ $\quad n=154$

iii) $(A)=34000$ $\quad(\alpha)=36000$ $\quad(\alpha B)=700$ $\quad(AB)=5300$

Solution:

i) $(A)=\frac{340}{17}$ $\quad(B)=\frac{13}{50}$ $\quad(AB)=\frac{1}{50}$ $\quad n=250$

$(A)(B)=\left(\frac{340}{17}\right)\left(\frac{13}{50}\right)=5.2$ $\quad(AB)=\frac{1}{50}=0.02$

$\left(AB\right) = \frac{(A)(B)}{n} = \frac{5.2}{250} = 0.021$

$\left(AB\right) < \frac{(A)(B)}{n}$ So, A and B are negatively associated

ii) $(A) = \frac{340}{17}$ $\quad(\beta) = 88$ $\quad(AB) + (A\beta) = 35$ $\quad(A\beta) = 20$ $\quad n = 154$

(A)= $(AB) + (A\beta) = 35$

$(AB) + (A\beta) = (A)$

$(AB) = 35 - (A\beta) = 35 - 20 = 15$

$(B) + (\beta) = n$

$(B) = n - (\beta) = 154 - 88 = 66$

$\frac{(A)(B)}{n} = \frac{(35 \times 66)}{154} = 15$

$\left(AB\right) = \frac{(A)(B)}{n}$ So, A and B are independent

iii) $(A) = 34000$ $\quad(\alpha) = 36000$ $\quad(\alpha B) = 700$ $\quad(AB) = 5300$

$(A) + (\alpha) = n$

$n = 34000 + 36000 = 70000$

$(AB) + (\alpha B) = (B)$

$(B) = 5300 + 700 = 6000$

$\frac{(A)(B)}{n} = \frac{(34000) \times (6000)}{70000} = 2914.29$

$\left(AB\right) > \frac{(A)(B)}{n}$ So, A and B are positively associated

Example: Test the independence by a simplest approach between gender and intelligence

Gender

Level of intelligent	Males	Females	Total
Intelligent	150	75	225
Non-intelligent	50	25	75
Total	200	100	300

Let "A" represent male and "B" represents the intelligent, then

$(A) = 200$ $\quad(B) = 225$ $\quad(AB) = 150$ $\quad n = 300$

$\frac{(A)(B)}{n} = \frac{(200) \times (225)}{300} = 150$

$\left(AB\right) = \frac{(A)(B)}{n}$ So, A and B are independent and we conclude that there is no relationship between Gender and Intelligent. It means that males and female equally intelligent or non-intelligent.

Procedure for independence of attributes in $(r \times c)$ contingency table

i) Set up our null and alternative hypothesis

H_0 *:The tattributes are independent Or There is no association*

H_1 *:The tattributes are dependent Or There is association*

ii) Assumption: The samples of attributes are random independent and drawn from approximately normal population of qualitative nature.

iii) Level of significance

$\alpha = Commonly\ used\ 5\%\ or\ 1\%$

iv) Test-statistic

$$\chi_c^{\ 2} = \sum \frac{\left(f_0 - f_e\right)^2}{f_e} = \sum \frac{\left(O_{ij} - e_{ij}\right)^2}{e_{ij}}$$

Under H_0 ,It has chi-square distribution with $V = (r-1)(c-1)$ degree of freedom

v) Critical region

it is naturally based on alternative hypothesis

$\chi_c^{\ 2} \geq \chi_\alpha^{\ 2}(v)$

vi) Calculation:
In this step we calculate the value of test statistic on the basis of sample information
vii) Conclusion:
If our calculated value does not fall in critical region, so we accept H_0 and we conclude that the attributes are independent at given level of significance.
Q: 17.41(b): The following table shows the number of recruits taking (i) A preliminary And (ii) a final test in car driving. Use a Chi-Square test to discuss Whether there is any association between the results of preliminary and those of the Final test.

Categories		Preliminary Pass	Preliminary Fail	Total
Final	Pass	605	195	740
	Fail	135	65	260
Total		800	200	1000

Solution:
i) Set up our null and alternative hypothesis
H_0 :*The tattributes are independent Or There is no association*
H_1 :*The tattributes are dependent Or There is association*
ii) Assumption: The samples of attributes are random independent and drawn from approximately normal population of qualitative nature.
iii) Level of significance
$\alpha = 0.05$
iv) Test-statistic
$$\chi_c^2 = \sum \frac{(f_0 - f_e)^2}{f_e} = \sum \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$
Under H_0 ,It has chi-square distribution with $V = (r-1)(c-1)$ degree of freedom
v) Critical region
$\chi_c^2 > \chi_\alpha^2(v)$
$\chi_c^2 > \chi_{0.05}^2(1)$
$\chi_c^2 > 3.84$ $\qquad v = (r-1)(c-1) = (2-1)(2-1) = 1$
vi) Calculation: First we calculate expected frequencies e

Categories		Preliminary Pass	Preliminary Fail	Total
Final	Pass	$\frac{800\times 740}{1000} = 592$	$\frac{200\times 740}{1000} = 148$	740
	Fail	$\frac{800\times 260}{1000} = 208$	$\frac{200\times 260}{1000} = 52$	260
Total		800	200	1000

Now further calculation

O_{ij}	e_{ij}	$(O_{ij} - e_{ij})$	$(O_{ij} - e_{ij})^2 / e_{ij}$
605	592	13	0.285
135	148	-13	1.142
195	208	-13	0.8125
65	52	13	3.250
$\sum O_{ij} = 1000$	$\sum e_{ij} = 1000$	$\sum (O_{ij} - e_{ij}) = 0$	$\sum (O_{ij} - e_{ij})^2 / e_{ij} = 5.489$

$$\chi_c^2 = \sum (O_{ij} - e_{ij})^2 / e_{ij} = 5.489$$
vii) Conclusion:
Since our calculated value fall’s in critical region, so we reject H_0 and we conclude that Preliminary and final results are associated at $\alpha = 0.05$.

Example: Find coefficient of association from the following data

Height of fathers

Height of sons	Tall	Short	Total
Tall	500	100	600
Short	100	400	500
Total	600	500	1100

Solution:

Let A denote the tall fathers and α denote the short fathers and B denote the tall sons and β denote the short fathers

Height of sons	A	α	Total
B	500=(AB)	100= (αB)	600=(B)
β	100= $(A\beta)$	400= $(\alpha\beta)$	500= (β)
Total	600=(A)	500= (α)	1100=n

$$Q = \frac{(AB)(\alpha\beta)-(A\beta)(\alpha B)}{(AB)(\alpha\beta)+(A\beta)(\alpha B)} = \frac{(500))(400)-(100)(100)}{(500))(400)+(100)(100)} = 0.905 = 0.91$$

It means that there is positive association between tall fathers and tall sons. Thus tall father have tall sons.

Q: 17.42 (a): Test the association between injection against typhoid and exemption from attack from the following contingency table:

Classes	Attacked	Not attacked	Total
Inoculated	528	25	553
Not inoculated	790	175	965
Total	1318	200	1518

Solution: Do yourself similarly Q.17.41(b)

Q.17.42(b): The following table gives the census data of orchards. Test the hypothesis that the two variables of classification are independent.

Classes	Shaded	Unshaded	Total
High Yielders	350	205	555
Low Yielders	250	195	445
Total	600	400	1000

Solution: Do yourself similarly Q.17.41(b)

Q: 43: Find Chi Square and test whether the two attributes are independent .Let $\alpha = 0.05$.

Attributes	A1	A2	A3	Total
B1	215	325	60	600
B2	135	175	90	400
Total	350	500	150	1000

Solution: Do yourself similarly Q.17.41(b)

Q.17.44(a):Test the null hypothesis that the two variables of classification are independent, using a 0.05Level of significance.

Classes	A_1	A_2	A_3
B1	337	291	302
B2	225	207	238

Solution: Do yourself similarly Q.17.41(b)

Q.17.44(b): The following is percentage distribution by income level and ownership of a random sample of 400 families in the city of Lahore.

	Income Level Less then Rs. 12,000	 Rs.12,000 to Rs.60,000	 More than Rs.60,000
Home Owner	5%	35%	10%
Renter	15%	25%	10%

Test the hypothesis that the home ownership is independent of the family income level using 1% level of significance.

Solution:

i) Set up our null and alternative hypothesis

H_0 :The tattributes are independent Or There is no association

H_1 :The tattributes are dependent Or There is association

ii) Assumption: The samples of attributes are random independent and drawn from approximately normal population of qualitative nature.

iii) Level of significance

$\alpha = 0.01$

iv) Test-statistic

$$\chi_c^2 = \sum \frac{(f_0 - f_e)^2}{f_e} = \sum \frac{(O_{ij} - e_{ij})^2}{e_{ij}}$$

Under H_0 ,It has chi-square distribution with $V = (r-1)(c-1)$ degree of freedom

v) Critical region

$\chi_c^2 > \chi_\alpha^2(v)$

$\chi_c^2 > \chi_{0.05}^2(2)$

$\chi_c^2 > 5.99$ $\quad v = (r-1)(c-1) = (2-1)(3-1) = 2$

vi) Calculation: First we find actual frequencies

	Income Level Less then Rs. 12,000	Rs.12,000 to Rs.60,000	More than Rs.60,000
Home Owner	$\frac{5}{100} \times 400 = 20$	$\frac{35}{100} \times 400 = 140$	$\frac{10}{100} \times 400 = 40$
Renter	$\frac{15}{100} \times 400 = 60$	$\frac{25}{100} \times 400 = 100$	$\frac{10}{100} \times 400 = 40$

Now we calculate expected frequencies e

	Income Level Less then Rs. 12,000	Rs.12,000 to Rs.60,000	More than Rs.60,000	Total
Home Owner	$\frac{80 \times 200}{400} = 40$	$\frac{240 \times 200}{400} = 120$	$\frac{80 \times 200}{400} = 40$	200
Renter	$\frac{80 \times 200}{400} = 40$	$\frac{200 \times 240}{400} = 120$	$\frac{80 \times 200}{400} = 40$	200
Total	80	240	80	400

Now further calculation

O_{ij}	e_{ij}	$(O_{ij} - e_{ij})$	$(O_{ij} - e_{ij})^2 / e_{ij}$
20	40	-20	10
140	120	20	3.33
40	40	0	0
60	40	20	10
100	120	-20	3.33
40	40	0	0
$\sum O_{ij} = 400$	$\sum e_{ij} = 400$	$\sum (O_{ij} - e_{ij}) = 0$	$\sum (O_{ij} - e_{ij})^2 / e_{ij} = 26.66$

$$\chi_c^2 = \sum (O_{ij} - e_{ij})^2 / e_{ij} = 26.66$$

vii) Conclusion:
Since our calculated value fall's in critical region, so we reject H_0 and we conclude that the attributes are associated at $\alpha = 0.01$.
Q: 17.45: A certain drug is claimed to be effective in curing colds. In an experiment on 164 people with colds, half of them were given the drug and half of them were given sugar pills. The Patients' reaction to the treatment is recorded in the following table. Test the hypothesis that the drug is no better than sugar pills for curing colds. Let α = 0.05.

Category	Helped	Harmed	No Effect
Drug	52	10	20
Sugar	44	12	26

Solution: Do yourself similarly Q.17.41(b)
Q: 17.46: A thousand households are taken at random and divided into three groups A' and C, according to the total monthly income. The following table shows the numbers in each Group having a color television receiver, a black and white receiver, or no television at all.

	A	B	C
Color television	56	51	93
black and white	118	207	375
None	26	42	32

Solution: Do yourself similarly Q.17.41(b)
Q: 17.47: Calculate Chi-Square from the following contingency table of attributes and test for Independence at α = 0.01

Attributes	A1	A2	A3
B1	44	22	4
B2	265	257	178
B3	41	91	98

Solution: Do yourself similarly Q.17.41(b)
Q: 17.48: Gilby classified 1725 school children according to intelligence and apparent family Economic level. A condensed classification fallows:

Classes	Dull	Intelligent	Very Capable
Very well clothed	81	322	233
Well clothed	141	457	153
Poorly clothed	127	163	48

Test the null hypothesis of independence of the two classifications at the 0.01 level of significance.
Solution: Do yourself similarly Q.17.41(b)
Q: 17.49: An insurance company wants to determine whether a policy holder's age is independent of whether or not the policy holder has filled an accident claim. A study of 1000 of its Policy holders gave the following results:

Age	Under 25	25-40	40-55	Over 55
Reported claim	93	72	53	63
No claim	115	155	265	184

Solution: Do yourself similarly Q.17.41(b)
Q: 17.50: A random sample of 200 married men, all retired, was classified according to education and number of children.

Education	Number of children		
	0-1	2-3	Over 3
Elementary	14	37	32
Secondary	19	42	17
College	12	17	10

Test the hypothesis, at the 0.05 level of significance, that the size of a family is independent of the level of education attained by the father.
Solution: Do yourself similarly Q.17.41(b)

Q: 17.51: Show that a chi-square test for a 2×2 contingency table is equivalent to testing the difference between the two proportions, using the normal approximation.
Proof: Let we have given the following 2×2 contingency table

	B_1	B_2	Total
A_1	A	b	$r_1 = a+b$
A_2	C	d	$r_2 = c+d$
Total	$n_1 = a+c$	$n_2 = b+d$	$n = n_1 + n_2 = r_1 + r_2$

So, we get

$n_1 = a+c \qquad n_2 = b+d \qquad r_1 = a+b \qquad r_2 = c+d \qquad n = a+b+c+d$

Suppose the two sample proportion are denoted by p_1, p_2 and overall estimate is denoted by "p"

$p_1 = \frac{a}{n_1}$ And $a = n_1 p_1$

$p_2 = \frac{b}{n_2}$ And $b = n_2 p_2$

$q_1 = \frac{c}{n_1}$ And $c = n_1 q_1$

$q_2 = \frac{d}{n_2}$ And $d = n_2 q_2$

$p = \frac{r_1}{n}$ And $r_1 = np$

$q = \frac{r_2}{n}$ And $r_2 = nq$

Now we substitute these values in 2×2 contingency table formula of χ^2 approximation we have

$$\chi_c^2 = \frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

$$\chi_c^2 = \frac{n(n_1 p_1 n_2 q_2 - n_2 p_2 n_1 q_1)^2}{n_1 n_2 r_1 r_2}$$

$$\chi_c^2 = \frac{n n_1^2 n_2^2 (p_1 q_2 - p_2 q_1)^2}{n_1 n_2 npnq}$$

$$\chi_c^2 = \frac{n_1 n_2 (p_1(1-p_2) - p_2(1-p_1))^2}{npq} \qquad \text{Therefore} \qquad q = 1-p$$

$$\chi_c^2 = \frac{(p_1 - p_1 p_2 - p_2 + p_2 p_1)^2}{\left(\frac{n}{n_1 n_2}\right) pq} \qquad \text{therefore} \qquad n = n_1 + n_2$$

$$\chi_c^2 = \frac{(p_1 - p_2)^2}{\left(\frac{n_1 + n_2}{n_1 n_2}\right) pq}$$

$$\chi_c^2 = \frac{(p_1 - p_2)^2}{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) pq}$$

$$\chi_c^2 = Z^2$$

By taking the square root

$$Z = \sqrt{\frac{(p_1 - p_2)^2}{\left(\frac{1}{n_1} + \frac{1}{n_2}\right) pq}}$$

$$Z=\frac{(p_1-p_2)}{\sqrt{\left(\frac{1}{n_1}+\frac{1}{n_2}\right)pq}}$$

Hence we calculate that the testing of association for 2×2 contingency table by χ^2 approximation is equalent to testing of two proportions by normal approximation.

Q: 17.52: Given the following contingency table for hair color and eye color, calculate the co-efficient of contingency and interpret the result.

Eye Color	Hair Color		
	Fair	Grey	Brown
Blue	69	49	28
Black	91	56	27
Dark Blue	57	24	33

Solution: Do yourself similarly Q.17.41(b)

Q.17.53 (a) Explain the use of Yates correction for continuity

Ans: **Yates correction for continuity**

In applying χ^2 approximation we are require to combine the smaller frequencies (less than 5) with larger once. But in case of two classes only we cannot pooled the smaller frequency into larger one for such a situation Frank Yates in 1934 showed that the χ^2 approximation is markedly improved if we use the following formula

$$\chi_c^2=\frac{n\left(|ad-bc|-\frac{1}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)}$$

This adjustment is known as Yates correction for contingency. It should be used only when there is one degree of freedom and one frequency is small. So, in 2×2 contingency with smaller frequencies the cell frequencies are adjusted by adding the ½ to the smaller and subtracting by the larger and keeping the marginal total unchanged with adjustment the formula for χ^2 becomes

$$\chi_c^2=\frac{n\left(|ad-bc|-\frac{n}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)}$$ If ad < bc Where n=a+b+c+d

This correction should be used if any expected frequency in 2×2 is less than 10

Q: 17.53(b): Out of a group of 320 people exposed to infection, 255 had not been immunized, and of these 95 contracted the disease. Of those who had been immunized, 15 were infected. Does it seem that treatment gave any protection against infection?

Solution:

i) Set up our null and alternative hypothesis

H_0 :*The tattributes are independent Or There is no association*

H_1 :*The tattributes are dependent Or There is association*

ii) Assumption: The samples of attributes are random independent and drawn from approximately normal population of qualitative nature.

iii) Level of significance

$\alpha=0.05$

iv) Test-statistic

$$\chi_c^2=\sum\frac{(f_0-f_e)^2}{f_e}=\sum\frac{(O_{ij}-e_{ij})^2}{e_{ij}}$$

Under H_0 ,It has chi-square distribution with $V=(r-1)(c-1)$ degree of freedom

v) Critical region

$\chi_c^2>\chi_\alpha^2(v)$

$\chi_c^2>\chi_{0.05}^2(1)$

$\chi_c^2>3.84$ $\qquad v=(r-1)(c-1)=(2-1)(2-1)=1$

vi) Calculation:

Infection

Categories	Infected	Not infected	Total
Immunized	15	50	65
Not immunized	95	110	255
Total	110	210	320

First we calculate expected frequencies e

Infection

Categories	Infected	Not infected	Total
Immunized	$\frac{110\times 65}{320}=22.34$	$\frac{210\times 65}{320}=42.66$	65
Not immunized	$\frac{110\times 255}{320}=87.66$	$\frac{210\times 255}{320}=167.34$	255
Total	110	210	320

Now further calculation

O_{ij}	e_{ij}	$(O_{ij}-e_{ij})^2 / e_{ij}$
15 95 50 160	22.34 87.66 42.66 167.34	2.41 0.62 1.26 0.32
$\sum O_{ij}=1000$	$\sum e_{ij}=1000$	$\sum (O_{ij}-e_{ij})^2 / e_{ij}=4.61$

$$\chi_c^{2}=\sum (O_{ij}-e_{ij})^2 / e_{ij}=4.61$$

vii) Conclusion:

Since our calculated value fall's in critical region, so we reject H_0 and we conclude that infection and vaccination are associated at $\alpha=0.05$.

What is the difference in the significance of the result of the Chi-square test according as Yates? Correction is or is not applied?

Q: 17.54 a) Prove that Chi-square for the table

$a+\frac{1}{2}$	$b-\frac{1}{2}$
$c-\frac{1}{2}$	$d-\frac{1}{2}$

$$\chi_c^2=\frac{n\left(|ad-bc|-\frac{n}{2}\right)^2}{(a+b)(c+d)(a+c)(b+d)} \qquad \text{When} \qquad ad<bc$$

If ad<bc and where n = a+b+c+d.

Proof: Let by the two 2×2 modify contingency table is

	B_1	B_2	Total
A_1	$a+\frac{1}{2}$	$b-\frac{1}{2}$	$a+b$
A_2	$c-\frac{1}{2}$	$d-\frac{1}{2}$	$c+d$
Total	$a+c$	$b+d$	$n=a+b+c+d$

As the shortcut methods is the simplest method of calculating χ^2 statistic with out using expected frequencies for 2×2 contingency table is

$$\chi_c^2=\frac{n(ad-bc)^2}{(a+b)(c+d)(a+c)(b+d)}$$

By substituting the modified observed frequencies we get

$$\chi_c^2 = \frac{n\left[\left(a+\frac{1}{2}\right)\left(d+\frac{1}{2}\right)-\left(b-\frac{1}{2}\right)\left(c-\frac{1}{2}\right)\right]^2}{(a+b)(c+d)(a+c)(b+d)}$$

$$\chi_c^2 = \frac{n\left[\left(ad+\frac{a}{2}+\frac{d}{2}+\frac{1}{4}\right)-\left(bc-\frac{b}{2}-\frac{c}{2}+\frac{1}{4}\right)\right]^2}{(a+b)(c+d)(a+c)(b+d)}$$

$$\chi_c^2 = \frac{n\left[ad+\frac{a}{2}+\frac{d}{2}+\frac{1}{4}-bc+\frac{b}{2}+\frac{c}{2}-\frac{1}{4}\right]^2}{(a+b)(c+d)(a+c)(b+d)}$$

$$\chi_c^2 = \frac{n\left[ad+\frac{a}{2}+\frac{d}{2}-bc+\frac{b}{2}+\frac{c}{2}\right]^2}{(a+b)(c+d)(a+c)(b+d)}$$

$$\chi_c^2 = \frac{n\left[ad-bc+\frac{a}{2}+\frac{d}{2}+\frac{b}{2}+\frac{c}{2}\right]^2}{(a+b)(c+d)(a+c)(b+d)}$$

$$\chi_c^2 = \frac{n\left[ad-bc+\frac{1}{2}(a+b+d+c\right]^2}{(a+b)(c+d)(a+c)(b+d)}$$

$$\chi_c^2 = \frac{n\left[ad-bc+\frac{1}{2}(n)\right]^2}{(a+b)(c+d)(a+c)(b+d)}$$

And if $ad<bc$ then

$$\chi_c^2 = \frac{n\left[|ad-bc|-\frac{n}{2}\right]^2}{(a+b)(c+d)(a+c)(b+d)}$$

Hence proved

Q.17.54(b): A random sample of 30 adults is classified according to sex and the number of hours they watch television during a week.

Classes	Male	Female
Over 25 hours	5	9
Under 25 hours	9	7

Using a 0.01 level of significance, test the hypothesis that a person's sex and time watching television are independent.

Solution:

i) Set up our null and alternative hypothesis

H_0 :*The tattributes are independent Or There is no association*

H_1 :*The tattributes are dependent Or There is association*

ii) Assumption: The samples of attributes are random independent and drawn from approximately normal population of qualitative nature.

iii) Level of significance

$\alpha = 0.05$

iv) Test-statistic

$$\chi_c^2 = \frac{n\left[|ad-bc|-\frac{n}{2}\right]^2}{(a+b)(c+d)(a+c)(b+d)}$$

Under H_0 ,It has chi-square distribution with $V=(r-1)(c-1)$ degree of freedom

Because here cell frequencies are less than 10, So we use Yates correction.

v) Critical region

$\chi_c{}^2 > \chi_\alpha{}^2(v)$

$\chi_c{}^2 > \chi_{0.05}{}^2(1)$

$\chi_c{}^2 > 3.84$ $\qquad v=(r-1)(c-1)=(2-1)(2-1)=1$

vi) Calculation:

Classes	Male	Female	Total
Over 25 hours	5=a	9=b	14=(a+b)
Under 25 hours	9=c	7=d	16=(c+d)
Total	(a+c)=14	(b+d)=16	N=30

$$\chi_c^2 = \frac{n\left[|ad-bc|-\frac{n}{2}\right]^2}{(a+b)(c+d)(a+c)(b+d)}$$

$$\chi_c^2 = \frac{30\left[|35-81|-\frac{30}{2}\right]^2}{(14)(16)(14)(16)} = 0.5745$$

vii) Conclusion:

Since our calculated value does not fall in critical region, so we accept H_0 and we conclude that Sex and watching television are not associated at $\alpha = 0.05$.

Q.17.55 (a): Describe the Fisher's exact test for a 2×2 contingency table

Ans: When the frequencies in 2×2 contingency table are fairly small, there will be some doubt about the adequately of the Chi-Square approximation. An exact test called Fisher exact test or Fisher Irwin exact test was proposed. Instead of comparing the observed and expected cell frequencies, the test is based on calculating the exact probabilities of cell frequencies for all possible 2×2 tables obtained by varying the smallest cell frequency from observed value to zero and with marginal frequencies fixed.

It was shown that under the null hypothesis of no association between rows and columns classification the exact probability of observing a table with frequencies

Attributes	B_1	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

Where all the marginal frequencies are fixed is given by

$$P = \frac{(a+b)!(c+d)!(a+c)!(b+b)!}{a!b!c!d!n!}$$

Supposing that "d" is the smallest frequency we obtain other 2×2 tables by decreasing cell frequencies and counting the process until "d" becomes zero we calculate the probabilities of observed table and other possible tables for all values of "d" from observed values down to zero.

Then total probabilities i.e. $P = P_0 + P_1 + P_2 + ... + P_d$ corresponding to one tail of the distribution and is comparable with half the probability calculated from χ_c^2 thus for a two sided test, we double the probability so obtained i.e. $\chi_c^2 = 2P$. If probability 2P is not negligible, we reject our null hypothesis.

Procedure:

i) Set up our null and alternative hypothesis

H_0 :The classification of two variable are independent

H_1 :The classification of two variable are dependent

ii) Assumption: Samples are drawn randomly and independently from a approximately normal population of qualitative nature.

iii) Level of significance

$\alpha =$ *Commonly used* (5% *or* 1%)

iv) Test-statistic

$\chi_c^2 = 2P$ Where $P_i = \frac{(a+b)!(c+d)!(a+c)!(b+b)!}{a!b!c!d!n!}$

And $P = P_0 + P_1 + P_2 + ... + P_d$

Under H_0, it has chi-square distribution

v) Calculation:

In this setup we calculate the value of test statistic on the basis of sample information

vi) Critical region

it is naturally based on alternative hypothesis

$2P \geq \alpha$

vii) Conclusion:

If our calculated value does not fall in critical region, Then we accept H_0 and conclude that the classification of two variables are independent.

Q: 17.55 (b): Suppose that a number of patients were treated for cancer with results as in the following table:

Toxity present	Tumer Regression	
	Yes	No
Yes	5	2
No	1	7

Use Fisher exact test to test the independence.

Solution:

i) Set up our null and alternative hypothesis

H_0 :*The classification of two variable are independent*

H_1 :*The classification of two variable are dependent*

ii) Assumption: Samples are drawn randomly and independently from a approximately normal population of qualitative nature.

iii) Level of significance

$\alpha = 5\% = 0.05$

iv) Test-statistic

$\chi_c^2 = 2P$ Where $P_i = \frac{(a+b)!(c+d)!(a+c)!(b+b)!}{a!b!c!d!n!}$

And $P = P_0 + P_1 + P_2 + \ldots + P_d$

Under H_0, it has chi-square distribution

v) Calculation:

1st table

Toxity present	Tumer Regression		
	Yes	No	Total
Yes	5=a	2=b	(a+b)=7
No	1=c	7=d	(c+d)=8
Total	(a+c)=6	(b+d)=9	(a+b+c+d)=n=15

Where smallest frequency is d=1 so the range (0 to 1) are the further table

$$P_1 = \frac{(a+b)!(c+d)!(a+c)!(b+b)!}{a!b!c!d!n!} = \frac{7!8!6!9!}{5!1!2!7!15!} = 0.034$$

2nd table

Toxity present	Tumer Regression		
	Yes	No	Total
Yes	6=a	1=b	(a+b)=7
No	0=c	8=d	(c+d)=8
Total	(a+c)=6	(b+d)=9	(a+b+c+d)=n=15

$$P_1 = \frac{(a+b)!(c+d)!(a+c)!(b+b)!}{a!b!c!d!n!} = \frac{7!8!6!9!}{6!0!1!8!15!} = 0.0014$$

$P = P_0 + P_1 = 0.034 + 0.0014 = 0.035$

$2P = 2(0.03539) = 0.071$

vi) Critical region

$2P \geq \alpha$

$0.071 \geq 0.05$

vii) Conclusion:

Since our calculated value fall's in critical region, so we reject H_0 and conclude that the classification of two variables are dependent at 5%.

Q.17.55©: Death in 6 months after fractured neck of femur in a specialized orthopaedic ward(A) and a general ward(B) are given below:

		Ward	
		A	B
Deaths	Yes	2	6
	No	18	14

Test the hypothesis of independence by using the Fisher-Irwin exact test.
Solution:
i) Set up our null and alternative hypothesis
H_0 : The classification of two variable are independent
H_1 : The classification of two variable are dependent
ii) Assumption: Samples are drawn randomly and independently from a approximately normal population.
iii) Level of significance
$\alpha = 5\% = 0.05$
iv) Test-statistic

$\chi_c^2 = 2P$ Where $P_i = \dfrac{(a+b)!(c+d)!(a+c)!(b+b)!}{a!b!c!d!n!}$

And $P = P_0 + P_1 + P_2 + ... + P_d$

Under H_0, it has chi-square distribution
v) Calculation:
1st table

		Ward		
		A	B	Total
Deaths	Yes	6=a	2=b	(a+b)=8
	No	14=c	18=d	(c+d)=32
Total		(a+c)=20	(b+d)=20	(a+b+c+d)=n=40

Where smallest frequency is d=2 so the range (0 to 2) are the further table

$$P_2 = \frac{(a+b)!(c+d)!(a+c)!(b+b)!}{a!b!c!d!n!} = \frac{20!20!8!32!}{6!2!14!18!40!} = 0.096$$

2nd table

		Ward		
		A	B	Total
Deaths	Yes	7=a	1=b	(a+b)=8
	No	13=c	19=d	(c+d)=32
Total		(a+c)=20	(b+d)=20	(a+b+c+d)=n=40

$$P_1 = \frac{(a+b)!(c+d)!(a+c)!(b+b)!}{a!b!c!d!n!} = \frac{20!20!8!32!}{7!1!13!19!40!} = 0.020$$

3rd table

		Ward		
		A	B	Total
Deaths	Yes	8=a	0=b	(a+b)=8
	No	12=c	20=d	(c+d)=32
Total		(a+c)=20	(b+d)=20	(a+b+c+d)=n=40

$$P_0 = \frac{(a+b)!(c+d)!(a+c)!(b+b)!}{a!b!c!d!n!} = \frac{20!20!8!32!}{8!0!12!20!40!} = 0.0016$$

$P = P_0 + P_1 + P_2 = 0.096 + 0.020 + 0.0016 = 0.1176$

$2P = 2(0.1176) = 0.235 = 0.24$

vi) Critical region
$2P \geq \alpha$
$0.24 \geq 0.05$
vii) Conclusion:
Since our calculated value fall's in critical region, so we reject H_0 and conclude that the classification of two variables are dependent at 5%.
Q:17.56: (a) The following data are intended to show dependence of brittleness in polyethylene bars on the duration of heat at a particular phase of the manufacturing process.

	Brittle	Tough
Treatment 1	2	8
Treatment 2	6	3

Use Fisher's exact test to test the null hypothesis that the brittleness of polyethylene bars

Does not vary with the two heat treatments.
Solution: Do yourself similarly Q.17.56©
Q.17.56 (b): Use Fisher's exact test to test the hypothesis that inoculation is independent of immunity from attack among population exposed to a certain disease from the following data.

	Inoculated	Not Inoculated
Attacked	9	2
Not Attacked	7	6

Solution: Do yourself similarly Q.17.56©
Q:17.57(a) In a study to determine whether or not the proportions of defective produced by worker was the same for the day, evening, or night shift worked, the following data were collected:

	Shift		
	Day	Evening	Night
Defective	45	55	70
Non Defective	905	890	870

Test the hypothesis at the 0.025 level of significance, that the proportion of defectives is the same for the all three shifts.
Solution:
i) Set up our null and alternative hypothesis
H_0 :*The proportion of all three shifts are same*
H_1 :*The proportion of all three shifts are not same*
ii) Assumption: Samples are drawn randomly and independently from a approximately normal population.
iii) Level of significance
$\alpha = 0.05$
iv) Test-statistic
$$\chi_c^2 = \sum \frac{(f_0 - f_e)^2}{f_e} = \sum \frac{(O_i - e_i)^2}{e_i}$$
Under H_0 ,It has chi-square distribution with $V = (k-1)$ degree of freedom
v) Calculation: First we calculate expected frequencies e

	Shift			
	Day	Night	Evening	Total
Defective	$\frac{950 \times 170}{2835} = 56.97$	$\frac{940 \times 170}{2835} = 56.37$	$\frac{945 \times 170}{2835} = 56.67$	170
Non Defective	$\frac{950 \times 2665}{2835} = 893.03$	$\frac{940 \times 2665}{2835} = 883.63$	$\frac{945 \times 2665}{2835} = 888.33$	2665
Total	950	940	945	2835

Now further calculation

O_i	e_i	$(O_i - e_i)^2 / e_i$
45	56.97	2.5150
55	56.37	0.0333
70	56.67	3.1355
905	893.03	0.1604
890	883.63	0.0459
870	888.333	0.3782
$\sum O_i = 2835$	$\sum e_i = 2835$	$\sum (O_i - e_i)^2 / e_i = 6.27$

$$\chi_c^2 = \sum (O_i - e_i)^2 / e_i = 6.27$$
vi) Critical region
$\chi_c^2 > \chi_\alpha^2(v)$
$\chi_c^2 > \chi_{0.05}^2(5)$

$\chi_c^2 > 12.59$ $\quad v = (k-1) = 6-1 = 5$

vii) Conclusion:

Since our calculated value does not fall in critical region, so we accept H_0 and we conclude that the proportion of all groups are same at $\alpha = 0.05$.

Q: 17.57(b): There are five classes, each having 50 students. The result of these five classes is given below.

Class	1	2	3	4	5	Total
Pass	42	45	43	45	45	220
Failure	8	5	7	5	5	30
Total	50	50	50	50	50	250

Test the null hypothesis, H_0: 4 Pass: 1 Failure, Using totals.

Solution:

i) Set up our null and alternative hypothesis

H_0 : *Data follow the ratio* 4 *Pass* : 1 *Failure*

H_1 : *Data do not follow the ratio* 4 *Pass* : 1 *Failure*

ii) Assumption: Samples are drawn randomly and independently from a approximately normal population.

iii) Level of significance $\alpha = 0.01$

iv) Test-statistic

$$\chi_c^2 = \sum \frac{(f_0 - f_e)^2}{f_e} = \sum \frac{(O_i - e_i)^2}{e_i}$$

Under H_0, it has chi-square distribution with $V = k-1$ degree of freedom

v) Critical region

$\chi_c^2 > \chi_\alpha^2(v)$

$\chi_c^2 > \chi_{0.01}^2(4)$

$\chi_c^2 > 13.28$ $\quad v = k-1 = 5-1 = 4$

vi) Calculation: Now further calculation

$O_i = Pass$	$e_i = n_i p = 50\frac{4}{5}$	$(O_i - e_i)^2 / e_i$	$O_i = Fail$	$e_i = n_i p = 50\frac{1}{5}$	$(O_i - e_i)^2 / e_i$
42 45 43 45 45	40 40 40 40 40	0.1 0.625 0.225 0.625 0.625	8 5 7 5 5	10 10 10 10 10	0.4 2.5 0.9 2.5 2.5
220	200	$\sum (O_i - e_i)^2 / e_i = 2.2$	30	50	$\sum (O_i - e_i)^2 / e_i = 8.8$

$$\chi_c^2 = \chi_1^2 + \chi_2^2 = 2.20 + 8.8 = 11.0$$

vii) Conclusion:

Since our calculated value fall's in critical region, so we reject H_0 and we conclude that the attributes are associated at $\alpha = 0.01$.

Q: 17.58: $\chi_c^2 = \dfrac{N^2}{AB}\left[\sum_{i=1}^{k} \dfrac{a_i^2}{c_i} - \dfrac{A^2}{N}\right]$

Use this formula to find χ^2 from the following data:

No. of rows of Kernels	8	10	12	14	Total
Present	18	37	27	0	82
Absent	15	26	43	4	88
Total	33	63	70	4	170

Solution: Do yourself similarly Q.17.60

Q: 17.59(a): Prove the Brandt-Snedecor formula for χ^2

Ans: Chi-Square (χ^2) as a test of $2\times m$ contingency (Brand Sendecor Formula)

The test procedure is given as independence of attributes but a similar method for the computation of χ^2 statistic for $2\times m$ contingency table for two independence and equal sizes has been derived by Brandt Sendecor. So calculate χ^2 test statistic which does not use the expected cell frequencies is given by

$$\chi_c^2 = \frac{n^2}{R_1R_2}\left[\sum_{i=1}^{m}\frac{X_i^2}{c_i} - \frac{R_1^2}{n}\right]$$

Where

R_1= Total of 1st row

R_2= Total of 2nd r0w

n= Grand total

C_i=Total of ith column

Theorem

Show that for $2\times m$ contingency table

$$\chi_c^2 = \frac{n^2}{R_1R_2}\left[\sum_{i=1}^{m}\frac{X_i^2}{c_i} - \frac{R_1^2}{n}\right]$$

Proof: Suppose we draw two independent random sample of size "m" from a population and we wish to test wether the two samples are independent.

Let the value of two samples be presented in the following $2\times m$

Or

Let we have a $2\times m$ contingency table

Observation/Samples	1 2 3… i… m	Total
I	X_1 X_2 X_3… X_i… X_m	R_1
II	Y_1 Y_2 Y_3… Y_i… Y_m	R_2
Total	C_1 C_2 C_3… C_i … C_m	

As $\chi_c^2 = \sum\frac{(O_i - e_i)^2}{e_i}$

$$\chi_c^2 = \sum\left[\frac{(O_i)^2}{e_i} + \frac{(e_i)^2}{e_i} - 2\frac{(O_ie_i)}{e_i}\right]$$

$$\chi_c^2 = \sum_{i=1}\frac{O_i^2}{e_i} + \sum_{i=1} e_i - 2\sum_{i=1} O_i \qquad \text{Therefore } \sum_{i=1} e_i = \sum_{i=1} O_i = n$$

$$\chi_c^2 = \sum_{i=1}\frac{O_i^2}{e_i} + n - 2n$$

$$\chi_c^2 = \sum_{i=1}\frac{O_i^2}{e_i} - n$$

Therefore expected value $e_i = \dfrac{R_1C_i}{n}$ for 1st sample $e_i = \dfrac{R_2C_i}{n}$ for 2nd sample

And $C_i = X_i + Y_i$ also $O_i^2 = X_i^2 + Y_i^2$ and $C_i - Y_i = X_i \quad or \quad C_i - X_i = Y_i$

$$\chi_c^2 = \sum_{i=1}\frac{X_i^2 + Y_i^2}{e_i} - n$$

$$\chi_c^2 = \sum_{i=1}\left[\frac{X_i^2}{\frac{R_1C_i}{n}} + \frac{Y_i^2}{\frac{R_2C_i}{n}}\right] - n$$

$$\chi_c^2 = n\sum_{i=1}\left[\frac{X_i^2}{R_1C_i} + \frac{(C_i - X_i)^2}{R_2C_i}\right] - n$$

$$\chi_c^2 = n\sum_{i=1}\left[\frac{X_i^2}{R_1C_i} + \frac{C_i^2 + X_i^2 - 2C_iX_i}{R_2C_i}\right] - n$$

$$\chi_c^2 = n\sum_{i=1}\left[\frac{X_i^2}{R_1C_i}+\frac{C_i^2}{R_2C_i}+\frac{X_i^2}{R_2C_i}-2\frac{C_iX_i}{R_2C_i}\right]-n$$

$$\chi_c^2 = n\sum_{i=1}\left[\frac{X_i^2}{C_i}\left(\frac{1}{R_1}+\frac{1}{R_2}\right)+\frac{C_i}{R_2}-2\frac{X_i}{R_2}\right]-n$$

$$\chi_c^2 = n\sum_{i=1}\left[\frac{X_i^2}{C_i}\left(\frac{R_2+R_1}{R_1R_2}\right)+\frac{C_i}{R_2}-2\frac{X_i}{R_2}\right]-n$$ Therefore $R_2+R_1=n$

$$\chi_c^2 = n\sum_{i=1}\left[\frac{X_i^2}{C_i}\left(\frac{n}{R_1R_2}\right)+\frac{C_i}{R_2}-2\frac{X_i}{R_2}\right]-n$$

$$\chi_c^2 = \frac{n^2}{R_1R_2}\sum_{i=1}\frac{X_i^2}{C_i}+\frac{n\sum_{i=1}C_i}{R_2}-\frac{2n\sum_{i=1}X_i}{R_2}-n$$ Therefore $\sum_{i=1}^{m}C_i=n$ and $\sum_{i=1}^{m}X_i=R_1$

$$\chi_c^2 = \frac{n^2}{R_1R_2}\sum_{i=1}\frac{X_i^2}{C_i}+\frac{nn}{R_2}-\frac{2nR_1}{R_2}-n$$

$$\chi_c^2 = \frac{n^2}{R_1R_2}\sum_{i=1}\frac{X_i^2}{C_i}+\frac{1}{R_2}\left(n^2-2nR_1-nR_2\right)$$

$$\chi_c^2 = \frac{n^2}{R_1R_2}\sum_{i=1}\frac{X_i^2}{C_i}+\frac{1}{R_2}\left(n^2-nR_1-nR_1-nR_2\right)$$

$$\chi_c^2 = \frac{n^2}{R_1R_2}\sum_{i=1}\frac{X_i^2}{C_i}+\frac{1}{R_2}\left(n^2-nR_1-n(R_1+R_2)\right)$$

$$\chi_c^2 = \frac{n^2}{R_1R_2}\sum_{i=1}\frac{X_i^2}{C_i}+\frac{1}{R_2}\left(n^2-nR_1-n(n)\right)$$

$$\chi_c^2 = \frac{n^2}{R_1R_2}\sum_{i=1}\frac{X_i^2}{C_i}+\frac{1}{R_2}\left(n^2-nR_1-n^2\right)$$

$$\chi_c^2 = \frac{n^2}{R_1R_2}\sum_{i=1}\frac{X_i^2}{C_i}+\frac{1}{R_2}\left(-nR_1\right)$$

$$\chi_c^2 = \frac{n^2}{R_1R_2}\sum_{i=1}\frac{X_i^2}{C_i}-\frac{nR_1}{R_2}\frac{nR_1}{nR_1}$$

$$\chi_c^2 = \frac{n^2}{R_1R_2}\sum_{i=1}\frac{X_i^2}{C_i}-\frac{n^2R_1^2}{nR_1R_2}$$

$$\chi_c^2 = \frac{n^2}{R_1R_2}\left[\sum_{i=1}\frac{X_i^2}{C_i}-\frac{R_1^2}{n}\right]$$

Hence proved

Q.17.59 (b): Two groups of freshmen applying to enter a university took the same college aptitude test. The group (A & B) differed in the type of high school education they had experienced.

Scores	0-9	10-19	20-29	30-39	40-49	50-59	60-69	70-79	80-89	90-99
Group A	71	68	66	47	51	30	43	39	33	18
Group B	22	8	14	12	3	13	3	14	12	10

The frequency distributions of scores for the two groups were as follows:
Calculate the value of χ^2 and determine whether there is a significant difference in college aptitude test between the groups.
Solution: Do yourself similarly Q.17.60
Q: 17.60: In an experiment on the effectiveness of a teaching machine, a machine instructed group of students was compared with teacher instructed group on an achievement test. The following scores were obtained:

	40-49	50-59	60-69	70-79	80-89	90-99	Total
Teacher -	21	40	55	38	10	2	166

instructed							
Machine-instructed	18	35	42	46	19	4	164
Total	39	75	97	84	29	6	330

Calculate χ_c^2 and determine whether there is a significant difference in achievements of the two groups

Solution:

i) Set up our null and alternative hypothesis

H_0 :The two groups are homogeneous

H_1 :The two groups are not homogeneous

ii) Assumption: Samples are drawn randomly and independently from a approximately normal population of qualitative nature.

iii) Level of significance

$\alpha = 0.05$

iv) Test-statistic

$$\chi_c^2 = \frac{N^2}{AB}\left[\sum_{i=1}^{k}\frac{a_i^2}{c_i} - \frac{A^2}{N}\right]$$

Under H_0, it has chi-square distribution with $V = k-1$ degree of freedom

v) Critical region

$\chi_c^2 > \chi_\alpha^2(v)$

$\chi_c^2 > \chi_{0.05}^2(5)$

$\chi_c^2 > 11.07$ $\qquad v = n-1 = 5$

vi) Calculation: First we find actual frequencies

	40-49	50-59	60-69	70-79	80-89	90-99	Total
Teacher –instructed (a_i)	21	40	55	38	10	2	166(A)
Machine- instructed	18	35	42	46	19	4	164(B)
Total (C_i)	39	75	97	84	29	6	330(N)

$$\chi_c^2 = \frac{(330)^2}{(166)(144)}\left[\frac{(21)^2}{39} + \frac{(40)^2}{75} + \frac{(55)^2}{97} + \frac{(38)^2}{84} + \frac{(10)^2}{29} + \frac{(2)^2}{6} - \frac{(166)^2}{330}\right] = 6.5166$$

vii) Conclusion:

Since our calculated value does not fall in critical region, so we accept H_0 and we conclude that there is no significance difference between the achievements of two groups at $\alpha = 0.05$

Theorem

Cox-Good Formula

Show that in $2 \times m$

$$\chi_c^2 = n\left[1 - \frac{n}{R_1 R_2}\sum_{i=1}^{m}\frac{X_i Y_i}{C_i}\right]$$

This formula was drived by our respected Sir Dr.M.A.Malik in which he found the varies of the both samples. This formula known as Dr..M.A.Malik formula

Proof: Let we have a $2 \times m$ contingency table

Observation/Samples	1 2 3… i… m	Total
I	X_1 X_2 X_3… X_i… X_m	R_1
II	Y_1 Y_2 Y_3… Y_i… Y_m	R_2
Total	C_1 C_2 C_3… C_i … C_m	

As $\chi_c^2 = \sum\frac{(O_i - e_i)^2}{e_i}$

$$\chi_c^2 = \sum\left[\frac{(O_i)^2}{e_i} + \frac{(e_i)^2}{e_i} - 2\frac{(O_i e_i)}{e_i}\right]$$

$$\chi_c^2 = \sum_{i=1} \frac{O_i^{\,2}}{e_i} + \sum_{i=1} e_i - 2\sum_{i=1} O_i \qquad \text{Therefore } \sum_{i=1} e_i = \sum_{i=1} O_i = n$$

$$\chi_c^2 = \sum_{i=1} \frac{O_i^{\,2}}{e_i} + n - 2n$$

$$\chi_c^2 = \sum_{i=1} \frac{O_i^{\,2}}{e_i} - n$$

Therefore expected value $e_i = \frac{R_1 C_i}{n}$ for 1st sample $e_i = \frac{R_2 C_i}{n}$ for 2nd sample

And $C_i = X_i + Y_i$ also $O_i^2 = X_i^{\,2} + Y_i^2$ and $C_i - Y_i = X_i$ *or* $C_i - X_i = Y_i$

$$\chi_c^2 = \sum_{i=1} \frac{X_i^{\,2} + Y_i^2}{e_i} - n$$

$$\chi_c^2 = \sum_{i=1} \left[\frac{X_i^{\,2}}{\frac{R_1 C_i}{n}} + \frac{Y_i^{\,2}}{\frac{R_2 C_i}{n}} \right] - n$$

$$\chi_c^2 = n\sum_{i=1} \left[\frac{X_i^{\,2}}{R_1 C_i} + \frac{Y_i^{\,2}}{R_2 C_i} \right] - n$$

$$\chi_c^2 = \frac{n}{R_1} \sum_{i=1}^{m} \frac{X_i^{\,2}}{C_i} + \frac{n}{R_2} \sum_{i=1}^{m} \frac{Y_i^{\,2}}{C_i} - n$$

Therefore

$$\sum_{i=1}^{m} \frac{X_i^{\,2}}{C_i} = \sum_{i=1}^{m} \frac{X_i X_i}{C_i} = \sum_{i=1}^{m} \frac{X_i (C_i - Y_i)}{C_i} = \sum_{i=1}^{m} \frac{X_i C_i - X_i Y_i}{C_i} = \sum_{i=1}^{m} \frac{X_i C_i}{C_i} - \sum_{i=1}^{m} \frac{X_i Y_i}{C_i} = \sum_{i=1}^{m} X_i - \sum_{i=1}^{m} \frac{X_i Y_i}{C_i} = R_1 - \sum_{i=1}^{m} \frac{X_i Y_i}{C_i}$$

Similarly

$$\sum_{i=1}^{m} \frac{Y_i^2}{C_i} = R_2 - \sum_{i=1}^{m} \frac{X_i Y_i}{C_i}$$

$$\chi_c^2 = \frac{n}{R_1} \left(R_1 - \sum_{i=1}^{m} \frac{X_i Y_i}{C_i} \right) + \frac{n}{R_2} \left(R_2 - \sum_{i=1}^{m} \frac{X_i Y_i}{C_i} \right) - n$$

$$\chi_c^2 = \left(R_1 \frac{n}{R_1} - \frac{n}{R_1} \sum_{i=1}^{m} \frac{X_i Y_i}{C_i} \right) + \left(R_2 \frac{n}{R_2} - \frac{n}{R_2} \sum_{i=1}^{m} \frac{X_i Y_i}{C_i} \right) - n$$

$$\chi_c^2 = n - \frac{n}{R_1} \sum_{i=1}^{m} \frac{X_i Y_i}{C_i} + n - \frac{n}{R_2} \sum_{i=1}^{m} \frac{X_i Y_i}{C_i} - n$$

$$\chi_c^2 = n - \frac{n}{R_1} \sum_{i=1}^{m} \frac{X_i Y_i}{C_i} - \frac{n}{R_2} \sum_{i=1}^{m} \frac{X_i Y_i}{C_i}$$

$$\chi_c^2 = n - \left(\frac{n}{R_1} + \frac{n}{R_2} \right) \sum_{i=1}^{m} \frac{X_i Y_i}{C_i}$$

$$\chi_c^2 = n - n\left(\frac{R_2 + R_1}{R_1 R_2} \right) \sum_{i=1}^{m} \frac{X_i Y_i}{C_i}$$

$$\chi_c^2 = n - n\left(\frac{n}{R_1 R_2} \right) \sum_{i=1}^{m} \frac{X_i Y_i}{C_i} \qquad \text{Therefore} \qquad R_2 + R_1 = n$$

$$\chi_c^2 = n\left[1 - \frac{n}{R_1 R_2} \sum_{i=1}^{m} \frac{X_i Y_i}{C_i} \right]$$

Hence proved

Theorem

Derivation of χ^2 by J-Scory method or Statistic

$$\chi_c^2 = n[R - !] \qquad \text{Where } R = \sum_{j=1}^{c} \frac{T_j}{B_j} \qquad \textit{And} \quad T_J = \sum_{i=1}^{r} \frac{O_{ij}^2}{A_i}$$

Proof: χ^2 test statistic for 2×2 contingency table is

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-e_{ij})^2}{e_{ij}}$$

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\left[\frac{(O_{ij})^2}{e_{ij}}+\frac{(e_{ij})^2}{e_{ij}}-2\frac{(O_{ij}e_{ij})}{e_{ij}}\right]$$

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{O_{ij}^{\ 2}}{e_{ij}}+\sum_{i=1}^{r}\sum_{j=1}^{c}e_{ij}-2\sum_{i=1}^{r}\sum_{j=1}^{c}O_{ij}$$ Therefore $\sum_{i=1}^{r}\sum_{j=1}^{c}O_{ij}=\sum_{i=1}^{r}\sum_{j=1}^{c}e_{ij}=n$

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{O_{ij}^{\ 2}}{e_{ij}}+n-2n$$

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{O_{ij}^{\ 2}}{e_{ij}}-n$$ Therefore $e_{ij}=\frac{A_iB_j}{n}$

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{O_{ij}^{\ 2}}{\frac{A_iB_j}{n}}-n$$

$$\chi^2 = n\sum_{i=1}^{c}\frac{1}{B_j}\sum_{i=1}^{r}\frac{O_{ij}^{\ 2}}{A_i}-n$$ Therefore $T_j=\sum_{i=1}^{r}\frac{O_{ij}^{\ 2}}{A_i}-n$

$$\chi^2 = n\sum_{i=1}^{c}\frac{1}{B_j}T_j-n$$ Therefore $R=\sum_{i=1}^{c}\frac{1}{B_j}T_j$

$$\chi^2 = nR-n$$

$$\chi^2 = n(R-1)$$

Hence proved

Theorem

Show that in a 2×2 contingency table where in the frequency table are

a	B
c	d

The value of χ^2 calculated on the hypothesis of independence given by

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

Proof: Let we have a 2×2 contingency

	B_1	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

Under the hypothesis of independence we calculate the expected frequencies

$$e_{ij} = \frac{ith\ \ row\ \ total\times jth\ \ column\ \ total}{total\,number\ of\ observation}$$

$$e_{11} = \frac{(a+b)\times(a+c)}{(a+b+c+d)}$$

$$e_{12} = \frac{(a+b)\times(b+d)}{(a+b+c+d)}$$

$$e_{21} = \frac{(c+d)\times(a+c)}{(a+b+c+d)}$$

$$e_{22} = \frac{(c+d)\times(b+d)}{(a+b+c+d)}$$

As we know that

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-e_{ij})^2}{e_{ij}}$$

$$\chi^2 = \frac{\left[a - \frac{(a+b)\times(a+c)}{(a+b+c+d)}\right]^2}{\frac{(a+b)\times(a+c)}{(a+b+c+d)}} + \frac{\left[b - \frac{(a+b)\times(b+d)}{(a+b+c+d)}\right]^2}{\frac{(a+b)\times(b+d)}{(a+b+c+d)}} + \frac{\left[c - \frac{(c+d)\times(a+c)}{(a+b+c+d)}\right]^2}{\frac{(c+d)\times(a+c)}{(a+b+c+d)}} + \frac{\left[d - \frac{(b+d)\times(c+d)}{(a+b+c+d)}\right]^2}{\frac{(b+d)\times(c+d)}{(a+b+c+d)}}$$

$$\chi^2 = \frac{\left[\frac{a(a+b+c+d)-(a+b)\times(a+c)}{(a+b+c+d)}\right]^2}{\frac{(a+b)\times(a+c)}{(a+b+c+d)}} + \frac{\left[\frac{b(a+b+c+d)-(a+b)\times(b+d)}{(a+b+c+d)}\right]^2}{\frac{(a+b)\times(b+d)}{(a+b+c+d)}} + \frac{\left[\frac{c(a+b+c+d)-(c+d)\times(a+c)}{(a+b+c+d)}\right]^2}{\frac{(c+d)\times(a+c)}{(a+b+c+d)}}$$

$$+ \frac{\left[\frac{d(a+b+c+d)-(b+d)\times(c+d)}{(a+b+c+d)}\right]^2}{\frac{(b+d)\times(c+d)}{(a+b+c+d)}}$$

$$\chi^2 = \frac{\left[\frac{a^2+ab+ac+ad-a^2-ac-ab-bc)}{(a+b+c+d)}\right]^2}{\frac{(a+b)\times(a+c)}{(a+b+c+d)}} + \frac{\left[\frac{ab+b^2+bc+bd-ab-b^2-ad-bd}{(a+b+c+d)}\right]^2}{\frac{(a+b)\times(b+d)}{(a+b+c+d)}}$$

$$+ \frac{\left[\frac{ac+bc+c^2+cd-ac-c^2-ad-cd}{(a+b+c+d)}\right]^2}{\frac{(c+d)\times(a+c)}{(a+b+c+d)}} + \frac{\left[\frac{ad+bd+cd+d^2)-bc-bd-cd-d^2}{(a+b+c+d)}\right]^2}{\frac{(b+d)\times(c+d)}{(a+b+c+d)}}$$

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

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

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

Therefore $[ad-bc]^2 = [bc-ad]^2$

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

$$\chi^2 = \frac{[ad-bc)]^2}{(a+b+c+d)}\left[\frac{1}{(a+b)\times(a+c)} + \frac{1}{(a+b)\times(b+d)} + \frac{1}{(c+d)\times(a+c)} + \frac{1}{(b+d)\times(c+d)}\right]$$

$$\chi^2 = \frac{[ad-bc)]^2}{(a+b+c+d)}\left[\frac{(b+d)(c+d)+(a+c)(c+d)+(a+b)(b+d)+(a+b)(a+c)}{(a+b)(a+c)(b+d)(c+d)}\right]$$

$$\chi^2 = \frac{[ad-bc)]^2}{(a+b+c+d)}\left[\frac{bc+bd+cd+d^2+ac+ad+c^2+cd+ab+ad+b^2+bd+a^2+ac+ab+bc}{(a+b)(a+c)(b+d)(c+d)}\right]$$

$$\chi^2 = \frac{[ad-bc)]^2}{(a+b+c+d)}\left[\frac{a^2+b^2+c^2+d^2+2ac+2bc+2bd+2cd+2ad+2ab}{(a+b)(a+c)(b+d)(c+d)}\right]$$

$$\chi^2 = \frac{[ad-bc)]^2}{(a+b+c+d)}\left[\frac{(a+b+c+d)^2}{(a+b)(a+c)(b+d)(c+d)}\right]$$

$$\chi^2 = \left[\frac{(a+b+c+d)[ad-bc)]^2}{(a+b)(a+c)(b+d)(c+d)}\right]$$

$$\chi^2 = \frac{(a+b+c+d)[ad-bc)]^2}{(a+b)(a+c)(b+d)(c+d)} = \frac{n[ad-bc)]^2}{(a+b)(a+c)(b+d)(c+d)}$$

Therefore (a+b+c+d)=n

Hence proved

Correlation coefficient

Procedure:

i) We set up our null and alternative hypothesis

$H_0 : \rho_1 = \rho_2 = \rho_3 = ... = \rho_k$ *Or Population Correlation coefficient are homogeneous*
H_1 : *At least two popultion Correlation coefficient are different*
ii) Assumption: A sample is drawn randomly and independently from a bivariate normal population
iii) Level of significance
$\alpha = (Comonly \;\; uesd \; 5\% \; or \; 1\%)$
iv) Test-statistic

$$\chi^2 = \sum_{i=1}^{k} (n_i - 3) Z_i^2 - \frac{\left[\sum_{i=1}^{k} (n_i - 3) Z_i\right]^2}{\sum_{i=1}^{k} (n_i - 3)}$$

Under H_0; it has χ^2 –distribution with $v = (k-1)df$

v) Critical region
It is naturally depend on alternative hypothesis
$\chi^2 > \chi_\alpha{}^2(v)$
vi) Calculation
In this step we calculate the value of "χ^2" test statistic on the basis of sample data.

S.no	r_i	n_i	$n_i - 1$	$Z_i = \frac{1}{2}\ln\frac{1+r_i}{1-r_i}$	$(n_i - 1)Z_i$	$(n_i - 1)Z_i^2$

vii) Conclusion
If our calculated value does not fall's in critical region then we accept H_0 other wise we reject it.
Example:
Random samples of 10, 15, 20 are drawn from a bivariate normal population yielding r=0.3, 0.4, 0.49 respectively from combine estimate of "ρ" and test that the hypothesis the correlation coefficient are homogeneous.
i) We set up our null and alternative hypothesis
$H_0 : \rho_1 = \rho_2 = \rho_3 = ... = \rho_k$ *Or Population Correlation coefficient are homogeneous*
H_1 : *At least two popultion Correlation coefficient are different*
ii) Assumption: A sample is drawn randomly and independently from a bivariate normal population
iii) Level of significance $\alpha = 5\% = 0.05$
iv) Test-statistic

$$\chi^2 = \sum_{i=1}^{k} (n_i - 3) Z_i^2 - \frac{\left[\sum_{i=1}^{k} (n_i - 3) Z_i\right]^2}{\sum_{i=1}^{k} (n_i - 3)}$$

Under H_0; it has χ^2 –distribution with $v = (k-1)df$
v) Critical region
$\chi^2 \geq \chi_\alpha{}^2(v)$
$\chi^2 \geq \chi_{0.05}{}^2(3-1=2)$
$\chi^2 \geq 5.99$
vi) Calculation
In this step we calculate the value of "χ^2" test statistic on the basis of sample data.

S.no	r_i	n_i	$n_i - 3$	$Z_i = \frac{1}{2}\ln\frac{1+r_i}{1-r_i}$	$(n_i - 3)Z_i$	$(n_i - 3)Z_i^2$
1	0.3	10	7	0.3095	2.1665	0.6705
2	0.4	15	12	0.4236	5.0832	2.1532
3	0.49	20	17	0.5361	9.1137	4.8858
			36		16.3634	7.7096

$$\chi^2 = \sum_{i=1}^{k}(n_i - 3)Z_i^2 - \frac{\left[\sum_{i=1}^{k}(n_i - 3)Z_i\right]^2}{\sum_{i=1}^{k}(n_i - 3)}$$

$$\chi^2 = 7.7096 - \frac{[16.3634]^2}{36} = 7.7096 - 7.4378 = 0.2718$$

vii) Conclusion

Since our calculated value does not fall's in critical region. So, we accept H_0 and conclude that several simple correlation coefficients are homogeneous at 5% level of significance.

Partial Correlation coefficient

Procedure:

Partial Correlation coefficient

Procedure:

i) We set up our null and alternative hypothesis

$H_0: Population\ partial\ Correlation\ coefficient\ are\ \text{hom}\, ogeneous$

$H_1: Population\ partial\ Correlation\ coefficient\ are\ not\ \text{hom}\, ogeneous$

ii) Assumption: A sample is drawn randomly and independently from a multinomial normal population

iii) Level of significance

$\alpha = (Comonly\ \ uesd\ 5\%\ or\ 1\%)$

iv) Test-statistic

$$\chi^2 = \sum_{i=1}^{k}(n_i - 3 - p)Z_i^2 - \frac{\left[\sum_{i=1}^{k}(n_i - 3 - p)Z_i\right]^2}{\sum_{i=1}^{k}(n_i - 3 - p)}$$

Under H_0; it has χ^2 –distribution with $v = (k-1)df$

Where "P" number of variable held as constant

v) Critical region

It is naturally depend on alternative hypothesis

$\chi^2 \geq \chi_\alpha{}^2(v)$

vi) Calculation

In this step we calculate the value of "χ^2" test statistic on the basis of sample data.

S.no	r_i	n_i	$n_i - 3 - p$	$Z_i = \frac{1}{2}\ln\frac{1+r_i}{1-r_i}$	$(n_i - 3 - p)Z_i$	$(n_i - 3 - p)Z_i^2$

vii) Conclusion

If our calculated value does not fall's in critical region then we accept H_0 other wise we reject it.

Example:

From the samples of 24, 29,33,38,42 are the partial correlation coefficients of order "3" are found to be 0.38, 0.45, 0.60, 0.50 and 0.42. Test the homogeneity of the partial correlation coefficient at 5% level of significance.

Solution:

i) We set up our null and alternative hypothesis

$H_0: Population\ partial\ Correlation\ coefficient\ are\ \text{hom}\, ogeneous$

$H_1: Population\ partial\ Correlation\ coefficient\ are\ not\ \text{hom}\, ogeneous$

ii) Assumption: A sample is drawn randomly and independently from a multinomial normal population

iii) Level of significance

$\alpha = 5\% = 0.05$

iv) Test-statistic

$$\chi^2 = \sum_{i=1}^{k}(n_i - 3 - p)Z_i^2 - \frac{\left[\sum_{i=1}^{k}(n_i - 3 - p)Z_i\right]^2}{\sum_{i=1}^{k}(n_i - 3 - p)}$$

Under H_0; it has χ^2 –distribution with $v = (k-1)df$

Where "P" number of variable held as constant

v) Critical region

$\chi^2 \geq \chi_\alpha{}^2(v)$

$\chi^2 \geq \chi_{0.05}{}^2(5-1=4)$

$\chi^2 \geq 9.49$

vi) Calculation

In this step we calculate the value of "χ^2" test statistic on the basis of sample data.

S.no	r_i	n_i	$n_i - 3 - p$ P=3	$Z_i = \frac{1}{2}\ln\frac{1+r_i}{1-r_i}$	$(n_i - 3 - p)Z_i$	$(n_i - 3 - p)Z_i^2$
1	0.38	24	18	0.4001	7.2018	2.8814
2	0.54	29	23	0.6042	13.8966	8.3969
3	0.60	33	27	0.6931	18.7137	12.9705
4	0.50	38	32	0.5493	17.5776	9.6554
5	0.42	42	36	0.4477	16.1172	7.2157
			136		73.5069	41.1199

$$\chi^2 = \sum_{i=1}^{k}(n_i - 3 - p)Z_i^2 - \frac{\left[\sum_{i=1}^{k}(n_i - 3 - p)Z_i\right]^2}{\sum_{i=1}^{k}(n_i - 3 - p)}$$

$$\chi^2 = 41.1199 - \frac{[73.5069]^2}{136} = 41.1199 - 39.7299 = 1.39$$

vii) Conclusion

Since our calculated value does not fall's in critical region. So, we accept H_0 and conclude that the partial correlation coefficients are homogeneous at 5% level of significance.