LARGE SAMPLE THEORY

BY

TANUJIT CHAKRABORTY

Indian Statistical Institute

Mail : tanujitisi@gmail.com

LARGE SAMPLE THEORY

– Tanujit Chakraborty.

Sometimes the determination of the exact distribution of a statistic is difficult for a finite value of n and also the assumptions made about the probability distribution may not be valid. In such cases, if the limiting distributions of the statistics exist, then the problems of testing of hypothesis and of setting confidence intervals may be easily solved for large samples. But these will be only approximate for a finite value of n.

The advantages of these large-sample approximate results are that we do not have to make too many assumptions about the parent population and that in most cases the limiting distribution is normal, so that normal theory can be applied to get approximate tests and confidence intervals. Another important limiting distribution used in connection with categorical data is the χ^2 distribution. These large-sample methods are useful in practical applications.

If the asymptotic distribution of a statistic is normal, then for large n we can perform approximate large-sample tests about the mean or variance of the statistic. We can also perform LR tests using the approximate χ^2-distn. In this chapter, we shall consider the case of categorical data and also discuss certain transformations of statistics that are used in large samples.

* THEORY OF LARGE SAMPLES *

Transformations of statistics to stabilize variance :—

Theorem:— If $\{T_n\}$ is a sequence of statistics such that

$$\sqrt{n}(T_n - \theta) \xrightarrow{L} Z \sim N(0, \sigma_T^2(\theta)), \text{ then}$$

$$\sqrt{n}\left[g(T_n) - g(\theta)\right] \xrightarrow{L} Z_1 \sim N\left(0, [\sigma_T(\theta) g'(\theta)]^2\right),$$

provided $g'(\theta)$ exists and non-zero.

Proof:- Consider the Taylor expansion of $g(T_n)$ around $g(\theta)$:

$$g(T_n) = g(\theta) + (T_n - \theta)\{g'(\theta) + \epsilon_n\}$$ where $\epsilon_n \to 0$ as $T_n \to \theta$.

Therefore, for any $\epsilon > 0$, $|\epsilon_n| < \epsilon$ whenever $|T_n - \theta| < \delta$. Hence, for any $\epsilon > 0$,

$$P\left[|\epsilon_n| < \epsilon\right] = P\left[|T_n - \theta| < \delta\right] = P\left[\frac{\sqrt{n}|T_n - \theta|}{\sigma_T(\theta)} < \frac{\sqrt{n}\,\delta}{\sigma_T(\theta)}\right]$$

$$= P\left[|\tau| < \frac{\sqrt{n}\,\delta}{\sigma_T(\theta)}\right], \text{ where } \tau \sim N(0,1).$$

$$= \left\{2\Phi\left(\frac{\sqrt{n}\,\delta}{\sigma_T(\theta)}\right) - 1\right\}$$

$$\longrightarrow 1 \text{ as } n \to \infty.$$

By definition, $\epsilon_n \xrightarrow{P} 0$.

Now, $\sqrt{n}\{g(T_n) - g(\theta)\} - \sqrt{n}(T_n - \theta)g'(\theta) = \sqrt{n}(T_n - \theta)\cdot\epsilon_n \xrightarrow{P} 0$

Hence $\sqrt{n}\{g(T_n) - g(\theta)\} \xrightarrow{L} \sqrt{n}(T_n - \theta)g'(\theta) \sim N\left(0, [\sigma_T(\theta)g'(\theta)]^2\right)$.

Stabilization of Variance :—

If $\{T_n\}$ is a sequence of statistics $\ni$ $\sqrt{n}(T_n - \theta) \overset{a}{\sim} N(0, \sigma_T^2(\theta))$, where $\sigma_T^2(\theta)$ depends on θ, then

$$\sqrt{n}\{g(T_n) - g(\theta)\} \overset{a}{\sim} N\left(0, [\sigma_T(\theta)g'(\theta)]^2\right), \text{ provided } g'(\theta) \text{ exists.}$$

We wish to find a transformation $g(T_n)$ whose variance is independent of θ (or, variance is stable w.r.t. θ).

To stabilize the variance, we choose the function $g(T_n)$ $\ni$

$[\sigma_T(\theta)g'(\theta)]^2 = c$ (independent of θ), the asymptotic variance of the transformed statistic $g(T_n)$ will be independent of θ.

Now, solving $\sigma_T(\theta)g'(\theta) = c$, we get

$$g'(\theta) = \frac{c}{\sigma_T(\theta)} \quad \text{and} \quad \int g'(\theta)\,d\theta = c\int \frac{d\theta}{\sigma_T(\theta)}$$

$$\Rightarrow g(\theta) = c\int \frac{d\theta}{\sigma_T(\theta)}$$ is a <u>variance stabilizing transformation.</u>

Hence $\sqrt{n}\{g(T_n) - g(\theta)\} \overset{a}{\sim} N(0, c^2) \Rightarrow \frac{\sqrt{n}\{g(T_n) - g(\theta)\}}{c} \overset{a}{\sim} N(0,1)$.

We reject $H_0: \theta = \theta_0$ Vs $H_1: \theta \neq \theta_0$ at level α if the observed

$$\left|\frac{\sqrt{n}\{g(T_n) - g(\theta_0)\}}{c}\right| > \tau_{\alpha/2}.$$

Application of the technique:—

(a) $\sin^{-1}$ transformation of the square root of the Binomial proportion:-

Let 'r' be the number of successes in 'n' Bernoullian trials with prob. of success 'p'. Then the statistic $r \sim Bin(n, p)$.

$E\left(\frac{r}{n}\right) = p$, $Var\left(\frac{r}{n}\right) = \frac{p(1-p)}{n}$.

Then by CLT, $\dfrac{\frac{r}{n} - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{a}{\sim} N(0,1)$

$$\Rightarrow \sqrt{n}\left(\frac{r}{n} - p\right) \overset{a}{\sim} N\left(0, \sigma^2(p)\right), \text{ as } n \to \infty, \ \sigma^2(p) = p(1-p).$$

Therefore, $\sqrt{n}\left\{g\left(\frac{r}{n}\right) - g(p)\right\} \overset{a}{\sim} N\left(0, [\sigma(p) g'(p)]^2\right)$ as $n \to \infty$.

If we choose the function $g(\cdot) \ni \sigma(p) g'(p) = c$, the asymptotic variance of $g\left(\frac{r}{n}\right)$ will be independent of p.

Hence, the variance stabilizing transformation is obtained by solving the equation

$$\sigma(p) g'(p) = c \Rightarrow g'(p) = \frac{c}{\sigma(p)} \text{ and}$$

$$g(p) = \int \frac{c\,dp}{\sigma(p)} = c \int \frac{dp}{\sqrt{p(1-p)}} \qquad \text{let, } p = \sin^2\theta$$

$$= 2c \int d\theta$$

$$= 2c\theta = 2c \sin^{-1}\sqrt{p}.$$

Choosing $c = \frac{1}{2}$, we get, $g(p) = \sin^{-1}(\sqrt{p})$. Hence, the variance stabilizing transformation is $g\left(\frac{r}{n}\right) = \sin^{-1}\sqrt{\frac{r}{n}}$ and

$$\sqrt{n}\left(\sin^{-1}\sqrt{\frac{r}{n}} - \sin^{-1}\sqrt{p}\right) \overset{a}{\sim} N\left(0, \frac{1}{4}\right) \text{ since } c = \frac{1}{2}.$$

Asymptotically $E\left(\sin^{-1}\sqrt{\frac{r}{n}}\right) = \sin^{-1}\sqrt{p}$ and $Var\left(\sin^{-1}\sqrt{\frac{r}{n}}\right) = \frac{1}{4n}$.

(b) Square Root Transformation of the Mean of a Poisson Sample:—

Let $X_1, \ldots, X_n$ be a r.s. from $P(\lambda)$ popln..

Then $E(\bar{X}) = \lambda$ and $Var(\bar{X}) = \frac{\lambda}{n}$. By CLT, $\dfrac{\bar{X} - \lambda}{\sqrt{\frac{\lambda}{n}}} \overset{a}{\sim} N(0,1)$ as $n \to \infty$.

$\therefore \sqrt{n}(\bar{X} - \lambda) \overset{a}{\sim} N(0, \sigma^2(\lambda) = \lambda)$ and

$\sqrt{n}\left(g(\bar{X}) - g(\lambda)\right) \overset{a}{\sim} N\left(0, [\sigma(\lambda) g'(\lambda)]^2\right)$ as $n \to \infty$.

We wish to determine a function $g(\cdot)$ such that $\sigma(\lambda) g'(\lambda)$ [constant] i.e. the asymptotic variance of $g(\bar{X})$ is independent of λ.

The variance stabilizing transformation is obtained by solving the equation

$$g'(\lambda) = \frac{c}{\sigma(\lambda)} \Rightarrow g(\lambda) = \int \frac{c}{\sigma(\lambda)} d\lambda = \int \frac{c}{\sqrt{\lambda}} d\lambda \text{ as } \sigma^2(\lambda) = \lambda.$$

$$\therefore g(\lambda) = 2c\sqrt{\lambda}.$$

Choose $c = \frac{1}{2}$, $g(\lambda) = \sqrt{\lambda}$. The variance stabilizing transformation is $g(\bar{X}) = \sqrt{\bar{X}}$ and $\sqrt{n}\left\{\sqrt{\bar{X}} - \sqrt{\lambda}\right\} \overset{a}{\sim} N\left(0, \frac{1}{4}\right)$

Therefore, $E\left(\sqrt{\bar{X}}\right) = \sqrt{\lambda}$ and $Var\left(\sqrt{\bar{X}}\right) = \frac{1}{4n}$, asymptotically.

(c) Logarithmic transformation of standard deviation of a normal sample

Let $X_1, X_2, \ldots, X_n$ be a random sample from $N(\mu, \sigma^2)$. Note that the sample variance $s^2 = \frac{1}{(n-1)} \sum_{i=1}^{n} (x_i - \bar{x})^2$ has mean

$E(s^2) = \sigma^2$ and $Var(s^2) = \frac{2\sigma^4}{n-1}$.

More-over, $\dfrac{s^2 - \sigma^2}{\sqrt{\frac{2\sigma^4}{n-1}}} \overset{a}{\sim} N(0,1)$.

$\Rightarrow \sqrt{n-1}\,(s^2 - \sigma^2) \overset{a}{\sim} N(0, 2\sigma^4)$ as $n \to \infty$, then we have

$\sqrt{n-1}\,\{ g(s^2) - g(\sigma^2) \} \overset{a}{\sim} N(0, 2\sigma^4 \{g'(\sigma^2)\}^2)$ as $n \to \infty$.

The variance stabilizing transformation is given by

$$\{g'(\sigma^2)\}^2 2\sigma^4 = c^2, \text{ say.}$$

$$\Rightarrow g'(\sigma^2) = \frac{c}{\sqrt{2\sigma^4}} \Rightarrow g(\sigma^2) = \int \frac{c}{\sqrt{2\sigma^4}} d(\sigma^2) = \frac{c}{\sqrt{2}} \int \frac{d(\sigma^2)}{\sigma^2}$$

$$= \frac{2c}{\sqrt{2}} \int \frac{d\sigma}{\sigma}$$

$$= \sqrt{2}\, c \ln \sigma.$$

Hence, $g(\sigma^2) = \ln \sigma$, choosing $c = \frac{1}{\sqrt{2}}$.

Hence, we have $\sqrt{n-1}\,\{ \ln s - \ln \sigma \} \overset{a}{\sim} N(0, \frac{1}{2})$ as $n \to \infty$.

$$\Rightarrow \ln s \overset{a}{\sim} N\left(\ln \sigma, \frac{1}{2(n-1)}\right) \text{ as } n \to \infty.$$

(Not in syallabi)

(d) ~~tanh~~$^{-1}$ transformation of the sample correlation coefficient based on a sample from a bivariate normal population:—

Let 'r' be the sample correlation coefficient based on n obsns. (X_i, Y_i), $i = 1(1)n$, taken from a bivariate normal population with correlation coefficient 'ρ'

It is known that $E(r) = \rho$, $Var(r) = (1-\rho^2)^2/n$.

Hence, $\dfrac{r - \rho}{\sqrt{\frac{(1-\rho^2)^2}{n}}} \overset{a}{\sim} N(0,1)$ as $n \to \infty$, i.e.

$\sqrt{n}\,(r - \rho) \overset{a}{\sim} N\left[0, \sigma^2(\rho) = (1-\rho^2)^2\right]$ and then

$\sqrt{n}\,\{g(r) - g(\rho)\} \overset{a}{\sim} N\left[0, \{g'(\rho)\sigma(\rho)\}^2\right]$.

We wise to choose the function $g(\cdot)$ such that the asymptotic variance of $g(r)$ is independent of ρ, that is,

$$g'(\rho)\sigma(\rho) = c \Rightarrow g'(\rho) = \frac{c}{\sigma(\rho)} \Rightarrow g(\rho) = c\int \frac{d\rho}{\sigma(\rho)}$$

$$\text{Hence, } g(\rho) = c\int \frac{d\rho}{(1-\rho^2)} = c \cdot \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) = c\tanh^{-1}(\rho).$$

Choosing $c=1$, we get, $g(\rho) = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right) = \tanh^{-1}(\rho)$.

Hence the variance stabillzing transformation is given by

$g(r) = \tanh^{-1}(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ and

$\sqrt{n}\left[\tanh^{-1}(r) - \tanh^{-1}(\rho)\right] \overset{a}{\sim} N(0,1)$.

Therefore, $E\left[\tanh^{-1}(r)\right] = \tanh^{-1}(\rho)$ and $Var\left[\tanh^{-1}(r)\right] = \frac{1}{n}$, asymptotically.

(e) Fisher's Z transformation:- (Based on sample correlation coefficient) We shall study the distribution of

$Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right) = \tanh^{-1}(r)$, through its moments.

Define, $\xi = g(\rho) = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$

$$Z = g(r) = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right).$$

By putting $Z - \xi = Y$, the distn. of Y can be derived from the distn. of r.

$$\mu_1'(Z) = \xi + \frac{\rho}{2(n-1)} + O\left(\frac{1}{n^2}\right)$$

$$\mu_2(Z) = \frac{1}{n-1} + \frac{4-\rho^2}{2(n-1)^2} + O\left(\frac{1}{n^3}\right).$$

$\therefore$ Y may be considered as normal variate with mean $= \frac{\rho}{2(n-1)}$ and variance $= \frac{1}{n-1} + \frac{4-\rho^2}{2(n-1)^2} \simeq \frac{1}{n-3}$.

$$\therefore Z \overset{a}{\sim} N\left(\xi + \frac{\rho}{2(n-1)}, \frac{1}{n-3}\right) \text{ as } n \to \infty.$$

Merits:— The variance stabilizing transformations have two fold merits:—

(i) the transformed statistic $g(T_n)$ tends to normality more rapidly than the original statistic (T_n).

(ii) The transformed statistic has asymptotic variance which is independent of parameter (θ) and no estimation of $\sigma(\theta) = c$ (constant) is required; thus providing better test or confidence interval than the original statistic (T_n).

(I) Approximate tests for the correlation coefficient of a Bivariate Normal population:—

(A) Single sample:— Let 'r' be the sample correlation coefficient based on n observations (X_i, Y_i), $i=1(1)n$ taken from a bivariate normal popln. with correlation coefficient 'ρ'.

(i) General theory:— We have $E(r) \simeq \rho$ and $Var(r) \simeq \frac{(1-\rho^2)^2}{n}$.

$$\text{By CLT, } \frac{r-\rho}{\sqrt{\frac{(1-\rho^2)^2}{n}}} = \frac{\sqrt{n}(r-\rho)}{(1-\rho^2)} \overset{a}{\sim} N(0,1).$$

The sampling distribution of r tends to normality fairly rapidly when ρ is not very different from zero. However, when ρ different widely from zero, this sampling distribution tends to normality so slowly that the use of normal approximation will not be admissable even if n is as large as 100.

To test $H_0: \rho = \rho_0$, we may use the statistic

$$Z = \frac{\sqrt{n}(r-\rho_0)}{1-\rho_0^2} \overset{a}{\sim} N(0,1), \text{ provided } n \text{ is fairly large.}$$

(ii) Use of Variance stabilizing transformation [Use of Fisher Z-transformation]

By Fisher Z-transformation, if $Z = \frac{1}{2}\ln\left(\frac{1+r}{1-r}\right)$ and $\xi = \frac{1}{2}\ln\left(\frac{1+\rho}{1-\rho}\right)$ then

$$Z \overset{a}{\sim} N\left(\xi + \frac{\rho}{2(n-1)}, \frac{1}{n-3}\right)$$

$$\Rightarrow \sqrt{n-3}\left(Z - \xi - \frac{\rho}{2(n-1)}\right) \overset{a}{\sim} N(0,1).$$ This statistic tends to normality even when n is as smaller as 10, although ρ may be widely different from zero.

To test $H_0: \rho = \rho_0$ we may use the statistic

$$\tau = \sqrt{n-3}\left(Z - \xi_0 - \frac{\rho_0}{2(n-1)}\right) \overset{a}{\sim} N(0,1), \text{ under } H_0, \text{ where}$$

$$\xi_0 = \frac{1}{2}\ln\left(\frac{1+\rho_0}{1-\rho_0}\right).$$

If $|\tau| > \tau_{\alpha/2}$, H_0 is rejected against $H_1: \rho \neq \rho_0$ at α level of significance.

(B) Two samples [Use of Fisher Z - transformation]

Let r_1 and r_2 be the sample correlation coefficients in two independent samples of size n_1 and n_2 respectively from two bivariate normal populations with correlation coefficients ρ_1 and ρ_2.

To test the homogeneity of two independent bivariate normal population with common mean and variance i.e. to test $H_0: \rho_1 = \rho_2$ $\boxed{\text{or}}$ the samples arose from two populations with the same correlation coefficient.

Define $Z_i = \frac{1}{2}\ln\left(\frac{1+r_i}{1-r_i}\right)$ and $\xi_i = \frac{1}{2}\ln\left(\frac{1+\rho_i}{1-\rho_i}\right)$.

Then $\sqrt{n_i-3}\left\{Z_i - \xi_i - \frac{\rho_i}{2(n_i-1)}\right\} \overset{a}{\sim} N(0,1)$, $i=1,2$, — independently.

However, under $H_0: \rho_1=\rho_2$ and let $\rho_1=\rho_2=\rho$,

$$E(Z_1-Z_2) = \frac{\rho}{2(n_1-1)} - \frac{\rho}{2(n_2-1)} = \frac{\rho(n_2-n_1)}{2(n_1-1)(n_2-1)} \simeq 0$$

if the sample sizes are not small or if n_1 and n_2 are not very different and $Var(Z_1-Z_2) = \frac{1}{n_1-3}+\frac{1}{n_2-3}$.

To test $H_0: \rho_1=\rho_2$, we may use the statistic

$$\tau = \frac{Z_1 - Z_2}{\sqrt{\frac{1}{n_1-3}+\frac{1}{n_2-3}}} \overset{a}{\sim} N(0,1), \text{ under } H_0.$$

If the observed $|\tau| > \tau_{\alpha/2}$, we reject $H_0: \rho_1=\rho_2$ against $H_1: \rho_1 \neq \rho_2$ at level α.

Scanned by CamScanner

(II) Approximate Tests for Standard Deviation of Normal Populations [Use of Logarithmic Transformation]

(A) Single Sample:- Let s be the S.D. of a r.s. of size 'n' taken from a normal popln. with variance σ^2. By logarithmic transformation,

$$\ln s \overset{a}{\sim} N\left(\ln\sigma, \frac{1}{2n}\right).$$

To test $H_0: \sigma = \sigma_0$, we may use the statistic

$$Z = \frac{\ln s - \ln\sigma_0}{\sqrt{\frac{1}{2n}}} \overset{a}{\sim} N(0,1), \text{ under } H_0: \sigma = \sigma_0.$$

(B) Two Samples:- Let s_i be the SD's of a r.s. taken from a normal popln. with variance σ_i^2, $i=1,2$; the samples are drawn independently. All these values s_1 and s_2 compatible with the hypothesis that the samples arose from two populations having the same variance ?

To test the homogeneity of two independent normal populations with common mean; that is to test $H_0: \sigma_1 = \sigma_2$.

Note that $\ln s_i \overset{a}{\sim} N\left(\ln\sigma_i, \frac{1}{2n_i}\right)$, $i=1,2$ independently.

Under $H_0: \sigma_1 = \sigma_2 = \sigma$(say), $E(\ln s_1 - \ln s_2) = 0$ and $Var(\ln s_1 - \ln s_2) = \frac{1}{2n_1} + \frac{1}{2n_2}$.

To test $H_0: \sigma_1 = \sigma_2$, one may use the statistic

$$Z = \frac{\ln s_1 - \ln s_2}{\sqrt{\frac{1}{2n_1} + \frac{1}{2n_2}}} \overset{a}{\sim} N(0,1), \text{ under } H_0.$$

If the observed $|Z| > \tau_{\alpha/2}$, we reject $H_0: \sigma_1 = \sigma_2$ against $H_1: \sigma_1 \neq \sigma_2$ at level α.

(III) Approximate Tests and Confidence intervals for proportion :—

(A) Single proportion :— Let 'p' be the proportion of members with a characteristic A in a population. Let f be the number of members with characteristic A in a r.s. of size n drawn WR from the population. Then $f \sim Bin(n, p)$, and $\hat{p} = \frac{f}{n}$ is the sample proportion of characteristic A.

$$E(\hat{p}) = E\left(\frac{f}{n}\right) = \frac{np}{n} = p, \quad Var(\hat{p}) = Var\left(\frac{f}{n}\right) = \frac{np(1-p)}{n^2} = \frac{p(1-p)}{n}.$$

(i) General Theory :— By CLT,

$$\frac{\hat{p} - E(\hat{p})}{\sqrt{Var(\hat{p})}} = \frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \overset{a}{\sim} N(0,1) \text{ for large } n.$$

Under the null hypothesis, $H_0: p = p_0$, then the test statistic will be

$$Z = \frac{\hat{p} - p_0}{\sqrt{\frac{p_0(1-p_0)}{n}}} \overset{a}{\sim} N(0,1), \text{ under } H_0.$$

Therefore $H_0: p = p_0$ is rejected against $H_1: p \neq p_0$ at level of significance α if the observed $|Z| > \tau_{\alpha/2}$.

If z_0 is the observed value of Z, then the p-value is $2P[Z > |z_0|]$

Confidence interval :— For large n, the estimate of $\sigma_{\hat{p}}^2 = \frac{p(1-p)}{n}$ is

$$\hat{\sigma}_{\hat{p}}^2 = \frac{\hat{p}(1-\hat{p})}{n} \text{ and we also have}$$

$$\frac{\sqrt{n}(\hat{p} - p)}{\sqrt{\hat{p}(1-\hat{p})}} \overset{a}{\sim} N(0,1).$$

An approximate $100(1-\alpha)\%$ C.I. for p is given by

$$P\left[\left|\frac{\sqrt{n}(\hat{p} - p)}{\sqrt{\hat{p}(1-\hat{p})}}\right| \leq \tau_{\alpha/2}\right] = 1-\alpha.$$

$$\Leftrightarrow P\left[\hat{p} - \tau_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}} \leq p \leq \hat{p} + \tau_{\alpha/2}\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right] = 1-\alpha.$$

(ii) Using Variance stabilizing Transformation :—

$\sin^{-1}$ transformation of square root of Binomial proportion, we have

$$\sin^{-1}\sqrt{\hat{p}} \overset{a}{\sim} N\left(\sin^{-1}\sqrt{p}, \frac{1}{4n}\right).$$

Hence, to test $H_0: p=p_0$, we may use the statistic

$$Z = \frac{\sin^{-1}\sqrt{\hat{p}} - \sin^{-1}\sqrt{p_0}}{\sqrt{\frac{1}{4n}}} \overset{a}{\sim} N(0,1), \text{ under } H_0: p=p_0.$$

We reject $H_0: p=p_0$ ag. $H_1: p\neq p_0$ at level of significance α if the observed $|Z| > \tau_{\alpha/2}$.

Confidence Interval :—

An approximate $100(1-\alpha)\%$ C.I. is given by

$$P\left[\left|\frac{\sin^{-1}\sqrt{\hat{p}} - \sin^{-1}\sqrt{p}}{\sqrt{\frac{1}{4n}}}\right| \leq \tau_{\alpha/2}\right] \simeq 1-\alpha.$$

$$\Leftrightarrow P\left[\sin^{-1}\sqrt{\hat{p}} - \tau_{\alpha/2}\cdot\sqrt{\frac{1}{4n}} \leq \sin^{-1}\sqrt{p} \leq \sin^{-1}\sqrt{\hat{p}} + \tau_{\alpha/2}\sqrt{\frac{1}{4n}}\right] \simeq 1-\alpha.$$

$$\Leftrightarrow P\left[\sin^2\left(\sin^{-1}\sqrt{\hat{p}} - \tau_{\alpha/2}\cdot\sqrt{\frac{1}{4n}}\right) \leq p \leq \sin^2\left(\sin^{-1}\sqrt{\hat{p}} + \tau_{\alpha/2}\sqrt{\frac{1}{4n}}\right)\right] \simeq 1-\alpha.$$

B. Two Proportions :—

Let p_1 and p_2 be two proportions of characteristic A in two populations. Let random samples of sizes n_1 and n_2, respectively, be obtained from the first and the second population through independent drawings. Let f_1, f_2 be the numbers of members with characteristic A in the random samples.

Then $f_1 \sim Bin(n_1, p_1)$ and

$f_2 \sim Bin(n_2, p_2)$, independently.

Then $\hat{p}_1 = \frac{f_1}{n_1}$ and $\hat{p}_2 = \frac{f_2}{n_2}$ be the two sample proportions of A.

We are to test the equality of two popln proportions, i.e. $H_0: p_1=p_2$.

General Theory:-

Note that $E(\hat{p}_1 - \hat{p}_2) = E(\hat{p}_1) - E(\hat{p}_2) = p_1 - p_2$ and

$Var(\hat{p}_1 - \hat{p}_2) = Var(\hat{p}_1) + Var(\hat{p}_2)$, since $\hat{p}_1$ and $\hat{p}_2$ are indep.

$$= \frac{p_1(1-p_1)}{n_1} + \frac{p_2(1-p_2)}{n_2}.$$

Hence, by CLT, $\dfrac{(\hat{p}_1 - \hat{p}_2) - (p_1 - p_2)}{\sqrt{\dfrac{p_1(1-p_1)}{n_1} + \dfrac{p_2(1-p_2)}{n_2}}} \overset{a}{\sim} N(0,1)$, for large n_1, n_2.

To test the hypothesis $H_0: p_1 = p_2$. Under H_0, let $p_1 = p_2 = p$ (unknown), then $E(\hat{p}_1 - \hat{p}_2) = 0$ and $Var(\hat{p}_1 - \hat{p}_2) = p(1-p)\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}$.

Here p is unknown and has to be estimated from the random samples.

Under H_0; $p_1 = p_2 = p$, we have $E(f_1 + f_2) = (n_1 + n_2)p$,

$$\Rightarrow E\left(\frac{f_1 + f_2}{n_1 + n_2}\right) = p.$$

$$\therefore \hat{p} = \frac{f_1 + f_2}{n_1 + n_2} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

To test $H_0: p_1 = p_2$, we may use the statistic

$$Z = \frac{\hat{p}_1 - \hat{p}_2}{\sqrt{\hat{p}(1-\hat{p})\left\{\frac{1}{n_1} + \frac{1}{n_2}\right\}}} \overset{a}{\sim} N(0,1), \text{ under } H_0.$$

If the observed $Z > \tau_\alpha$, we reject $H_0: p_1 = p_2$ ag. $H_1: p_1 > p_2$ at level of significance α, etc.

<ii> Using Variance-stabilizing Transformation:-

Using $\sin^{-1}$ transformation of square root of Binomial proportion, we have

$$\sin^{-1}\sqrt{\hat{p}_1} \overset{a}{\sim} N\left(\sin^{-1}\sqrt{p_1}, \frac{1}{4n_1}\right)$$

$$\& \sin^{-1}\sqrt{\hat{p}_2} \overset{a}{\sim} N\left(\sin^{-1}\sqrt{p_2}, \frac{1}{4n_2}\right), \text{ independently,}$$

since $\hat{p}_1$ and $\hat{p}_2$ are independent.

Hence, $\sin^{-1}\sqrt{\hat{p}_1} - \sin^{-1}\sqrt{\hat{p}_2} \overset{a}{\sim} N\left(\sin^{-1}\sqrt{p_1} - \sin^{-1}\sqrt{p_2}, \frac{1}{4n_1} + \frac{1}{4n_2}\right)$

Under $H_0: p_1 = p_2$, $\sin^{-1}\sqrt{\hat{p}_1} - \sin^{-1}\sqrt{\hat{p}_2} \overset{a}{\sim} N\left(0, \frac{1}{4n_1} + \frac{1}{4n_2}\right)$.

To test $H_0: p_1 = p_2$, we may use the statistic

$$Z = \frac{\sin^{-1}\sqrt{\hat{p}_1} - \sin^{-1}\sqrt{\hat{p}_2}}{\sqrt{\frac{1}{4n_1} + \frac{1}{4n_2}}} \overset{a}{\sim} N(0,1), \text{ under } H_0.$$

Confidence Interval:- An approximate $100(1-\alpha/2)\%$ C.I. for p_i is given by,

$$P\left[\sin^2\left(\sin^{-1}\sqrt{\hat{p}_i} - \frac{\tau_{\alpha/4}}{2\sqrt{n_i}}\right) \le p_i \le \sin^2\left(\sin^{-1}\sqrt{\hat{p}_i} + \frac{\tau_{\alpha/4}}{2\sqrt{n_i}}\right)\right] = 1 - \frac{\alpha}{2}$$

(IV) Approximate Tests and confidence intervals for Poisson Parameter :-

(A) Single sample:- Let $X_1, \ldots, X_n$ be a r.s. from $P(\lambda)$. Note that

$$E\left(\sum_{i=1}^{n} X_i\right) = n\lambda \text{ and } Var\left(\sum_{i=1}^{n} X_i\right) = n\lambda.$$

$$\frac{\sum_{i=1}^{n} X_i - E\left(\sum_{i=1}^{n} X_i\right)}{\sqrt{Var\left(\sum_{i=1}^{n} X_i\right)}} = \frac{n\bar{X} - n\lambda}{\sqrt{n\lambda}} = \frac{\sqrt{n}(\bar{X}-\lambda)}{\sqrt{\lambda}} \overset{a}{\sim} N(0,1) \text{ as } n \to \infty.$$

Therefore, the test for λ, should be based on the sufficient statistic $\bar{X} \overset{a}{\sim} N\left(\lambda, \frac{\lambda}{n}\right)$.

(i) General theory:- $\because \frac{\left(\sum_{i=1}^{n} X_i - n\lambda\right)}{\sqrt{n\lambda}} \overset{a}{\sim} N(0,1)$, as $n \to \infty$.

$$\Rightarrow \frac{\sqrt{n}(\bar{X}-\lambda)}{\sqrt{\lambda}} \overset{a}{\sim} N(0,1) \text{ as } n \to \infty.$$

To test $H_0: \lambda = \lambda_0$, we may use the statistic

$$Z = \frac{\sqrt{n}(\bar{X}-\lambda_0)}{\sqrt{\lambda_0}} \overset{a}{\sim} N(0,1), \text{ under } H_0.$$

$\therefore$ We reject $H_0: \lambda = \lambda_0$ Vs. $H_1: \lambda \neq \lambda_0$ at level of significance α, if the observed $|Z| > \tau_{\alpha/2}$.

(ii) Using Variance Stabilizing Transformation:-

By square root transformation of Poisson mean, we have

$$\sqrt{\bar{X}} \overset{a}{\sim} N\left(\sqrt{\lambda}, \frac{1}{4n}\right), \text{ for large } n.$$

To test $H_0: \lambda = \lambda_0$, we may use the statistic

$$Z = \frac{\sqrt{\bar{X}} - \sqrt{\lambda_0}}{\sqrt{\frac{1}{4n}}} \overset{a}{\sim} N(0,1), \text{ under } H_0.$$

If the observed $|Z| > \tau_{\alpha/2}$, we reject $H_0: \lambda = \lambda_0$ against $H_1: \lambda \neq \lambda_0$ at level α.

Confidence Interval:- An approximate $100(1-\alpha)\%$ C.I. for λ is given by:

$$P\left[-\tau_{\alpha/2} \leq \frac{\sqrt{\bar{X}} - \sqrt{\lambda}}{\sqrt{\frac{1}{4n}}} \leq \tau_{\alpha/2}\right] = 1-\alpha.$$

$$\Leftrightarrow P\left[\left(\sqrt{\bar{X}} - \frac{\tau_{\alpha/2}}{2\sqrt{n}}\right)^2 \leq \lambda \leq \left(\sqrt{\bar{X}} + \frac{\tau_{\alpha/2}}{2\sqrt{n}}\right)^2\right] = 1-\alpha.$$

Hence the observed value of $\left[\left(\sqrt{\bar{X}} - \frac{\tau_{\alpha/2}}{2\sqrt{n}}\right)^2, \left(\sqrt{\bar{X}} + \frac{\tau_{\alpha/2}}{2\sqrt{n}}\right)^2\right]$ is a C.I. for λ with confidence coefficient $(1-\alpha)$.

(B) Two samples:— Let $x_{11}, x_{12}, \ldots, x_{1n_1}$ be a r.s. from $P(\lambda_1)$ and $x_{21}, x_{22}, \ldots, x_{2n_2}$ be a r.s. from $P(\lambda_2)$, drawing independently.
We are to test the homogeneity of two independent Poisson distributions.

(i) General Theory:— By CLT, $\bar{x}_1 \overset{a}{\sim} N\left(\lambda_1, \frac{\lambda_1}{n_1}\right)$ and $\bar{x}_2 \overset{a}{\sim} N\left(\lambda_2, \frac{\lambda_2}{n_2}\right)$, independently.

Hence, $\bar{x}_1 - \bar{x}_2 \overset{a}{\sim} N\left(\lambda_1 - \lambda_2, \frac{\lambda_1}{n_1} + \frac{\lambda_2}{n_2}\right)$.

Under $H_0: \lambda_1 = \lambda_2$, let $\lambda_1 = \lambda_2 = \lambda$ (unknown)

$\therefore E(\bar{x}_1 - \bar{x}_2) = 0$ and $Var(\bar{x}_1 - \bar{x}_2) = \lambda\left(\frac{1}{n_1} + \frac{1}{n_2}\right)$

$$E\left(\sum_{i=1}^{n_1} x_{1i} + \sum_{j=1}^{n_2} x_{2j}\right) = (n_1 + n_2)\lambda$$

i.e. $$\hat{\lambda} = \frac{n_1 \bar{x}_1 + n_2 \bar{x}_2}{n_1 + n_2}.$$

To test, $H_0: \lambda_1 = \lambda_2$, we may use the statistic

$$Z = \frac{(\bar{x}_1 - \bar{x}_2)}{\sqrt{\hat{\lambda}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}} \overset{a}{\sim} N(0,1) \text{ as } n \to \infty.$$

We reject $H_0: \lambda_1 = \lambda_2$ vs. $H_1: \lambda_1 \neq \lambda_2$ at α level of significance if the observed $|Z| > \tau_{\alpha/2}$.

(ii) Using Variance Stabilizing Transformation:—

By square root transformation of Poisson mean, we have
$\sqrt{\bar{x}_1} \overset{a}{\sim} N\left(\sqrt{\lambda_1}, \frac{1}{4n_1}\right)$ and $\sqrt{\bar{x}_2} \overset{a}{\sim} N\left(\sqrt{\lambda_2}, \frac{1}{4n_2}\right)$ independently.

Hence $\sqrt{\bar{x}_1} - \sqrt{\bar{x}_2} \overset{a}{\sim} N\left(\sqrt{\lambda_1} - \sqrt{\lambda_2}, \frac{1}{4n_1} + \frac{1}{4n_2}\right)$.

Under $H_0: \lambda_1 = \lambda_2$, $\left(\sqrt{\bar{x}_1} - \sqrt{\bar{x}_2}\right) \sim N\left(0, \frac{1}{4n_1} + \frac{1}{4n_2}\right)$.

Hence, we use the statistic $$Z = \frac{\sqrt{\bar{x}_1} - \sqrt{\bar{x}_2}}{\sqrt{\frac{1}{4n_1} + \frac{1}{4n_2}}}.$$

To test $H_0: \lambda_1 = \lambda_2$, If the observed $|Z| > \tau_{\alpha/2}$, we reject $H_0: \lambda_1 = \lambda_2$ ag. $H_1: \lambda_1 \neq \lambda_2$ at level of significance α.

Confidence Interval:— Let an estimate of $\sqrt{\lambda}$ is, $\widehat{\sqrt{\lambda}} = \frac{n_1\sqrt{\bar{x}_1} + n_2\sqrt{\bar{x}_2}}{n_1 + n_2}$,

$$E(\widehat{\sqrt{\lambda}}) = \sqrt{\lambda}, \quad Var(\widehat{\sqrt{\lambda}}) = \frac{1}{4(n_1+n_2)}$$

$$\therefore \frac{\widehat{\sqrt{\lambda}} - \sqrt{\lambda}}{\sqrt{\frac{1}{4(n_1+n_2)}}} \overset{a}{\sim} N(0,1) \text{ as } n_1, n_2 \to \infty.$$

An $100(1-\alpha)\%$ CI for λ is given by $P\left[\left|\frac{\widehat{\sqrt{\lambda}} - \sqrt{\lambda}}{\sqrt{\frac{1}{4(n_1+n_2)}}}\right| \leq \tau_{\alpha/2}\right] = 1-\alpha$.

$$\Rightarrow P\left[\left(\widehat{\sqrt{\lambda}} - \tau_{\alpha/2} \cdot \frac{1}{2\sqrt{n_1+n_2}}\right)^2 < \lambda < \left(\widehat{\sqrt{\lambda}} + \tau_{\alpha/2} \cdot \frac{1}{2\sqrt{n_1+n_2}}\right)^2\right] = 1-\alpha.$$

PEARSONIAN χ^2- distribution [Frequency χ^2]

Let us suppose that the elements of an infinite population are divided into k mutually exclusive classes or categories and the probability that an individual falling in the i^{th} class is p_i, $i=1(1)k$. where $\sum_{i=1}^{k} p_i = 1$. A r.s. of size n is drawn from this population and it is found that f_i members in the sample belong to the i^{th} class, $i=1(1)k$.

The probability of obtaining f_i members from the i^{th} class, $i=1(1)k$, in a random sample of size 'n' is given by the multinomial distribution is

$$p(f_1, \ldots, f_k) = \frac{\lfloor n}{\lfloor f_1 \lfloor f_2 \ldots \lfloor f_k} p_1^{f_1} p_2^{f_2} \ldots p_k^{f_k}, \text{ where } \sum_{i=1}^{k} f_i = n$$

and $\sum_{i=1}^{k} p_i = 1$. We assume that the probability distribution given by $p(f_1, \ldots, f_k)$ is completely known.

Here, the observed frequency in the i^{th} class is f_i and the expected frequency is $E(f_i) = np_i$, $i=1(1)k$.

As a measure of discrepency between the observed frequencies and expected frequencies, Karl Pearson suggested the following statistic:

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

In the statistical literature, this statistic is refered to as a Pearsonian χ^2 or frequency χ^2.

Derivation of distribution of Pearsonian χ^2 statistic, for large samples:

If the sample size n is sufficiently large so that f_i's are not small, then using Stirling's approximation to factorials, we have

$$p(f_1, \ldots, f_k) = \frac{\sqrt{2\pi}\, e^{-n} \cdot n^{n+1/2}}{\prod_{i=1}^{k} \left\{ \sqrt{2\pi}\, e^{-f_i} f_i^{f_i + 1/2} \right\}} \cdot p_1^{f_1} p_2^{f_2} \ldots p_k^{f_k}$$

$$= c \cdot \prod_{i=1}^{k} \left(\frac{np_i}{f_i} \right)^{f_i + 1/2}, \text{ where } c \text{ is a constant, independent of } f_i.$$

Define $x_i = \frac{f_i - np_i}{\sqrt{np_i}} = \frac{f_i - e_i}{\sqrt{e_i}}$, where $e_i = np_i$, so that $\chi^2 = \sum_{i=1}^{k} x_i^2$.

Now, $\ln\left(\frac{p(f_1,\cdots,f_k)}{C}\right) \simeq \sum_{i=1}^{k}\left(f_i+\frac{1}{2}\right)\log\left(\frac{e_i}{f_i}\right)$

$$\simeq \sum_{i=1}^{k}\left(e_i+\frac{1}{2}+x_i\sqrt{e_i}\right)\log\left(1+\frac{x_i}{\sqrt{e_i}}\right)^{-1}$$

$$\simeq -\sum_{i=1}^{k}\left(e_i+\frac{1}{2}+x_i\sqrt{e_i}\right)\left\{\frac{x_i}{\sqrt{e_i}}-\frac{x_i^2}{2e_i}+\cdots\right\}$$

[if $e_i = np_i$ is large, then x_i will be smaller compared to $\sqrt{e_i}$, then the expansion of $\log\left(1+\frac{x_i}{\sqrt{e_i}}\right)$ is valid]

$$\simeq -\frac{1}{2}\sum_{i=1}^{k}x_i^2-\sum_{i=1}^{k}x_i\sqrt{e_i}+O\left(\frac{1}{\sqrt{e_i}}\right)$$

$$\simeq -\frac{1}{2}\sum_{i=1}^{k}x_i^2+O\left(\frac{1}{\sqrt{e_i}}\right)$$

[Since, $\sum_{i=1}^{k}x_i\sqrt{e_i} = \sum_{i=1}^{k}(f_i-e_i) = \sum_{i=1}^{k}f_i - n\sum_{i=1}^{k}p_i = 0$]

Since n is large and $e_i = np_i$ are large, so $O\left(\frac{1}{\sqrt{e_i}}\right)\to 0$ as $n\to\infty$. Therefore $p(f_1,\cdots,f_k)\simeq C.e^{-\frac{1}{2}\sum_{i=1}^{k}x_i^2}$ and this shows that x_i's are distributed as independent standard normal variates in large samples. Hence the distribution of $\chi^2=\sum_{i=1}^{k}x_i^2=\sum_{i=1}^{k}\frac{(f_i-np_i)^2}{np_i}$ is that of the sum of squares of k standard normal variates subject to one homogeneous linear constraint $\sum_{i=1}^{k}x_i\sqrt{e_i}=0$, i.e. χ^2 distn. with $(k-1)$ d.f., approximately for large n.

Heuristic Proof :- The multinomial PMF $p(f_1, f_2, \ldots, f_k)$ can be written as

$$\frac{e^{-np_1} \cdot \frac{(np_1)^{f_1}}{\lfloor f_1} \cdots\cdots e^{-np_k} \cdot \frac{(np_k)^{f_k}}{\lfloor f_k}}{e^{-n(p_1+p_2+\cdots+p_k)} \cdot \frac{\{n(p_1+p_2+\cdots+p_k)\}^n}{\lfloor n}}, \text{ is}$$

equivalent to the conditional PMF of k poisson variates f_i with parameter $\lambda_i = np_i$, $i=1(1)k$ given that $\sum_{i=1}^{k} f_i = n$.

If each individual cell frequency is large, then $x_i = \frac{f_i - np_i}{\sqrt{np_i}}$ is approximately a $N(0,1)$ variate if np_i is sufficiently large.

Then $\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i)^2}{np_i} = \sum_{i=1}^{k} x_i^2$ approximately comes out to be the sum of squares of k $N(0,1)$ variates subject to the linear constraint $\sum_{i=1}^{k} \sqrt{np_i}\, x_i = \sum_{i=1}^{k} (f_i - np_i) = 0$.

Hence the χ^2 statistic follows approximately χ^2 distn. with $(k-1)$ d.f.

Remarks :- In order to apply χ^2 approximation of the statistic the following conditions must be satisfied :

(i) The members of the sample should be independent.

(ii) Constraints on cell frequencies, if any, should be linear.

(iii) Since, a discrete distribution is being approximated by a continuous χ^2- distribution, which is valid in large sample 'n' should be sufficiently large, say, ≥ 50 and, all expected frequencies $e_i = np_i \geq 5$. If some of $np_i = e_i$ are < 5, it is advisable to pool the smaller groups, so that every group contains at least 5 expected frequency, the χ^2 approximation to the statistic is applied.

(iv) Since the distribution of the Pearsonian χ^2 does not depend on the form of the distribution of the population from which the sample has been drawn, any test based on the statistic may be regarded as a non-parametric test.

Application of Pearsonian χ^2 statistic :—

We shall see presently how this statistic may be used to solve various problems in hypothesis-testing, in large samples.

A χ^2-test for Goodness of fit :— We now proceed to study the problem of testing the agreement between a probability distn. and actual observations. We assume here that the popln. follows a probability distribution which is completely specified. Here one wishes to verify whether the data form a random sample from the given popln. distn..

Let us suppose that the elements of the popln. are divided into k mutually exclusive classes and the probability, evaluated under the assumed probability distribution, that an individual falling in the i^{th} class is p_i^0, $i=1(1)k$. If f_i is the frequency in the i^{th} class in a random sample of size n, then

$$(f_1, f_2, \dots, f_k) \sim \text{Multinomial}(n, p_1^0, \dots, p_k^0) \text{ where}$$

$$\sum_{i=1}^{k} f_i = n \text{ and } \sum_{i=1}^{k} p_i^0 = 1. \text{ Note that } E(f_i) = np_i^0, \; i=1(1)k.$$

Since our task here is to see how well the expected frequencies np_i^0's are in agreement with or how well they fit the observed frequencies f_i, as a measure of goodness between the observed and expected frequencies, we may use Pearsonian χ^2-statistic :

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - np_i^0)^2}{np_i^0} = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i} = \sum_{i=1}^{k} \frac{O_i^2}{E_i} - n,$$

which approximately follows χ^2 distn. with d.f. $(k-1)$.

If the fit between the observed and expected frequencies are good, the value of χ^2-statistic would be small. The greater the differences between the observed and the expected frequencies np_i, under the hypothesis, the larger will be the value χ^2-statistic. Hence a very high value of χ^2-statistic should indicate falsity of the given hypothesis. If the observed $\chi^2 > \chi^2_{\alpha, k-1}$, we say that our sample shows a significant deviation from the hypothetical population distribution and we shall reject the hypothesis, at least until further data available.

χ^2-test for goodness of fit when some parameters of hypothetical population are unknown: —

This is the more useful form of the problem of testing for goodness of fit. Suppose that the distribution function $F(x; \alpha_1, \ldots, \alpha_s)$ containing 's' unknown parameters $\alpha_1, \ldots, \alpha_s$ but otherwise of known mathematical form.

Then $P[X \in A_i] = p_i = p_i(\alpha_1, \ldots, \alpha_s)$, depends on the parameter $\alpha_1, \alpha_2, \ldots, \alpha_s$ and the parameters are estimated from the sample. If the estimators of p_i are denoted by $\hat{p}_i = p_i(\hat{\alpha}_1, \ldots, \hat{\alpha}_s)$ then the Pearsonian χ^2 statistic reduces to

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - n\hat{p}_i)^2}{n\hat{p}_i} = \sum_{i=1}^{k} \frac{f_i^2}{n\hat{p}_i} - n.$$

The sampling distn. of χ^2-statistic will more or less depend on the method of estimation chosen.

Therefore $$\chi^2 = \sum_{i=1}^{k} \frac{(f_i - n\hat{p}_i)^2}{n\hat{p}_i} \sim \chi^2_{k-1-s}, \text{ approximately as } n \to \infty.$$

[As the estimation of each parameter imposes a homogeneous linear constraints on the (approximate) standard normal variables $\frac{f_i - n\hat{p}_i}{\sqrt{n\hat{p}_i}}$ and as there are 's' parameters we have 's' linear constraints on $\frac{f_i - n\hat{p}_i}{\sqrt{n\hat{p}_i}}$.]

☒ It is infact, only necessary to reduce the number of d.f. of the limeting distribution of $\chi^2 = \sum_{i=1}^{k} \frac{(f_i - n\hat{p}_i)^2}{n\hat{p}_i}$ by one unit for each independent parameters estimated from the sample.

[B] χ^2 test as a test of independence in a contingency table :—

Suppose that the individuals of a population be classified according two categorical variables A and B into r and s classes respectively, say $A_1, A_2, \ldots, A_r$
$B_1, B_2, \ldots, B_s$

Let f_{ij} be the observed frequency in the class A_iB_j in a random sample of size 'n' from the population and the probability that an individual is in the class A_iB_j is p_{ij}.

Note that $\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} = n$ and $\sum_{i=1}^{r}\sum_{j=1}^{s} p_{ij} = 1$. Again $\sum_{j=1}^{s} f_{ij} = f_{io}$ and $\sum_{j=1}^{s} p_{ij} = p_{io}$, are the marginal frequency and marginal probability of A_i, $i=1(1)r$. Similarly, $\sum_{i=1}^{r} f_{ij} = f_{oj}$ and $\sum_{i=1}^{r} p_{ij} = p_{oj}$ are the marginal frequency and marginal probability of B_j, $j=1(1)s$.

We want to test whether A and B are independent; i.e. to test

$H_0 : p_{ij} = p_{io} \times p_{oj} \ \forall (i,j)$.

If the probabilities p_{ij} of the cells in the contingency table are assigned, then to test the hypothesis that the data are in agreement with these hypothetical probabilities p_{ij}, the pearsonian χ^2-statistic

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(f_{ij} - np_{ij})^2}{np_{ij}}$$

can be used and $\chi \sim \chi^2_{rs-1}$ distribution, the only restriction being $\sum_{i=1}^{r}\sum_{j=1}^{s} f_{ij} = n$, under H_0, we have $p_{ij} = p_{io}p_{oj}$, $\forall (i,j)$ and the statistic reduces to

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(f_{ij} - np_{io}p_{oj})^2}{np_{io}p_{oj}}.$$

Here p_{io}'s and p_{oj}'s are unknown and are to be estimated from the sample.

There are exactly $(r+s)$ parameters p_{io} and p_{oj}, with $\sum_{i=1}^{r} p_{io} = \sum_{j=1}^{s} p_{oj} = 1$; i.e. $(r+s-2)$ independent parameters which are to be estimated from the sample.

By maximum likelihood method, we get

$$\hat{p}_{io} = \frac{f_{io}}{n}, \quad \hat{p}_{oj} = \frac{f_{oj}}{n}.$$

Hence, under H_0,

$$\chi^2 = \sum_{i=1}^{r}\sum_{j=1}^{s} \frac{(f_{ij} - n\hat{p}_{io}\hat{p}_{oj})^2}{n\hat{p}_{io}\hat{p}_{oj}} = n\sum_{i=1}^{r}\sum_{j=1}^{s} \frac{\left(f_{ij} - \frac{f_{io}f_{oj}}{n}\right)^2}{f_{io}f_{oj}}$$

$$= n\left\{\sum_i\sum_j \frac{f_{ij}^2}{f_{io}f_{oj}} - 1\right\}$$, tends to χ^2-distn. with

d.f. $(rs-1) - (r+s-2) = (r-1)(s-1)$;

Hence the statistic χ^2 measures the departure from independence, we reject H_0 if observed $\chi^2 > \chi^2_{\alpha;(r-1)(s-1)}$ at level of significance α, provided n is large enough.

$\boxed{c}$. χ^2 test of Homogeneity of Parallel samples:-

[Test of l samples arisen from the same population]

consider samples ($l \geq 2$) from l independent multinomial popln., the prob. that an object in the jth population belong to the ith class is p_{ij}, $j=1(1)l$, $i=1(1)k$, then the hypothesis to be tested is H_0: $\{p_{ij} = p_{io} \ \forall\ j=1(1)l\} \ \forall\ i=1(1)k$.

If p_{ij}'s are known, n_i's are sufficiently large, the χ^2 statistic of departure of the frequencies from their expected frequencies is

$$\chi^2 = \sum_{i=1}^{k} \frac{(f_{i1} - n_1p_{i1})^2}{n_1p_{i1}} + \cdots + \sum_{i=1}^{k} \frac{(f_{il} - n_lp_{il})^2}{n_lp_{il}}$$

$$= \sum_{j=1}^{l}\sum_{i=1}^{k} \frac{(f_{ij} - n_jp_{ij})^2}{n_jp_{ij}}$$

which has $l(k-1)$ d.f. and $\overset{a}{\sim} \chi^2_{l(k-1)}$.

Under H_0, the MLE of p_{io} is $\hat{p}_{io} = \frac{1}{n}\sum_{j=1}^{l} f_{ij}$ and $\sum_{i=1}^{k} p_{io} = 1$.

If this value $\hat{p}_{io}$ is substituted for p_{io}, under H_0, then—

$$\chi^2 = \sum_{j=1}^{l}\sum_{i=1}^{k} \frac{(f_{ij} - n_j\hat{p}_{io})^2}{n_j\hat{p}_{io}} = \sum_{j=1}^{l}\sum_{i=1}^{k} \frac{\left(f_{ij} - n_j\cdot\frac{f_{io}}{n}\right)^2}{n_j\cdot\frac{f_{io}}{n}}$$

$$\overset{a}{\sim} \chi^2_{l(k-1)-k-1}$$

Hence our test statistic is $\chi^2 = \sum_{j=1}^{l}\sum_{i=1}^{k} \frac{\left(f_{ij} - n_j\frac{f_{io}}{n}\right)^2}{n_j\frac{f_{io}}{n}}$

$\overset{a}{\sim} \chi^2_{(l-1)(k-1)}$ for large samples under H_0.

Combination of Probabilities from independent tests of significance :-

Sometimes, we have a number of independent tests of significance for the same hypothesis, giving different probabilities for the statistic to be more extreme in the direction of the alternative, under the null hypothesis.

For right tailed test, we find the p-value, $p = P_{H_0}[T > t]$, where T is the test statistic and t is its observed value.

For left tailed test and both tailed test, the p-values are

$P[T < t]$ and $P[|T| > |t|]$.

Among the different independent tests, some of which may give significant and others non-significant results. Our problem is to combine on pool various probabilities, to get a single probability, from which to decide on the significance of the aggregate of these tests.

Let $f_T(t)$ be the PDF of T. For left tailed test, p value is

$$p = P[T < t] = \int_{-\infty}^{t} f_T(t)\, dt = F_T(t) \sim R(0,1).$$

Hence, $-2\log_e p \sim \chi^2_{(2)}$.

Therefore for k independent tests of significance, the p-values are $p_1, p_2, \ldots, p_k$ and $-2\log_e p_i \overset{iid}{\sim} \chi^2_{(2)}$, $i = 1(1)k$.

Then $P_\lambda = \sum_{i=1}^{k} (-2\log_e p_i) \sim \chi^2_{(2k)}$. If the observed P_λ exceeds $\chi^2_{\alpha, 2k}$, the combined result will be said to be significant and we shall finally reject the null hypothesis at $100\alpha\%$ level of significance.

It is also important to note that, for right-tailed and both-tailed tests the p-values are $P[T > t]$ and $P[|T| > |t|]$. Here $p' = P[T > t] = 1 - F_T(t) \sim R(0,1)$ and $p'' = P[|T| > |t|] = 1 - F_{|T|}(|t|) \sim R(0,1)$. Therefore, we also have $\sum_{i=1}^{k} -2\log_e p_i'$ and $\sum_{i=1}^{k} -2\log_e |p_i''|$ as $\chi^2_{(2k)}$ r.v.

<u>Ex.</u> To test a hypothesis H_0, we use $(p+q+r)$ different test statistics $T_1, T_2, \ldots, T_{p+q+r}$. Suppose the tests based on $T_1, \ldots, T_p$ are left-tailed tests, $T_{p+1}, \ldots, T_{p+q}$ are right tailed tests and $T_{p+q+1}, \ldots, T_{p+q+r}$ are two-tailed tests. How can we combine the results of these $p+q+r$ tests to get an overall decision? State clearly the assumptions you have made.

<u>Hints:-</u> The p-value of the test based on the statistic T_i is

$$p_i = \begin{cases} P[T_i < t_i], & i = 1(1)p. \\ P[T_i > t_i], & i = \overline{p+1}(1)\overline{p+q} \\ P[|T_i| > |t_i|], & i = \overline{p+q+1}(1)\overline{p+q+r} \end{cases}$$

$\sim R(0,1) \;\forall\; i = 1(1)\overline{p+q+r}$.

Then $P_\lambda = \sum_{i=1}^{p+q+r} (-2\ln p_i) \sim \chi^2_{2(p+q+r)}$, etc.

Test for Independence in 2X2 contingency Table :—

suppose that the two classes of attributes A and B are denoted by A_1, A_2 and B_1, B_2 respectively. Also, let the observed distn. of a random sample of N observations are given below:

Table : 2X2 contingency table

$B \backslash A$	A_1	A_2	Total
B_1	a	c	a+c
B_2	b	d	b+d
Total	a+b	c+d	N

To test the independence of the attributes A and B. Here

$$H_0 : p_{ij} = p_{i0} \times p_{0j} \quad \forall (i,j).$$

(a) Fisher's Exact probability Test :—

The fisher's exact probability test in an extremely useful non-parametric technique for analysing 2x2 contingency table when each cell frequencies are small.

Assuming the independence of the attributes, the exact probability of observing a particular set of frequencies in a 2x2 table, when the marginal totals are regarded as fixed, is given by

$$p_a = \frac{\binom{a+c}{a}\binom{b+d}{b}}{\binom{N}{a+b}} = \frac{\lfloor\underline{a+b}\ \lfloor\underline{c+d}\ \lfloor\underline{a+c}\ \lfloor\underline{b+d}}{\lfloor\underline{N}\ \lfloor\underline{a}\ \lfloor\underline{b}\ \lfloor\underline{c}\ \lfloor\underline{d}}.$$

For a statistical test of the null hypothesis, we compute the 'p'value' = Probality under Ho of observed table or of one even more extreme. If 'p-value' is less than α, then we shall reject Ho at 100α % level of significance.

If a is the smallest cell value in the table then the 'p-value' is

$$\sum_{x=0}^{a} p_x = \sum_{x=0}^{a} \frac{\lfloor\underline{a+b}\ \lfloor\underline{c+d}\ \lfloor\underline{a+c}\ \lfloor\underline{b+d}}{\lfloor\underline{N}\ \lfloor\underline{x}\ \lfloor\underline{a+b-x}\ \lfloor\underline{a+c-x}\ \lfloor\underline{d-a+x}}$$

For more extreme cases we have the 2X2 contingency table as

			Total
	x	a+c−x	a+c
	a+b−x	d−a+x	b+d
Total	a+b	c+d	N

with fixed marginals.

(b) Pearsonian χ^2-test :-

Under H_0 : $p_{ij} = p_{io} \times p_{oj} \quad \forall \; i,j = 1,2.$

The χ^2-statistic for 2×2 contingency table become

$$\chi^2 = \sum_{i=1}^{2}\sum_{j=1}^{2} \frac{(f_{ij} - n\hat{p}_{oi}\hat{p}_{oj})^2}{n\hat{p}_{oi}\hat{p}_{oj}}$$

$$= \frac{\left\{a - \frac{(a+b)(a+c)}{N}\right\}^2}{\frac{(a+b)(a+c)}{N}} + \frac{\left\{b - \frac{(a+b)(b+d)}{N}\right\}^2}{\frac{(a+b)(b+d)}{N}}$$

$$+ \frac{\left\{c - \frac{(a+c)(c+d)}{N}\right\}^2}{\frac{(a+c)(c+d)}{N}} + \frac{\left\{d - \frac{(c+d)(b+d)}{N}\right\}^2}{\frac{(c+d)(b+d)}{N}}$$

$$= \frac{(ad - bc)^2 N}{(a+b)(a+c)(c+d)(b+d)} \sim \chi_1^2 \text{, asymptotically.}$$

As in the case of general contingency table the above χ^2 provides an approximate test of independence of two attributes.

Yate's correction for continuity in a 2X2 contingency table :-

The distn. of χ^2 is a continuous distribution while the distribution of frequencies is by its very nature discontinuous. The continuous χ^2 distn., may be regarded as the limit to which the true discontinuous distn. tends as the sample size increases. One of the reasons why we assume that all cell frequencies are greater than 5, is to avoid irregularities due to this continuity.

If any one of the theoretical cell frequencies is less than 5, then use of pooling method for χ^2-test result in χ^2 with 0 d.f. which is meaningless.

In this case, we apply a correction, due to Yates which is known as Yate's correction for continuity.

The modification suggested by Yates for small cell frequencies compensates for the difference between the discrete distribution of cell frequencies and the approximate continuous χ^2-distribution.

To make the continuity correction, one cell frequency, say 'a' is replaced by $(a+\frac{1}{2})$ and $(a-\frac{1}{2})$ according as $ad \lesseqgtr bc$ and the other cell frequencies are then adjusted so as to leave the marginal totals of the contingency table unchanged. When $ad < bc$, the modification is done as below:

B \ A	A_1	A_2	Total
B_1	$a+\frac{1}{2}$	$c-\frac{1}{2}$	$a+c$
B_2	$b-\frac{1}{2}$	$d+\frac{1}{2}$	$b+d$
Total	$a+b$	$c+d$	N

The corrected χ^2 is given by

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

with d.f. 1.

If $ad > bc$, we have the corrected χ^2 given by

$$\chi^2 = \frac{N\left\{ad-bc-\frac{N}{2}\right\}^2}{(a+b)(a+c)(b+d)(c+d)} \text{ with d.f. 1.}$$

$$\therefore \chi^2_{corrected} = \frac{N\left\{|ad-bc|-\frac{N}{2}\right\}^2}{(a+b)(a+c)(b+d)(c+d)} \text{ with d.f. 1.}$$

The continuity correction invariably improves the test of significance for the independence of the attributes in a 2X2 contingency table.

x —————— x