X^2 distribution
1. Definition 4
2. Derivation 6
3. Moment Generating Function 9
4. Cumulant Generating Function 10
5. Additive Property 12
6. Application 13
7. Theorem 1 14
8. Theorem 2 18

t distribution
1. Definition 22
2. Constants of t distribution 23
3. Applications 28

F distribution
1. Definition 28
2. Derivation 30
3. Constants of F distribution 33
4. Mode of F distribution 37
5. Application 39

Relation between t and F distribution 40

Relation Between F and X^2 distribution 41

Estimator 44
Sufficiency 46
Likelihood Function 57
Maximum Likelihood Estimator 58

Hypothesis Testing 73

Contingency Table 81

Non-parametric Test 93
Meadian Test 97

Randomly Solved Questions (Hypothesis & Contingency Table) 100

There are 4 sampling distributions:

a) χ^2 (chi-square) distribution

b) T distribution

c) F distribution

d) Z distribution

$X \sim N(\mu, \sigma^2)$

Here, X is a random variable which follows normal distribution with mean μ and variance σ^2.

$Z \sim N(0,1)$ denotes standard normal distribution.

$Z = \frac{X - \mu}{\sigma}$ denotes transformation from normal distribution to standard normal distribution.

Thus if $X \sim N(\mu; \sigma^2)$

Then $Z = \frac{X-\mu}{\sigma} \sim N(0;1)$

and $Z^2 = \left(\frac{X-\mu}{\sigma}\right)^2$, is a chi-square variate with 1 degree of freedom.

Example:

$$f(x) = \frac{x}{2}; \quad 0 \le x \le 2$$
$$= 0; \quad \text{otherwise}$$

Then, find the distribution of $2x+3$.

Solution:

Let, $z = 2x+3$

or, $2x = z-3$

or, $x = \frac{z-3}{2}$ [Jacobian Transformation]

or, $dx = \frac{1}{2} dz$

or, $\frac{dx}{dz} = \frac{1}{2}$

or, $|J| = \frac{1}{2}$

Now,

$$f(z) = \frac{\frac{z-3}{2}}{2} \cdot |J|$$

$$= \frac{z-3}{4} \cdot \frac{1}{2}$$

$$= \frac{z-3}{8}$$

when, $x = 0$
then, $z = 2 \cdot 0 + 3 = 3$

when, $x = 2$
then, $z = 2 \cdot 2 + 3$
$= 7$

$$\therefore f(z) = \frac{z-3}{8}; \quad 3 \leq z \leq 7$$

$$= 0; \quad \text{otherwise.}$$

<u>Question:</u> Define and derive χ^2 distribution.

<u>Definition:</u> In probability theory and statistics, the χ^2 distribution with k degree of freedom (d.f.) is the distribution of a sum of the squares of k independent standard normal variates. It is a continuous probability distribution having the probability density function (pdf) is as

$$f(x) = \frac{1}{2^{n/2}\Gamma\frac{n}{2}} e^{-x/2} x^{n/2-1} \quad ; \quad x \geq 0$$

$$= 0 \; ; \quad \text{otherwise}$$

Thus if,

$$X \sim N(\mu, \sigma^2)$$

Then, $z = \frac{X-\mu}{\sigma} \sim N(0,1)$

and $z^2 = \left(\frac{X-\mu}{\sigma}\right)^2$ is a χ^2 distribution.

In general if, $x_i \; (i=1,2,3,\ldots,K)$ are K-independent normal variates with mean μ_i and variance $\sigma_i^2 \; (i=1,2,3,\ldots,K)$

then, $\chi^2 = \sum_{i=1}^{K} \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2$ is a χ^2 variate with K d.f.

Derivation:

Method of Moment Generating Function

If X_i $(i = 1, 2, 3, \ldots, N)$ are n independent normal variates with mean μ_i and variance σ_i^2, the we want the distribution of

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{X_i - \mu_i}{\sigma_i}\right)^2 = \sum_{i=1}^{n} U_i^2 \text{ ; where } U_i = \frac{X_i - \mu_i}{\sigma_i}$$

Since X_i's are independent. U_i's are also independent.

By definition of moment generating function, we have

Pi (capital Letter) (Product)

$$M_{\chi^2}(t) = M_{\Sigma U_i^2}(t) = \prod_{i=1}^{n} M_{U_i^2}(t) = \left[M_{U_i^2}(t)\right]^n$$

Since U_i's $\sim N(0,1)$ are identicall distributed.

Now

pdf of normal distribution

$$M_{U_i^2}(t) = E\left(e^{U_i^2 t}\right) = \int_{-\infty}^{\infty} \exp(tU_i^2)\, f(x_i)\, dx_i$$

$$= \int_{-\infty}^{\infty} e^{tU_i^2} \frac{1}{\sigma\sqrt{2\pi}} \exp\{-(x_i-\mu)^2/2\sigma^2\}\, dx_i$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} \exp(tu_i^2) \exp(-u_i^2/2)\, du_i$$

where, $u_i = \dfrac{x_i - \mu}{\sigma}$

or, $du_i = \dfrac{1}{\sigma} dx_i$

or, $dx_i = \sigma \cdot du_i$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-u_i^2/2 + tu_i^2}\, du_i$$

$$= \frac{1}{\sqrt{2\pi}} \int_{-\infty}^{\infty} e^{-\frac{1}{2}(1-2t)u_i^2}\, du_i$$

$$= \frac{1}{\sqrt{2\pi}} \cdot \frac{\sqrt{\pi}}{\left(\frac{1-2t}{2}\right)^{1/2}} = \frac{1}{(1-2t)^{1/2}}$$

$$\left[\int_{-\infty}^{\infty} e^{-a^2x^2}\, dx = \frac{\sqrt{\pi}}{|a|}\right]$$

$$\therefore M_{U_i^2}(t) = (1-2t)^{-1/2}$$

$\therefore M_{\sum_{i=1}^{n} u_i^2}(t) = (1-2t)^{-n/2}$

or, $M_{\chi^2}(t) = (1-2t)^{-n/2}$

which is the mgf of a Gamma variate with parameters $\frac{1}{2}$ and $\frac{1}{2}n$ d.f.

Hence by uniqueness theorem of mgf's,

$$\chi^2 = \sum_{i=1}^{n} \left(\frac{x_i - \mu_i}{\sigma_i}\right)^2$$

is a Gamma variate with parameters $\frac{1}{2}$ and $\frac{1}{2}n$ d.f.

Probability differential function is

$$df(x) = \frac{e^{-x/2}\, x^{n/2-1}}{\Gamma\frac{n}{2}} \left(\frac{1}{2}\right)^{n/2} dx \qquad f(x) = \frac{a^\lambda}{\Gamma\lambda} e^{-ax} x^{\lambda-1}$$

$$\therefore dP(\chi^2) = \frac{\left(\frac{1}{2}\right)^{n/2}}{\Gamma\frac{n}{2}} \left[\exp\left(-\frac{1}{2}\chi^2\right)\right] (\chi^2)^{n/2-1}\, d\chi^2 \qquad [0 \le \chi^2 < \infty]$$

which is the required pdf of χ^2 distribution with n d.f.

M.G.F. of χ^2 distribution:

We know that pdf of χ^2 variate is as

$$f(x) = \frac{1}{2^{n/2}\Gamma\frac{n}{2}} e^{-x/2} x^{n/2-1} \quad ; \quad 0 \le x \le \infty$$

By definition of mgf, we have

$$M_x(t) = E(e^{tx}) = \int_0^\infty e^{tx} f(x)\, dx$$

$$= \int_0^\infty e^{tx} \frac{1}{2^{n/2}\Gamma\frac{n}{2}} e^{-x/2} x^{n/2-1}\, dx$$

$$= \frac{1}{2^{n/2}\Gamma\frac{n}{2}} \int_0^\infty e^{tx} \cdot e^{-x/2} x^{n/2-1}\, dx$$

$$= \frac{1}{2^{n/2}\Gamma\frac{n}{2}} \int_0^\infty \exp\left[-\left(\frac{1-2t}{2}\right)x\right] \cdot x^{n/2-1}\, dx$$

$$= \frac{1}{2^{n/2}\Gamma\frac{n}{2}} \frac{\Gamma\frac{n}{2}}{\left(\frac{1-2t}{2}\right)^{n/2}}$$

[Using Gamma Integral]

$$\int_0^\infty e^{-ax} x^{n-1}\, dx = \frac{\Gamma n}{a^n}$$

$$= \frac{1}{2^{n/2}(1-2t)^{n/2}} \cdot 2^{n/2}$$

$$= \frac{1}{(1-2t)^{n/2}}$$

$$\therefore M_x(t) = (1-2t)^{-n/2}$$

which is the mgf of χ^2 distribution.

Cumulant Generating Function of χ^2- distribution:

$$K_{\chi^2}(t) = \log M_x(t) = \log (1-2t)^{-n/2}$$

$$= -\frac{n}{2} \log (1-2t)$$

$$= \frac{n}{2}\left[2t + \frac{(2t)^2}{2} + \frac{(2t)^3}{3} + \frac{(2t)^4}{4} + \dots\right]$$

Now, K_r th cumulant is the coefficient of $\frac{t^r}{r!}$ in $K(t) = n\,2^{r-1}(r-1)!$

$\therefore K_1 = n$

$K_2 = 2n$

$K_3 = 8n$

$K_4 = 48n$

Hence,

μ_r' = moment about origin

μ_r = moment about mean or central moment

$$\text{Mean} = \mu_1 = K_1 = n$$

$$\text{variance} = \mu_2 = K_2 = 2n$$

$$\mu_3 = K_3 = 8n$$

$$\mu_4 = K_4 + 3K_2^2$$

$$= 48n + 3(2n)^2 = 48n + 12n^2$$

$$\beta_1 = \frac{\mu_3^2}{\mu_2^3} = \frac{(8n)^2}{(2n)^3} = \frac{64n^2}{8n^3} = \frac{8}{n}$$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{12}{n} + 3$$

B1 = measure of skewness
B2 = measure of kurtosis

B1 = 0 symmetric (normal)
B1 > 0 positively skewed
B1 < 0 negatively skewed

Remark:

If $\beta_2 = 3$, then the curve is mesokurtic. (normal)

If $\beta_2 > 3$, " " " " Leptokurtic.

If $\beta_2 < 3$, " " " " Platykurtic.

Here, $\beta_2 = \frac{12}{n} + 3 > 3$ as $n > 0$.

Additive Property of χ^2-variates:

Theory: The sum of independent χ^2 variates is also a χ^2 variate.

More precisely if $x_1, x_2, \ldots, x_K$ are the K independent χ^2 variates, the $x_1 + x_2 + \ldots + x_K$ is also a χ^2 variate.

Proof:

Let X_i $(i = 1, 2, \ldots, K)$ be K independent χ^2 ~~inde~~ variates with n_i $(i = 1, 2, \ldots, K)$ degree of freedom respectively.

Then by definition of mgf

$$M_{X_i}(t) = (1-2t)^{-n_i/2} \quad ; \quad i = 1, 2, \ldots, K$$

The mgf of the sum $\sum_{i=1}^{K} X_i$ is given by

$$M_{\Sigma X_i}(t) = M_{X_1}(t) \cdot M_{X_2}(t) \cdot \ldots \cdot M_{X_K}(t)$$

$$= (1-2t)^{-n_1/2} \cdot (1-2t)^{-n_2/2} \cdot \ldots \cdot (1-2t)^{-n_K/2}$$

$$= (1-2t)^{-\frac{1}{2}(n_1 + n_2 + \cdots + n_K)}$$

$= (1-2t)^{-n/2}$

where $n = n_1 + n_2 + \dots + n_k$

which is the mgf of χ^2 variate with $n = n_1 + n_2 + \dots + n_k$ degree of freedom.

Application of χ^2 distribution:

χ^2 distribution has a wide range of application in statistics as well as engineering. Some of these applications are given below

1. To test if the hypothetical value of the population variance is $\sigma^2 = \sigma_0^2$ (say).

2. To test the 'goodness of fit' of a particular distribution.

3. To test the independence of attributes by using contingency table.

4. To test the homogeneity of independent

estimates of the population variance.

5. To combine various probabilities obtained from independent experiments to give a single test of significance.

6. To test the homogeneity of independent estimates of the population correlation coefficient.

Theorem:

If χ_1^2 and χ_2^2 are two independent χ^2 variate with n_1 and n_2 df respectively, then $\frac{\chi_1^2}{\chi_2^2}$ is a $\beta_2\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$ variate.

Proof:

Since χ_1^2 and χ_2^2 are independent χ^2 variates with n_1 and n_2 df respectively, then the joint distribution function can be written as

$$f(\chi_1^2, \chi_2^2) = f(\chi_1^2)\, f(\chi_2^2)$$

$$= \frac{1}{2^{n_1/2}\Gamma\frac{n_1}{2}} e^{-\tilde{x}_1/2} (\tilde{x}_1)^{\frac{n_1}{2}-1} \cdot$$

$$\frac{1}{2^{n_2/2}\Gamma\frac{n_2}{2}} e^{-\tilde{x}_2/2} (\tilde{x}_2)^{\frac{n_2}{2}-1}$$

$$= \frac{1}{2^{(n_1+n_2)/2}\,\Gamma\frac{n_1}{2}\,\Gamma\frac{n_2}{2}} e^{-(\tilde{x}_1+\tilde{x}_2)/2}$$

$$(\tilde{x}_1)^{\frac{n_1}{2}-1} \cdot (\tilde{x}_2)^{\frac{n_2}{2}-1}$$

Let us make transformation as,

$u = \tilde{x}_1 / \tilde{x}_2$ and $v = \tilde{x}_2$

so that $\tilde{x}_1 = uv$ and $\tilde{x}_2 = v$

The Jacobian J is obtained as

$$J = \frac{\partial(\tilde{x}_1, \tilde{x}_2)}{\partial(u,v)} = \begin{vmatrix} v & u \\ 0 & 1 \end{vmatrix} = v$$

Thus the joint density function of u and v becomes

$$f(u,v) = \frac{1}{2^{(n_1+n_2)/2}\,\Gamma\frac{n_1}{2}\,\Gamma\frac{n_2}{2}}\, e^{-\frac{1}{2}(1+u)v}$$

$$(uv)^{\frac{n_1}{2}-1} \cdot v^{\frac{n_2}{2}-1} \cdot v \qquad [\because J = v]$$

$$= \frac{1}{2^{\frac{n_1+n_2}{2}}\,\Gamma\frac{n_1}{2}\,\Gamma\frac{n_2}{2}}\, e^{-\frac{1}{2}(1+u)v}\, u^{\frac{n_1}{2}-1} \cdot v^{\frac{n_1}{2}-1+\frac{n_2}{2}-1+1}$$

$$= \frac{1}{2^{\frac{n_1+n_2}{2}}\,\Gamma\frac{n_1}{2}\,\Gamma\frac{n_2}{2}}\, u^{\frac{n_1}{2}-1}\, e^{-\frac{1}{2}(1+u)v}\, v^{\frac{n_1}{2}+\frac{n_2}{2}-1}$$

To get the marginal pdf of u, we have to integrate the joint pdf of u and v with respect to v over the range 0 to ∞.

$$\therefore f(u) = \int_v f(u,v)\,dv$$

$$= \int_0^\infty \frac{1}{2^{\frac{n_1+n_2}{2}} \Gamma\frac{n_1}{2} \Gamma\frac{n_2}{2}} u^{\frac{n_1}{2}-1} e^{-\frac{1}{2}(1+u)v} v^{\frac{n_1}{2}+\frac{n_2}{2}-1}\,dv$$

$$= \frac{1}{2^{(n_1+n_2)/2} \Gamma\frac{n_1}{2} \Gamma\frac{n_2}{2}} u^{\frac{n_1}{2}-1} \int_0^\infty e^{-\frac{1}{2}(1+u)v} v^{\frac{n_1}{2}+\frac{n_2}{2}-1}\,dv$$

$$= \frac{u^{\frac{n_1}{2}-1}}{2^{(n_1+n_2)/2} \Gamma\frac{n_1}{2} \Gamma\frac{n_2}{2}} \cdot \frac{\Gamma\frac{n_1+n_2}{2}}{\left(\frac{1+u}{2}\right)^{\frac{n_1+n_2}{2}}}$$

Note for Gamma distribution
(not a part of this proof)

$$\int_0^\infty f(a,\lambda) = \int_0^\infty \frac{a^\lambda e^{-ax} x^{\lambda-1}}{\Gamma\lambda}\,dx = 1$$

Here, $a = \frac{1+u}{2}$

$\lambda = \frac{n_1}{2} + \frac{n_2}{2}$

$$= \frac{u^{\frac{n_1}{2}-1} \Gamma\left(\frac{n_1+n_2}{2}\right) 2^{\frac{n_1+n_2}{2}}}{2^{(n_1+n_2)/2} \Gamma\left(\frac{n_1}{2}\right) \Gamma\left(\frac{n_2}{2}\right) (1+u)^{(n_1+n_2)/2}}$$

$$= \frac{\Gamma\left(\frac{n_1+n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right) \Gamma\left(\frac{n_2}{2}\right)} \cdot \frac{u^{n_1/2-1}}{(1+u)^{(n_1+n_2)/2}} \quad ; 0 \le u < \infty$$

$$= \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \cdot \frac{u^{(n_1/2)-1}}{(1+u)^{(n_1+n_2)/2}} \quad ; 0 \le u < \infty \qquad \beta(m,n) = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}$$

Hence, $u = \frac{\chi_1^2}{\chi_2^2}$ is a $\beta_2\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$ variate.

<u>Theorem:</u>

If χ_1^2 and χ_2^2 are independent χ^2-variates with n_1 and n_2 df respectively, then

$$U = \frac{\chi_1^2}{\chi_1^2 + \chi_2^2} \quad \text{and} \quad V = \chi_1^2 + \chi_2^2$$

are independently distributed, U as a $\beta_1\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$ variate and V as a χ^2 variate with (n_1+n_2) degree of freedom

Proof:

Since χ_1^2 and χ_2^2 are independent χ^2 variates with n_1 and n_2 df respectively, then the joint distribution function can be written as

$$f(\chi_1^2, \chi_2^2) = f(\chi_1^2) f(\chi_2^2)$$

$$= \frac{1}{2^{(n_1+n_2)/2} \Gamma\left(\frac{n_1}{2}\right) \Gamma\left(\frac{n_2}{2}\right)} e^{-(\chi_1^2+\chi_2^2)/2} (\chi_1^2)^{\frac{n_1}{2}-1} (\chi_2^2)^{\frac{n_2}{2}-1}$$

Let us transform to u and v defined as follows:

$$u = \frac{\chi_1^2}{\chi_1^2 + \chi_2^2} \quad \text{and} \quad v = \chi_1^2 + \chi_2^2$$

so that $\chi_1^2 = uv$ and $\chi_2^2 = v - \chi_1^2 = (1-u)v$

The Jacobian J is obtained as

$$J = \frac{\partial(\chi_1^2, \chi_2^2)}{\partial(u, v)} = \begin{vmatrix} v & u \\ -v & 1-u \end{vmatrix} = v$$

Thus the joint density function of u and v becomes

$$f(u,v) = \frac{1}{2^{(n_1+n_2)/2}\,\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \cdot e^{-v/2}$$

$$(uv)^{n_1/2-1} \cdot \{(1-u)v\}^{n_2/2-1} \cdot |J|$$

$$= \frac{1}{2^{(n_1+n_2)/2}\,\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \cdot u^{(n_1/2)-1} \cdot (1-u)^{(n_2/2)-1}$$

$$e^{-v/2} \cdot v^{(n_1/2)-1} \cdot v^{(n_2/2)-1} \cdot v \qquad [\text{as } J = v]$$

$$= \frac{1}{2^{(n_1+n_2)/2}\,\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \cdot u^{(n_1/2)-1} \cdot (1-u)^{(n_2/2)-1}$$

$$e^{-v/2} \cdot v^{\{(n_1+n_2)/2\}-1}$$

$$= \frac{\Gamma\frac{n_1+n_2}{2}}{\Gamma\frac{n_1}{2}\Gamma\frac{n_2}{2}} u^{(n_1/2)-1} (1-u)^{(n_2/2)-1} \cdot$$

$$\frac{1}{2^{(n_1+n_2)/2}\Gamma\frac{n_1}{2}+\frac{n_2}{2}} e^{-v/2} v^{\{(n_1+n_2)/2\}-1}$$

since the joint probability function of u and v is the product of their respective probability function, u and v are independently distributed, with

$$f(u) = \frac{1}{B\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} u^{(n_1/2)-1} (1-u)^{(n_2/2)-1}$$

$$f(v) = \frac{1}{2^{(n_1+n_2)/2}\Gamma\frac{n_1}{2}\Gamma\frac{n_2}{2}} e^{-v/2} v^{\{(n_1+n_2)/2\}-1}$$

Hence, u as a $\beta_1\left(\frac{n_1}{2}, \frac{n_2}{2}\right)$ variate and v as a χ^2 variate with (n_1+n_2) df.

t - distribution

Definition:

Let $x_i, (i = 1, 2, \ldots, n)$ be a random sample of size n from a normal population with mean μ and variance σ^2. Then student's t is defined by the statistic

$$t = \frac{\bar{x} - \mu}{s/\sqrt{n}}$$

where, $\bar{x} = \frac{1}{n} \sum_{i=1}^{n} x_i$ is the sample mean

$s^2 = \frac{1}{n-1} \sum_{i=1}^{n} (x_i - \bar{x})$ is an unbiased estimate of population variance σ^2,

and it follows student's t-distribution with $\lambda = (n-1)$ df and with pdf

$$f(t) = \frac{1}{\sqrt{\lambda}\, B\left(\frac{1}{2}, \frac{\lambda}{2}\right)} \cdot \frac{1}{\left\{1 + \frac{t^2}{\lambda}\right\}^{(\lambda+1)/2}};$$

$$-\infty < t < \infty$$

Remark:

If we take $\lambda = 1$, we get

$$f(t) = \frac{1}{B\left(\frac{1}{2}, \frac{1}{2}\right)} \cdot \frac{1}{(1+t^2)}$$

$$= \frac{1}{\pi} \cdot \frac{1}{(1+t^2)} \; ; \quad -\infty < t < \infty$$

which is the pdf of standard Cauchy distribution.

Constants of t-distribution:

Since t-distribution is symmetrical about the line $t=0$, all the moments of odd order about origin vanish i.e. zero.

i.e. μ'_{2r+1} (about origin) $= 0$; $r = 0, 1, 2, \ldots$

In particular μ'_1 (about origin) $= 0 =$ Mean

Hence central moments coincide with moments about origin.

$$\mu_{2r+1} = 0 \; ; \quad r = 0, 1, 2, \ldots$$

The moments of even order are given by

$$\mu_{2r} = \mu_{2r}' \text{ (about origin)}$$

$$= \int_{-\infty}^{\infty} t^{2r} f(t)\, dt \qquad \left[\mu_r' = E(x^r) = \int x^r f(x)\, dx\right]$$

$$= 2\int_{0}^{\infty} t^{2r} f(t)\, dt \qquad [\text{since it is symmetric}]$$

$$= 2 \cdot \frac{1}{B\left(\frac{1}{2}, \frac{n}{2}\right)\sqrt{n}} \int_0^{\infty} \frac{t^{2r}}{\left\{1+\frac{t^2}{n}\right\}^{(n+1)/2}}\, dt$$

Put, $1+\frac{t^2}{n} = \frac{1}{y}$

or, $t^2 = \frac{n(1-y)}{y}$

or, $2t\,dt = -\frac{n}{y^2}\,dy$

or, $dt = -\frac{n}{2ty^2}\,dy$

When $t=0$, $y=1$ and when $t=\infty$, $y=0$

Therefore,

$$\mu_{2r} = \frac{2}{\sqrt{n}\,\beta\left(\frac{1}{2},\frac{n}{2}\right)} \int_1^0 \frac{t^{2r}}{\left(\frac{1}{y}\right)^{(n+1)/2}} \cdot -\frac{n}{2ty^2}\,dy$$

$$= \frac{\sqrt{n}}{\beta\left(\frac{1}{2},\frac{n}{2}\right)} \int_0^1 (t^2)^{(2r-1)/2} \cdot y^{\{(n+1)/2\}-2} \cdot dy$$

$$= \frac{\sqrt{n}}{\beta\left(\frac{1}{2},\frac{n}{2}\right)} \int_0^1 \left[n\left(\frac{1-y}{y}\right)\right]^{r-\frac{1}{2}} \cdot y^{\frac{n+1}{2}-2}\,dy$$

$$= \frac{n^r}{\beta\left(\frac{1}{2},\frac{n}{2}\right)} \int_0^1 y^{\frac{n}{2}-r-1} (1-y)^{r-\frac{1}{2}}\,dy$$

$$= \frac{n^r}{\beta\left(\frac{1}{2},\frac{n}{2}\right)} \beta\left(\frac{n}{2}-r,\ r+\frac{1}{2}\right) \quad \left[\beta(m,n) = \int_0^1 x^{m-1}(1-x)^{n-1}\,dx\right.$$

$$= n^r \cdot \frac{\cancel{\Gamma\left(\frac{1}{2}+\frac{n}{2}\right)}}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}\right)} \cdot \frac{\Gamma\left(\frac{n}{2}-r\right)\Gamma\left(r+\frac{1}{2}\right)}{\cancel{\Gamma\left(\frac{n}{2}-r+r+\frac{1}{2}\right)}} \quad \left[\beta(m,n) = \frac{\Gamma(m)\Gamma(n)}{\Gamma(m+n)}\right.$$

$$= n^r \cdot \frac{\Gamma\left(\frac{n}{2}-r\right)\Gamma\left(r+\frac{1}{2}\right)}{\Gamma\left(\frac{1}{2}\right)\Gamma\left(\frac{n}{2}\right)}$$

$$= n^r \cdot \frac{\left(r-\frac{1}{2}\right)\left(r-\frac{3}{2}\right)\cdots\frac{3}{2}\cdot\frac{1}{2}\cdot\sqrt{\frac{1}{2}}\,\Gamma\left(\frac{n}{2}-r\right)}{\sqrt{\frac{1}{2}}\left(\frac{n}{2}-1\right)\left(\frac{n}{2}-2\right)\cdots\left(\frac{n}{2}-r\right)\Gamma\left(\frac{n}{2}-r\right)}$$

$$= n^r \cdot \frac{(2r-1)(2r-3)\cdots 3\cdot 1}{(n-2)(n-4)\cdots(n-2r)} \quad ; \frac{n}{2} > r$$

In particular $r = 1, 2, \ldots$, we have

$$\mu_2 = n\,\frac{1}{(n-2)} = \frac{n}{n-2} \quad ; \; n > 2$$

and, $$\mu_4 = n^2\,\frac{3\cdot 1}{(n-2)(n-4)} = \frac{3n^2}{(n-2)(n-4)} \quad ; \; n > 4$$

Hence, $\beta_1 = \dfrac{\mu_3^2}{\mu_2^3} = 0$ [since all odd moments are zero i.e. $\mu_3 = 0$

$$\beta_2 = \frac{\mu_4}{\mu_2^2} = \frac{3n^2}{(n-2)(n-4)} \cdot \frac{(n-2)^2}{n^2} = 3 \cdot \frac{n-2}{n-4}$$

Remark:

As $n \to \infty$, $\beta_1 = 0$ and

$$\beta_2 = \lim_{n \to \infty} 3 \cdot \frac{n-2}{n-4} = 3 \lim_{n \to \infty} \left[\frac{1 - (2/n)}{1 - (4/n)} \right] = 3$$

we know that

β_1 is the measure of skewnees

As $\beta_1 = 0$, the distribution is symmetric.

and β_2 is the measure of kurtosis

as $\beta_2 = 3$, the distribution is mesokurtic.

Remember:

$\beta_1 = 0 \rightarrow$ symmetric

$\beta_1 < 0 \rightarrow$ negatively skewed

$\beta_1 > 0 \rightarrow$ positively skewed

$\beta_2 = 3 \rightarrow$ Mesokurtic

$\beta_2 < 3 \rightarrow$ Platykurtic

$\beta_3 > 3 \rightarrow$ Leptokurtic

Applications of t-distribution:

There are several applications of t-distribution in statistical hypothesis testing. Some of them are as follows:

1. To test of significance of single mean.
2. To test of significance of difference of means
3. To test of significance of difference of means from correlated population.

F-distribution

If X and Y are two independent chi-square variates with n_1 and n_2 df respectively, the F-statistic is defined by

$$F = \frac{X/n_1}{Y/n_2}$$

In other words, F is defined as the ratio of two independent chi-square variates divided by their respective degrees of freedom and it follows Snedecor's F-distribution with (n_1, n_2) df with pdf given by

$$f(F) = \frac{\left(\frac{n_1}{n_2}\right)^{n_1/2}}{\beta\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \cdot \frac{F^{\frac{n_1}{2}-1}}{\left\{1 + \frac{n_1}{n_2} F\right\}^{(n_1+n_2)/2}} ; \quad 0 \le F < \infty$$

It does not involve any population parameters but depends only on the degrees of freedom n_1 and n_2. It is denoted as $F \sim F(n_1, n_2)$.

Derivation of F-distribution:

Let x and y be two independent chi-square variates with n_1 and n_2 df then their joint probability differential is given by

$$dF(x,y) = dF(x) \cdot dF(y)$$

$$= \left\{ \frac{1}{2^{n_1/2} \Gamma\frac{n_1}{2}} e^{-x/2} x^{\frac{n_1}{2}-1} dx \right\} \times$$

$$\left\{ \frac{1}{2^{n_2/2} \Gamma\frac{n_2}{2}} e^{-y/2} y^{\frac{n_2}{2}-1} dy \right\}$$

$$= \frac{1}{2^{(n_1+n_2)/2} \Gamma\frac{n_1}{2} \Gamma\frac{n_2}{2}} \cdot e^{-(x+y)/2}$$

$$x^{\frac{n_1}{2}-1} \cdot y^{\frac{n_2}{2}-1} dx\, dy$$

Let us make the following transformation of variables:

$$F = \frac{x/n_1}{y/n_2} \quad \text{and} \quad u = y$$

$$\text{or,} \quad x = \frac{n_1}{n_2} F u \quad \text{and} \quad y = u$$

Jacobian of transformation J is given by

$$J = \frac{\partial(x,y)}{\partial(F,u)} = \begin{vmatrix} \frac{n_1}{n_2} u & 0 \\ \frac{n_1}{n_2} F & 1 \end{vmatrix} = \frac{n_1}{n_2} u$$

Thus the distribution of the transformed variable is

$$dG(F,u) = \frac{1}{2^{(n_1+n_2)/2} \Gamma\frac{n_1}{2} \Gamma\frac{n_2}{2}} \cdot \exp\left\{-\frac{u}{2}\left(1+\frac{n_1}{n_2}F\right)\right\}$$

$$\left(\frac{n_1}{n_2} F u\right)^{\frac{n_1}{2}-1} \cdot u^{\frac{n_2}{2}-1} \cdot \underbrace{\boxed{\frac{n_1}{n_2} u}}_{J} \, du \, dF$$

$$= \frac{\left(\frac{n_1}{n_2}\right)^{\frac{n_1}{2}}}{2^{(n_1+n_2)/2}\,\Gamma\frac{n_1}{2}\,\Gamma\frac{n_2}{2}} \exp\left\{-\frac{u}{2}\left(1+\frac{n_1}{n_2}F\right)\right\} u^{\frac{n_1+n_2}{2}-1} \cdot F^{\frac{n_1}{2}-1}\, du\, dF$$

To get the distribution of F we need to integrate u over 0 to ∞

$$g(F)\,dF = \frac{\left(\frac{n_1}{n_2}\right)^{n_1/2} F^{(n_1/2)-1}\, dF}{2^{(n_1+n_2)/2}\,\Gamma\frac{n_1}{2}\,\Gamma\frac{n_2}{2}} \cdot \left[\int_0^\infty \exp\left\{-\frac{u}{2}\left(1+\frac{n_1}{n_2}F\right)\right\} u^{\frac{n_1+n_2}{2}-1}\, du\right]$$

$$= \frac{\left(\frac{n_1}{n_2}\right)^{n_1/2} F^{(n_1/2)-1}\, dF}{2^{(n_1+n_2)/2}\,\Gamma\frac{n_1}{2}\,\Gamma\frac{n_2}{2}} \cdot \frac{\Gamma\frac{n_1+n_2}{2}}{\left[\frac{1}{2}\left(1+\frac{n_1}{n_2}F\right)\right]^{\frac{n_1+n_2}{2}}}$$

$$\therefore\ g(F) = \frac{\left(\frac{n_1}{n_2}\right)^{n_1/2}}{\beta\left(\frac{n_1}{2},\frac{n_2}{2}\right)} \cdot \frac{F^{(n_1/2)-1}}{\left\{1+\frac{n_1}{n_2}F\right\}^{\frac{n_1+n_2}{2}}};\quad 0 \le F < \infty$$

which is the required probability function of F- distribution with (n_1, n_2) df

Constants of F-distribution:

$$\mu_r' \text{ (about origin)} = E(F^r)$$

$$= \int_0^\infty F^r f(F)\, dF$$

$$= \int_0^\infty \frac{\left(\frac{n_1}{n_2}\right)^{n_1/2}}{\beta\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \cdot \frac{F^{(n_1/2)-1}}{\left\{1+\frac{n_1}{n_2}F\right\}^{(n_1+n_2)/2}} \cdot F^r\, dF$$

To evaluate the integral, put

$$\frac{n_1}{n_2} F = y$$

or, $F = \frac{n_2}{n_1} y$

or, $dF = \frac{n_2}{n_1} dy$

when, $F = 0$, then $y = 0$

$F = \infty$, then $y = \infty$

Now,

$$\mu_r' = \frac{\left(\frac{n_1}{n_2}\right)^{n_1/2}}{\beta\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \int_0^\infty \frac{\left(\frac{n_1}{n_2} y\right)^{r+(n_1/2)-1}}{(1+y)^{(n_1+n_2)/2}} \cdot \left(\frac{n_2}{n_1}\right) dy$$

$$\text{or, } \mu_r' = \frac{\left(\frac{n_1}{n_2}\right)^{n_1/2} \cdot \left(\frac{n_2}{n_1}\right)^{r+(n_1/2)-1} \cdot \left(\frac{n_2}{n_1}\right)}{\beta\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \cdot$$

$$\int_0^\infty \frac{y^{r+(n_1/2)-1}}{\{1+y\}^{r+(n_1/2)+\{(n_2/2)-r\}}} \, dy$$

$$\text{or, } \mu_r' = \frac{\left(\frac{n_2}{n_1}\right)^r}{\beta\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \cdot \beta\left(r+\frac{n_1}{2}, \frac{n_2}{2}-r\right)$$

$$\boxed{\because \beta(m,n) = \int_0^\infty \frac{x^{m-1}}{(1+x)^{m+n}} \, dx}$$

$$\text{or, } \mu_r' = \left(\frac{n_2}{n_1}\right)^r \frac{\Gamma\left(\frac{n_1}{2}+\frac{n_2}{2}\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \cdot \frac{\Gamma\left(r+\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}-r\right)}{\Gamma\left(r+\frac{n_1}{2}+\frac{n_2}{2}-r\right)}$$

$$\text{or, } \mu_r' = \left(\frac{n_2}{n_1}\right)^r \frac{\Gamma\left(r+\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}-r\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}$$

$$\boxed{\because \beta(m,n) = \frac{\Gamma m \, \Gamma n}{\Gamma(m+n)}}$$

*

In particular,

$$\mu_1' = \left(\frac{n_2}{n_1}\right)^1 \frac{\Gamma\left(1+\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}-1\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}$$

or, $$\mu_1' = \frac{n_2}{n_1} \cdot \frac{\frac{n_1}{2}\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}-1\right)}{\Gamma\left(\frac{n_1}{2}\right)\left(\frac{n_2}{2}-1\right)\Gamma\left(\frac{n_2}{2}-1\right)}$$

$$\boxed{\because \Gamma n = (n-1)\Gamma(n-1)}$$

or, $$\mu_1' = \frac{n_2}{n_1} \cdot \frac{n_1}{2} \cdot \frac{1}{\frac{n_2-2}{2}}$$

or, $$\mu_1' = \frac{n_2}{n_2-2}$$

Here μ_1' denotes mean of F-distribution, which is independent of n_1.

$$\mu_2' = \left(\frac{n_2}{n_1}\right)^2 \frac{\Gamma\left(\frac{n_1}{2}+2\right)\Gamma\left(\frac{n_2}{2}-2\right)}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)}$$

or, $$\mu_2' = \left(\frac{n_2}{n_1}\right)^2 \frac{\left(\frac{n_1}{2}+1\right)\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}-2\right)}{\Gamma\left(\frac{n_1}{2}\right)\left(\frac{n_2}{2}-1\right)\left(\frac{n_2}{2}-2\right)\Gamma\left(\frac{n_2}{2}-2\right)}$$

$$\text{or, } \mu_2' = \left(\frac{n_2}{n_1}\right)^2 \frac{\left(\frac{n_1}{2}+1\right)\left(\frac{n_1}{2}\right)}{\left(\frac{n_2}{2}-1\right)\left(\frac{n_2}{2}-2\right)}$$

$$\text{or, } \mu_2' = \frac{n_2^2(n_1+2)}{n_1(n_2-2)(n_2-4)}$$

Hence the Variance is

$$\therefore \mu_2 = \mu_2' - (\mu_1')^2 = \frac{n_2^2(n_1+2)}{n_1(n_2-2)(n_2-4)} - \frac{n_2^2}{(n_2-2)^2}$$

$$\text{or, } \mu_2 = \frac{n_2^2}{(n_2-2)}\left[\frac{n_1+2}{n_1(n_2-4)} - \frac{1}{(n_2-2)}\right]$$

$$\text{or, } \mu_2 = \frac{n_2^2}{(n_2-2)}\left[\frac{n_1n_2+2n_2-2n_1-4-n_1n_2+4n_1}{n_1(n_2-4)(n_2-2)}\right]$$

$$\text{or, } \mu_2 = \frac{2n_2^2(n_1+n_2-2)}{n_1(n_2-2)^2(n_2-4)}$$

Mode of F- distribution:

We have pdf of F- distribution as

$$f(F) = \frac{\left(\frac{n_1}{n_2}\right)^{n_1/2}}{\beta\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \cdot \frac{F^{(n_1/2)-1}}{\left\{1+\frac{n_1}{n_2}F\right\}^{(n_1+n_2)/2}} \quad \cdots (1)$$

$$\therefore \log f(F) = c + \left(\frac{n_1}{2}-1\right)\log F - \left(\frac{n_1+n_2}{2}\right)\log\left(1+\frac{n_1}{n_2}F\right)$$

where. $c = \log\left[\frac{\left(\frac{n_1}{n_2}\right)^{n_1/2}}{\beta\,\frac{n_1}{2}, \frac{n_2}{2}}\right]$

which is independent of F.

Now, $\frac{\partial}{\partial F}\left[\log f(F)\right] = \left(\frac{n_1}{2}-1\right)\cdot\frac{1}{F} - \left(\frac{n_1+n_2}{2}\right)\cdot\frac{1}{\left(1+\frac{n_1}{n_2}F\right)}\cdot\frac{n_1}{n_2}$

or, $\frac{1}{f(F)}\cdot f'(F) = \left(\frac{n_1-2}{2}\right)\cdot\frac{1}{F} - \left(\frac{n_1+n_2}{2}\right)\cdot\frac{1}{\frac{n_2+n_1F}{n_2}}\cdot\frac{n_1}{n_2}$

or, $f'(F) = f(F)\left\{\left(\frac{n_1-2}{2}\right)\cdot\frac{1}{F} - \frac{n_1^2+n_1n_2}{2(n_2+n_1F)}\right\}$

Put. $\frac{\partial}{\partial F} f(F) = f'(F) = 0$

$$\therefore \frac{n_1 - 2}{2F} - \frac{n_1^2 + n_1 n_2}{2(n_2 + n_1 F)} = 0$$

$$\text{or, } \frac{2(n_2 + n_1 F)}{2F} = \frac{n_1^2 + n_1 n_2}{n_1 - 2}$$

$$\text{or, } \frac{n_2}{F} + n_1 = \frac{n_1^2 + n_1 n_2}{n_1 - 2}$$

$$\text{or, } F = \frac{n_2 (n_1 - 2)}{n_1 (n_2 + 2)}$$

It can be easily verified that at this point $f''(F) < 0$.

Hence $\quad \text{Mode} = \frac{n_2 (n_1 - 2)}{n_1 (n_2 + 2)}$

Application of F-distribution:

1. F-test for testing the equality of two population variances.
2. F-test for testing the significance of an observed multiple correlation coefficient.
3. F-test for testing the significance of an observed sample correlation ratio
4. F-test for testing the linearity of regression.
5. F-test for testing the equality of several means test.
6. F-test is used in ANOVA (Analysis of variance).

Relation between t and F distributions:

In F-distribution with (n_1, n_2) df

We take $n_1 = 1$, $n_2 = n$ and $t^2 = F$

i.e. $dF = 2t\,dt$.

Thus the probability differential of F

transforms to

$$dG(t) = \frac{(1/n)^{1/2}}{\beta\left(\frac{1}{2}, \frac{n}{2}\right)} \cdot \frac{(t^2)^{\frac{1}{2}-1}}{\left\{1+\frac{t^2}{n}\right\}^{\frac{(n+1)}{2}}} \, 2t\,dt; \quad 0 \le t^2 < \infty$$

$$= \frac{1}{\sqrt{n}\,\beta\left(\frac{1}{2}, \frac{n}{2}\right)} \cdot \frac{1}{\left\{1+\frac{t^2}{n}\right\}^{\frac{(n+1)}{2}}} \, dt; \quad -\infty < t < \infty$$

which is the pdf of t distribution with n df. Therefore we have the following relation between t and F distribution.

Note: The factor 2 disappearing since the total propability in the range $(-\infty, \infty)$ is unity.

If a statistic t follows t-distribution with n df, the t^2 follows F-distribution with $(1, n)$ df

Symbolically,

if $t \sim t_n$

then $t^2 \sim F(1, n)$

Relation between F and χ^2 distribution:

In F-distribution if we let $n_2 \to \infty$, then $\chi^2 = n_1 F$ follows χ^2-distribution with n_1 df.

Proof:

$$f(F) = \frac{\left(\frac{n_1}{n_2}\right)^{n_1/2}}{\beta\left(\frac{n_1}{2}, \frac{n_2}{2}\right)} \cdot \frac{F^{(n_1/2)-1}}{\left\{1 + \frac{n_1}{n_2} F\right\}^{(n_1+n_2)/2}}$$

$$\text{or, } f(F) = \frac{(n_1/n_2)^{n_1/2} \, F^{(n_1/2)-1}}{\Gamma\left(\frac{n_1}{2}\right) \Gamma\left(\frac{n_2}{2}\right)} \cdot \frac{\Gamma\left(\frac{n_1}{2} + \frac{n_2}{2}\right)}{\left\{1 + \frac{n_1}{n_2} F\right\}^{(n_1+n_2)/2}}$$

separate all n_2's and determine limit

$$\lim_{n_2 \to \infty} \frac{\Gamma\left(\frac{n_1}{2}+\frac{n_2}{2}\right)}{n_2^{n_1/2}\,\Gamma\left(\frac{n_2}{2}\right)} \to \frac{(n_2/2)^{n_1/2}}{n_2^{n_1/2}} = \frac{1}{2^{n_1/2}}$$

$$\left[\because \frac{\Gamma(n+k)}{\Gamma(n)} \to n^k \text{ as } n \to \infty\right]$$

Also $\lim_{n_2 \to \infty} \left\{1+\frac{n_1}{n_2}F\right\}^{(n_1+n_2)/2}$

$$= \lim_{n_2 \to \infty} \left[\left(1+\frac{n_1}{n_2}F\right)^{n_2}\right]^{1/2} \times \lim_{n_2 \to \infty} \left(1+\frac{n_1}{n_2}F\right)^{n_1/2}$$

$$= e^{n_1 F/2} \cdot 1$$

$$= e^{\chi^2/2} \quad [\because n_1 F = \chi^2]$$

$$\left[\lim_{n \to \infty} \left(1+\frac{a}{n}\right)^n = e^a, \quad \lim_{n \to \infty} \left(1+\frac{a}{n}\right)^m = 1\right]$$

Now,

$$df(F) = \frac{(n_1/n_2)^{n_1/2} F^{(n_1/2)-1}}{\Gamma\left(\frac{n_1}{2}\right)\Gamma\left(\frac{n_2}{2}\right)} \cdot \frac{\Gamma\left(\frac{n_1}{2}+\frac{n_2}{2}\right)}{\left\{1+\frac{n_1}{n_2}F\right\}^{(n_1+n_2)/2}} dF$$

Hence in the limit, the pdf of $x^2 = n_1 F$

i.e. $dF = \frac{1}{n_1} \cdot dx^2$ becomes

$$df(x^2) = \frac{(n_1/2)^{n_1/2} \cdot e^{-x^2/2}}{\sqrt{n_1/2}} \cdot \left(\frac{x^2}{n_1}\right)^{(n_1/2)-1} \cdot \frac{1}{n_1} dx^2$$

$$= \frac{1}{2^{n_1/2} \sqrt{\frac{n_1}{2}}} \cdot e^{-x^2/2} (x^2)^{(n_1/2)-1} dx^2;$$

$$0 < x^2 < \infty$$

which is the pdf of x^2 distribution with n_1 df.

Estimator

Estimators: We wish to determine the functions of the sample observations:

$T_1 = \hat{\theta}_1(x_1, x_2, \ldots, x_n)$, $T_2 = \hat{\theta}_2(x_1, x_2, \ldots, x_n) \ldots,$

$T_K = \hat{\theta}_K(x_1, x_2, \ldots, x_n)$

such that their distribution is concentrated as closely as possible near the true value of the parameter.

Then, the estimating functions are called estimators.

Characteristics of Estimators: The following are some of the criteria that should be satisfied by a good estimator.

i) Consistency: $T_n \to \theta$ as $n \to \infty$

ii) Unbiasedness: $E(\hat{\mu}) = \mu$

iii) Efficiency and [If $V(\hat{\mu}) < V(\hat{\mu}')$, then $\hat{\mu}$ is more efficient than $\hat{\mu}'$]

iv) sufficiency

Example 15.1: $x_1, x_2, \ldots, x_n$ is a random sample from a normal population $N(\mu, 1)$. Show that $t = \frac{1}{n}\sum_{i=1}^{n} x_i^2$, is an unbiased estimator of μ^2+1.

Solution:

Here given,

$$E(x_i) = \mu, \quad V(x_i) = 1 \; ; \quad \forall\, i = 1, 2, \ldots, n$$

We know

$$E(x_i^2) = V(x_i) + \{E(x_i)\}^2 = 1 + \mu^2$$

Now,

$$E(t) = E\left[\frac{1}{n}\sum_{i=1}^{n} x_i^2\right] = \frac{1}{n}\sum_{i=1}^{n} E(x_i^2)$$

$$= \frac{1}{n}\sum_{i=1}^{n}(1+\mu^2) \qquad \left[\because E(x_i^2) = 1+\mu^2\right]$$

$$= \frac{1}{n}(n + n\mu^2)$$

$$= 1+\mu^2$$

Hence t is an unbiased estimator of μ^2+1.

[Solved]

Sufficiency: An estimator is said to be sufficient for a parameter, if it contains all the information in the sample regarding the parameter. More precisely, if $T = t(x_1, x_2, \ldots, x_n)$ is an estimator of a parameter θ, based on a sample $x_1, x_2, \ldots, x_n$ of size n from the population with pdf $f(x, \theta)$ such that the conditional probability distribution of $x_1, x_2, \ldots, x_n$ given T, is independent of θ, then T is sufficient estimator for θ.

Factorization Theorem (Neyman): The necessary and sufficient condition for a distribution to admit sufficient statistic is provided by the 'factorization theorem' due to Neyman.

Statement: $T = t(x)$ is sufficient for θ if and only iff the joint density function L (say), of the sample values can be expressed in the form

$$L = g_\theta[t(x)] \cdot h(x)$$

where $g_\theta[t(x)]$ depends on x and θ only and $h(x)$ is independent of θ.

Example 15.13: Let $x_1, x_2, \ldots, x_n$ be a random sample from a uniform population on $[0, \theta]$. Find a sufficient estimator of θ.

Solution:

We know that, pdf of uniform population is

$$f(x_i, \theta) = \frac{1}{\theta} \; ; \qquad 0 \le x_i \le \theta$$

The joint distribution of $x_1, x_2, \ldots, x_n$ is

$$L = f(x_1, \theta) f(x_2, \theta) \ldots f(x_n, \theta)$$

or, $L = \frac{1}{\theta} \cdot \frac{1}{\theta} \cdots \frac{1}{\theta}$

or, $L = \frac{1}{\theta^n} \cdot 1 \cdots$ (1)

If $t = \max(x_1, x_2, \ldots, x_n) = x_{(n)}$

then pdf of T is given by

$$g(t, \theta) = n\left[F(x_{(n)})\right]^{n-1} \cdot f(x_{(n)}, \theta)$$

We have

$$f(x_{(n)}, \theta) = \frac{1}{\theta}$$

$$F(x) = P(X \leq x) = \int_0^x f(x, \theta)\, dx = \int_0^x \frac{1}{\theta} \cdot dx = \frac{x}{\theta}$$

Now,

$$g(t, \theta) = n\left[\frac{x_{(n)}}{\theta}\right]^{n-1} \cdot \left(\frac{1}{\theta}\right)$$

$$= \frac{n}{\theta^n}\left[x_{(n)}\right]^{n-1}$$

Rewriting (1), we get

$$L = \frac{n\left[x_{(n)}\right]^{n-1}}{\theta^n} \cdot \frac{1}{n\left[x_{(n)}\right]^{n-1}}$$

$$= g(t, \theta) \cdot h(x_1, x_2, \ldots, x_n)$$

Here $g(t, \theta)$ depends on x and θ only and $h(x)$ is independent of θ

Thus, $t = \max(x_1, x_2, \ldots, x_n) = x_{(n)}$ is the sufficient statistic of θ.

[Solved]

Question: What is point estimation?

Answer: Point estimation, in statistics, the process of finding an approximate value of some parameter - such as the mean - of a population from random samples of the population.

Question: What are the methods of point estimation?

Answer:

1. Method of Maximum Likelihood Estimation (MLE)
2. Method of variance
3. Method of Least square
4. Method of moments
5. Method of minimum chi-square
6. Method of inverse probability.

Question: Show that a sufficient statistic is unique.

Solution:

Let $x_1, x_2, \ldots, x_n$ be a sample from a density function $f(x, \theta)$.

Let t be a sufficient statistic for θ.

Let there be another distinct statistic t_1 which is also sufficient statistic for θ.

Thus

$$h(t, t_1 / \theta) = g(t, \theta)\, h(t_1 / t)$$
$$= g(t_1, \theta)\, h(t / t_1)$$

t and t_1 are thus functionally related as

$$t = K(t_1, \theta)$$

Since t and t_1 are function of sample values only, t is functionally related to t_1 only. Thus a sufficient statistic is unique.

Example 15.14: Let $x_1, x_2, \ldots, x_n$ be a random sample from $N(\mu, \sigma^2)$ population. Find sufficient estimators for μ and σ^2.

Solution:

We know that, the pdf of normal population is

$$f(x; \mu, \sigma^2) = \left(\frac{1}{\sigma\sqrt{2\pi}}\right) \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

We know,

$$L = f(x_1; \mu, \sigma^2)\, f(x_2; \mu, \sigma^2) \ldots f(x_n; \mu, \sigma^2)$$

$$\text{or, } L = \prod_{i=1}^{n} f(x_i; \mu, \sigma^2)$$

$$\text{or, } L = \prod_{i=1}^{n} \left[\frac{1}{\sigma\sqrt{2\pi}} \exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\}\right]$$

$$\text{or, } L = \left(\frac{1}{\sigma\sqrt{2\pi}}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\left(\sum_{i=1}^{n} x_i^2 - 2\mu\Sigma x_i + n\mu^2\right)\right\}$$

$$\text{or, } L = \left(\frac{1}{\sigma}\right)^n \exp\left\{-\frac{1}{2\sigma^2}\left(\Sigma x_i^2 - 2\mu\Sigma x_i + n\mu^2\right)\right\} \cdot \left(\frac{1}{\sqrt{2\pi}}\right)^n$$

$$\text{or, } L = g(t_1, t_2; \mu, \sigma^2) \cdot (h\, x_1, x_2, \ldots, x_n)$$

where, $t_1 = \Sigma x_i$ and $t_2 = \Sigma x_i^2$ are sufficient estimators for μ and σ^2. And $h(x) = \left(\frac{1}{\sqrt{2\pi}}\right)^n$.

3. If $f(x,\theta) = e^{-(x-\theta)}$; $x>0$, $-\infty<\theta<\infty$

Find sufficient statistic for θ.

<u>Solution:</u>

Given $f(x,\theta) = e^{-(x-\theta)}$

The joint distribution of $x_1, x_2, \ldots, x_n$ is

$$L = f(x_1,\theta) f(x_2,\theta) \cdots f(x_n,\theta)$$

$$\text{or, } L = \prod_{i=1}^{n} f(x_i,\theta)$$

$$\text{or, } L = \prod_{i=1}^{n} \left\{ e^{-(x_i-\theta)} \right\}$$

$$\text{or, } L = e^{-\sum_{i=1}^{n} x_i} \cdot e^{n\theta} \quad \ldots (1)$$

If $t = \min(x_1, x_2, \ldots, x_n) = x_{(1)}$

then pdf of T is given by

$$g(t,\theta) = n\left[1 - F(x_{(1)})\right]^{n-1} \cdot f(x_{(1)},\theta)$$

Here,

$$f(x_{(1)}) = e^{-(x_{(1)} - \theta)}$$

$$F(x_{(1)}) = \int_{\theta}^{x} e^{-(x_{(1)} - \theta)} dx$$

$$= \left| \frac{e^{-(x_{(1)} - \theta)}}{-1} \right|_{\theta}^{x}$$

$$= 1 - e^{-(x_{(1)} - \theta)}$$

Now, $g(t, \theta) = n\left[e^{-(x_{(1)} - \theta)}\right]^{n-1} \cdot e^{-(x_{(1)} - \theta)}$

or, $g(t, \theta) = ne^{-n(x_{(1)} - \theta)}$

From eqation ①, we get

$$L = e^{n\theta} \cdot e^{-\sum_{i=1}^{n} x_i}$$

or, $L = ne^{-n(x_{(1)} - \theta)} \cdot \left[\frac{e^{-\sum_{i=1}^{n} x_i}}{ne^{-nx_{(1)}}} \right]$

or, $L = g(t, \theta)\left[\dfrac{e^{-\Sigma x_i}}{n e^{-n x_{(1)}}}\right]$

or, $L = g(t, \theta) \cdot h(x_1, x_2, \ldots, x_n)$

Here $g(t, \theta)$ depends on θ and x but $h(x)$ is independent of θ.

Thus $t = \min(x_1, x_2, \ldots, x_n) = x_{(1)}$ is the sufficient statistic of θ.

4. $f(x,\theta) = \theta x^{\theta-1}$; $0<x<1$, $\theta>0$

Show that $t_1 = \prod_{i=1}^{n} x_i$, is sufficient for θ.

Solution:

Given

$$f(x,\theta) = \theta x^{\theta-1} ; \quad 0<x<1, \quad \theta>0$$

The joint distribution of $x_1, x_2, \ldots, x_n$ is

$$L = f(x_1,\theta) f(x_2,\theta) \cdots f(x_n,\theta)$$

$$\text{or, } L = \prod_{i=1}^{n} f(x_i,\theta)$$

$$\text{or, } L = \prod_{i=1}^{n} \left\{\theta x_i^{\theta-1}\right\}$$

$$\text{or, } L = \theta^n \prod_{i=1}^{n} \left(x_i^{\theta-1}\right)$$

$$\text{or, } L = \theta^n \left(\prod_{i=1}^{n} x_i\right)^{\theta} \cdot \frac{1}{\left(\prod_{i=1}^{n} x_i\right)}$$

or, $L = g(t, \theta) \cdot h(x_1, x_2, \ldots, x_n)$

Here $g(t, \theta)$ depends on x and θ but $h(x)$ is independent of θ.

Hence, $t_1 = \prod_{i=1}^{n} x_i$ is sufficient estimator for θ.

Likelihood Function:

Let $x_1, x_2, \ldots, x_n$ be a random sample of size n from a population with density function $f(x, \theta)$. Then the likelihood function of the sample values $x_1, x_2, \ldots, x_n$. usually denoted by $L = L(\theta)$ is their joint density function, given by

$$L = f(x_1, \theta)\, f(x_2, \theta) \ldots f(x_n, \theta)$$

$$\text{or, } L = \prod_{i=1}^{n} f(x_i, \theta)$$

L gives the relative likelihood that the random variables assume a particular set of values $x_1, x_2, \ldots, x_n$. For a given sample $x_1, x_2, \ldots, x_n$, L becomes a function of the variable θ, the parameter.

The principle of maximum likelihood estimator (MLE) is

$$L = L(\theta); \quad \frac{\partial L}{\partial \theta} = 0; \quad \frac{\partial^2 L}{\partial \theta^2} < 0$$

Example: In a random sampling a sample of size n from normal population $N(\mu, \sigma^2)$, find the MLE of

i) μ when σ^2 is known

ii) σ^2 when μ is known

iii) the simultaneous estimation of μ and σ^2

solution:

Given $x \sim N(\mu, \sigma^2)$

We know that, the pdf of normal distribution is

$$f(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x-\mu)^2\right\}$$

We know,

$$L = f(x_1; \mu, \sigma^2)\, f(x_2; \mu, \sigma^2) \ldots f(x_n; \mu, \sigma^2)$$

or, $$L = \prod_{i=1}^{n} f(x_i; \mu, \sigma^2)$$

or, $$L = \prod_{i=1}^{n} \left[\frac{1}{\sqrt{2\pi\sigma^2}} \exp\left\{-\frac{1}{2\sigma^2}(x_i-\mu)^2\right\}\right]$$

or, $$L = \left(\frac{1}{2\pi\sigma^2}\right)^{n/2} \exp\left\{-\frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2\right\}$$

or, $\text{Log}\, L = -\frac{n}{2}\log(2\pi) - \frac{n}{2}\log\sigma^2 - \frac{1}{2\sigma^2}\sum_{i=1}^{n}(x_i-\mu)^2$

Case - i:

When σ^2 is known, the MLE of μ is

$$\frac{\partial \log L}{\partial \mu} = 0$$

or, $-0-0-\frac{1}{2\sigma^2}\sum_{i=1}^{n} 2(x_i-\mu)(-1) = 0$

or, $\sum_{i=1}^{n}(x_i-\mu) = 0$

or, $\sum(x_i) - n\mu = 0$

or, $n\mu = \sum x_i$

or, $\hat{\mu} = \frac{1}{n}\sum x_i = \bar{x}$

Hence MLE of μ is the sample mean $\bar{x}$.

Case-ii:

When μ is known, the MLE for σ^2 is

$$\frac{\partial \log L}{\partial \sigma^2} = 0$$

or, $-\frac{n}{2}\cdot\frac{1}{\sigma^2} + \frac{1}{2\sigma^4}\sum_{i=1}^{n}(x_i-\mu)^2 = 0$

or, $n - \frac{1}{\sigma^2} \sum (x_i - \mu)^2 = 0$

or, $\hat{\sigma}^2 = \frac{1}{n} \sum (x_i - \mu)^2$

Case III:

The likelihood equations for simultaneous estimation of μ and σ^2 are

$$\frac{\partial \log L}{\partial \mu} = 0 \quad \text{and} \quad \frac{\partial \log L}{\partial \sigma^2} = 0$$

thus giving

$$\hat{\mu} = \bar{x}$$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \mu)^2$$

or, $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x})^2 = s^2$

which is the sample variance.

1. If $f(x, \lambda) = \lambda e^{-\lambda x}$; $0 < x < \infty$

Find MLE of λ.

Solution:

Given $f(x, \lambda) = \lambda e^{-\lambda x}$

We know,

$$L = f(x_1, \lambda) f(x_2, \lambda) \dots f(x_n, \lambda)$$

or, $$L = \prod_{i=1}^{n} \left\{ \lambda e^{-\lambda x_i} \right\}$$

or, $$L = \lambda^n e^{-\lambda \sum_{i=1}^{n} x_i}$$

or, $$\log L = n \log \lambda - \lambda \sum_{i=1}^{n} x_i$$

or, $$\frac{\partial \log L}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

or, $$0 = \frac{n}{\lambda} - \sum_{i=1}^{n} x_i$$

or, $\frac{n}{\lambda} = \sum_{i=1}^{n} x_i$

or, $\lambda = \frac{n}{\sum x_i}$

or, $\lambda = \frac{1}{\frac{\sum x_i}{n}}$

or, $\hat{\lambda} = \frac{1}{\bar{x}}$

Hence MLE for λ is the inverse of sample mean $\bar{x}$.

2. If $f(x,\theta) = \frac{1}{\theta} e^{-x/\theta}$; $\theta > 0$, $x > 0$

Find MLE of θ.

Solution:

Given $f(x,\theta) = \frac{1}{\theta} e^{-x/\theta}$

we know,

$$L = f(x_1,\theta) f(x_2,\theta) \dots f(x_n,\theta)$$

$$\text{or, } L = \prod_{i=1}^{n} \left\{ \frac{1}{\theta} e^{-x_i/\theta} \right\}$$

$$\text{or, } L = \left(\frac{1}{\theta}\right)^n \exp\left\{ -\frac{1}{\theta} \sum_{i=1}^{n} x_i \right\}$$

$$\text{or, } \log L = -n \log \theta - \frac{1}{\theta} \sum_{i=1}^{n} x_i$$

$$\text{or, } \frac{\partial \log L}{\partial \theta} = -\frac{n}{\theta} + \frac{1}{\theta^2} \sum_{i=1}^{n} x_i$$

or, $0 = -\frac{n}{\theta} + \frac{1}{\theta^2}\sum_{i=1}^{n} x_i$

or, $n = \frac{1}{\theta}\sum_{i=1}^{n} x_i$

or, $\theta = \frac{1}{n}\sum_{i=1}^{n} x_i$

or, $\hat{\theta} = \bar{x}$

Hence MLE for θ is the sample mean $\bar{x}$.

5. $P(X=x) = f(x,\lambda) = \dfrac{e^{-\lambda}\lambda^{x}}{x!}$; $x = 0, 1, 2, \ldots$

① Find MLE of λ. Also find its variance

Solution:

Given

$$P(X=x) = f(x,\lambda) = \frac{e^{-\lambda}\lambda^{x}}{x!} \; ; \quad x = 0, 1, 2, \ldots$$

The joint distribution of $x_1, x_2, \ldots, x_n$ is

$$L = f(x_1,\lambda)\, f(x_2,\lambda) \ldots f(x_n,\lambda)$$

$$\text{or, } L = \prod_{i=1}^{n} f(x_i,\lambda)$$

$$\text{or, } L = \prod_{i=1}^{n} \left(\frac{e^{-\lambda}\lambda^{x_i}}{x_i!} \right)$$

$$\text{or, } L = \frac{e^{-n\lambda}\lambda^{\sum x_i}}{x_1!\, x_2! \ldots x_n!}$$

or, $\log L = -n\lambda + \left(\sum_{i=1}^{n} x_i\right) \log\lambda - \sum_{i=1}^{n} \log(x_i!)$

or, $\log L = -n\lambda + n\cdot\left(\frac{1}{n}\sum x_i\right)\log\lambda - \sum \log(x_i!)$

or, $\log L = -n\lambda + n\bar{x}\log\lambda - \sum_{i=1}^{n} \log(x_i!)$

or, $\frac{\partial \log L}{\partial \lambda} = -n + \frac{n\bar{x}}{\lambda}$

or, $0 = -n + \frac{n\bar{x}}{\lambda}$

or, $\hat{\lambda} = \bar{x}$

Hence the MLE for $\hat{\lambda}$ is sample mean $\bar{x}$.

The variance of the estimate is given by

$$\frac{1}{V(\hat{\lambda})} = E\left[-\frac{\partial^2 \log L}{\partial \lambda^2}\right]$$

$$\text{or, } \frac{1}{V(\hat{\lambda})} = E\left[-\frac{\partial}{\partial \lambda}\left(-n + \frac{n\bar{x}}{\lambda}\right)\right]$$

$$\text{or, } \frac{1}{V(\hat{\lambda})} = E\left[-\left(-\frac{n\bar{x}}{\lambda^2}\right)\right]$$

$$\text{or, } \frac{1}{V(\hat{\lambda})} = \frac{n}{\lambda^2} E(\bar{x}) = \frac{n}{\lambda}$$

$$\therefore V(\hat{\lambda}) = \frac{\lambda}{n}$$

ii) Show that the sample mean $\bar{x}$, is sufficient for estimating the paramter λ of the poisson distribution.

Solution:

We have

$$\frac{\partial \log L}{\partial \lambda} = -n + \frac{n\bar{x}}{\lambda}$$

$$= n\left(\frac{\bar{x}}{\lambda} - 1\right)$$

$$= \psi(\bar{x}, \lambda).$$

where, $\psi(\bar{x}, \lambda)$ is a function of $\bar{x}$ and λ only.

Hence, sample mean $\bar{x}$ is sufficient for estimating λ.

6. Find MLE of θ in the following case:

$$f(x,\theta) = {}^nC_x \, \theta^x (1-\theta)^{n-x} \; ; \; x = 0, 1, 2, \ldots, n$$

<u>Solution:</u>

Given $f(x,\theta) = {}^nC_x \, \theta^x (1-\theta)^{n-x}$

We know,

$$L = f(x_1, \theta) \, f(x_2, \theta) \ldots f(x_n, \theta)$$

$$\text{or, } L = \prod_{i=1}^{n} f(x_i, \theta)$$

$$\text{or, } L = \prod_{i=1}^{n} \left\{ {}^nC_{x_i} \, \theta^x (1-\theta)^{n-x} \right\}$$

$$\text{or, } L = \left[\prod_{i=1}^{n} {}^nC_{x_i} \right] \theta^{\sum_{i=1}^{n} x_i} (1-\theta)^{n^2 - \sum x_i}$$

$$\text{or, } \log L = \sum_{i=1}^{n} \log\left({}^nC_{x_i}\right) + \sum x_i \log \theta + (n^2 - \sum x_i) \log(1-\theta)$$

$$\text{or, } \frac{\partial \log L}{\partial \theta} = 0 + \frac{\sum x_i}{\theta} - \frac{n^2 - \sum x_i}{1-\theta}$$

or, $0 = \frac{\sum x_i}{\theta} - \frac{n^2 - \sum x_i}{1-\theta}$ $\left[\because \frac{\partial \log L}{\partial \theta} = 0\right]$

or, $\theta = \frac{\sum x_i}{n^2}$

or, $\theta = \frac{1}{n} \cdot \frac{\sum x_i}{n}$

or, $\hat{\theta} = \frac{\bar{x}}{n}$

Hence MLE for θ is the sample mean $\bar{x}$ divided by sample size n.

7. Find MLE of θ for a random sample of size n from the distribution

$$f(x,\theta) = \begin{cases} (\theta+1)x^{\theta}; & 0 \le x \le 1, \ \theta > -1 \\ 0, & \text{otherwise} \end{cases}$$

Show that it is also sufficient statistic for θ.

<u>Solution:</u>

Given,

$$f(x,\theta) = \begin{cases} (\theta+1)x^{\theta}; & 0 \le x \le 1, \ \theta > -1 \\ 0, & \text{otherwise} \end{cases}$$

We know,

$$L = f(x_1,\theta)\, f(x_2,\theta) \ldots f(x_n,\theta)$$

or, $L = \prod_{i=1}^{n} f(x_i,\theta)$

or, $L = \prod_{i=1}^{n} (\theta+1)\, x_i^{\theta}$

or, $L = (\theta+1)^n \prod_{i=1}^{n} x_i^{\theta}$

or, $\log L = n\log(\theta+1) + \theta\sum_{i=1}^{n}\log x_i$

or, $\frac{\partial \log L}{\partial \theta} = \frac{n}{\theta+1} + \sum_{i=1}^{n}\log x_i$

or, $0 = \frac{n}{\theta+1} + \sum \log x_i \quad \left[\because \frac{\partial \log L}{\partial \theta} = 0\right]$

or, $\hat{\theta} = -\frac{n}{\sum_{i=1}^{n}\log x_i} - 1$

<u>Check sufficiency:</u>

$$\frac{\partial \log L}{\partial \theta} = \frac{n}{\theta+1} + \sum \log x_i$$

$$= \left(\frac{n}{\theta+1} + \sum \log x_i\right)\cdot 1$$

$$= \psi(x_i, \theta)\cdot h(x_1, x_2, \ldots, x_n)$$

where, $\psi(x_i, \theta)$ is a function of x_i and θ only. But $h(x)$ is independent of θ.

Hence, $\hat{\theta} = \left[-\frac{n}{\sum \log x_i} - 1\right]$ is a sufficient statistic for θ.

Hypothesis Testing

Definition: A statistical Hypothesis is a statement or assumption regarding one or more population parameters. Our aim in hypothesis testing is to verify whether the hypothesis is true or not on the basis of sample data.

Null hypothesis and Alternative hypothesis: The null hypothesis, denoted by H_0, is the hypothesis that is to be tested: The alternative hypothesis, denoted by H_1, is the hypothesis that, in some sense, contradicts the null hypothesis.

Type I error: Reject null hypothesis when it is true is called type I error.

Type II error: Accept null hypothesis when alternative hypothesis is true is called type II error.

Decision	H_0 is true	H_1 is true
Reject H_0	Type I error $P(\text{Type I error}) = \alpha$	Correct decision $P(\text{Correct decision}) = 1-\beta$
Accept H_0	Correct decision $P(\text{Correct decision}) = 1-\alpha$	Type II error $P(\text{Type II error}) = \beta$

<u>Level of significance</u>: The probability of committing type I error is usually denoted by α and is commonly referred to as the level of significance of a test.

* $P(\text{Reject } H_0 \mid H_1) = 1-\beta$ is called power of the test.

<u>One tailed test</u>:

$$H_0 : \mu = \mu_0$$

$H_1 : \mu < \mu_0$ (left tailed test) or

$H_1 : \mu > \mu_0$ (Right tailed test)

Two tailed test:

$$H_0 : \mu = \mu_0$$
$$H_1 : \mu \neq \mu_0$$

Steps in a statistical test:

1. set up the null hypothesis (H_0) and its alternative (H_1). It is a one-tailed test if the alternative hypothesis states the direction of difference. If no direction of difference is given, it is a two-tailed test.

2. Choose the desired level of significance α. While $\alpha = 0.05$ and $\alpha = 0.01$ are the most common, many others are also used.

3. Compute the appropriate test statistic (normal, t etc.) from the sample data.

4. Find the critical value(s) using normal integral tables corresponding to the critical region established.

5. With the critical values determined in step 4, compare the test statistic computed in step 3.

6. Make decision: reject the null hypothesis if the computed test statistic falls in the critical region and accept the alternative (or withhold decision).

<u>Select test statistic:</u>

When $n \geq 30$ or standard deviation (σ) is known, use z-test.

When $n < 30$ and σ is unknown, use t-test

<u>Formula:</u>

$$z = \frac{\bar{x} - \mu_0}{\sqrt{\sigma^2 / n}} = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma}$$ is called calculated value of z (cal. z).

$$t = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s}$$

<u>Example:</u> A dairy farm study concludes that school going children consume an average at most 19.4 ounces of milk per day. In a survey of 140 randomly selected children, it is found that they drink an average of 18.5 ounces with a standard deviation of 6.8 ounces.

Do you support the claim of the farm study that the children consume 19.4 ounces milk per day?

<u>Solution:</u>

The null and alternative hypotheses are

$$H_0 : \mu = 19.4$$

$$H_1 : \mu < 19.4$$

The sample data give

$$\bar{x} = 18.5$$

$$s = 6.8$$

$$n = 140$$

since $n \geq 30$, we use z-test

$$z = \frac{\sqrt{n}\,(\bar{x} - \mu_0)}{s} = \frac{\sqrt{140}\,(18.5 - 19.4)}{6.8} = -1.57$$

$\therefore$ Calculated value of z is -1.57

The tabulated value of z at 5% level of significance is -1.64 (for left tailed test).

Since, cal.(z) > tabulated value of z,

hence the calculated value fall in the acceptence region, therefore we fail to reject H_0.

Thus the school going children may consume an average 19.4 ounces milk per day.

Example: A random sample of 15 men from a large community yields a mean height of 66.4 inches. The value of s calculated from the sample is 3.1 inches. Do the data support the hypothesis that the mean height of men in the community is 65.0 inches?

Solution:

$H_0: \mu = \mu_0 = 65$

$H_1: \mu \neq 65$

The sample data give

$\bar{x} = 66.4$

$s = 3.1$

$n = 15$

since $n < 30$ and σ is unknown, we use t-test

$$t = \frac{\bar{x} - \mu_0}{s/\sqrt{n}} = \frac{66.4 - 65}{3.1/\sqrt{15}} = 1.75$$

∴ Calculated value of t is 1.75

The tabulated value of t at 5% level of significance is -2.145 to 2.145 (for two tailed test).

Since the calculated value of t lies between -2.145 to 2.145 at 5% level of significance with (15-1) or 14 df, we fail to reject H_0.

Contingency Table

Question: What is manifold classification?

Answer: Let us consider two attributes A and B, A divided into r classes $A_1, A_2, \ldots, A_r$ and B divided into s classes $B_1, B_2, \ldots, B_s$. Such a classification in which attributes are divided into more than two classes is known as manifold classification.

A \ B	B_1	B_2	$\cdots$	B_j	$\cdots$	B_s	
A_1	A_1B_1	A_1B_2	$\cdots$	A_1B_j	$\cdots$	A_1B_s	$\sum_{j=1}^{s} A_1B_j$ is called marginal frequency
A_2	A_2B_1	A_2B_2	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_i	A_iB_1	$\vdots$	$\cdots$	A_iB_j	$\cdots$	A_iB_s	$\sum_{i=1,j=1}^{r,s} A_iB_j$
$\vdots$	$\vdots$	$\vdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$	$\cdots$
A_r	A_rB_1	$\vdots$	$\cdots$	A_rB_j	$\cdots$	A_rB_s	$\cdots$
	$\sum_{i=1}^{r} A_iB_1$	$\cdots$	$\cdots$	$\cdots$	$\cdots$		$N \leftarrow$ Total population

$r \times s$ contingency table

Question: What is contingency table?

Answer: A table of data in which the row entries tabulate the data according to one variable (attribute) and the colum entries tabulate the data according to another variable (attribute) and which is used especially in the study of the correlation between variables.

$r \times s$ manifold contingency table where (A_i) is the number of persons possessing the attribute A_i $(i=1,2,\ldots,r)$ and (B_j) is the number of persons possessing attribute B_j $(j=1,2,\ldots,s)$ and (A_iB_j) is the number of persons possessing both the attributes

$$\sum_{i=1}^{r} A_i = \sum_{j=1}^{s} B_j = N$$

Under the null hypothesis, that attributes are independent

$$P(A_i) = \frac{(A_i)}{N}, \quad P(B_j) = \frac{(B_j)}{N} \quad \left[\begin{array}{l} i = 1, 2, \ldots, r \\ j = 1, 2, \ldots s \end{array} \right.$$

$$\therefore P(A_i B_j) = P(A_i)\, P(B_j) = \frac{(A_i)}{N} \cdot \frac{(B_j)}{N}$$

$\therefore (A_i B_j)_0$ = Expected number of persons possessing both attributes

$$= N \cdot P(A_i B_j) = \frac{(A_i)\,(B_j)}{N}$$

The exact test statistic

$$\chi^2 = \sum_{i=1}^{r} \sum_{j=1}^{s} \frac{\{(A_i B_j) - (A_i B_j)_0\}^2}{(A_i B_j)_0} \sim \chi^2 \text{ distribution}$$

with $(r-1)(s-1)$ degree of freedom.

If cal. $\chi^2 \geq$ Tab. χ^2 at 5% level of significance with $(r-1)(s-1)$ df, then we may reject H_0 otherwise fail to reject.

Example: For 2×2 contingency table

a	b
c	d

, prove that

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

Proof:

The test statistic

$$\chi^2 = \sum \frac{(O-E)^2}{E} \sim \chi^2 \text{ distribution with } (2-1)(2-1) = 1 \times 1 = 1 \text{ df.}$$

where,

O = observed cell frequency

E = expected cell frequency

$$\therefore \chi^2 = \frac{\{a-E(a)\}^2}{E(a)} + \frac{\{b-E(b)\}^2}{E(b)} + \frac{\{c-E(c)\}^2}{E(c)} + \frac{\{d-E(d)\}^2}{E(d)}$$

We know,

$$\text{Expected cell frequency, } E = \frac{\text{Row total} \times \text{Column total}}{\text{Grand total}}$$

			Total
	a	b	$a+b$
	c	d	$c+d$
Total	$a+c$	$b+d$	$N = a+b+c+d$

Now,

$$E(a) = \frac{(a+b)(a+c)}{N}, \quad E(b) = \frac{(a+b)(b+d)}{N}$$

$$E(c) = \frac{(a+c)(c+d)}{N}, \quad E(d) = \frac{(c+d)(b+d)}{N}$$

$$a - E(a) = a - \frac{(a+b)(a+c)}{N} = a - \frac{(a+b)(a+c)}{a+b+c+d}$$

$$\text{or, } a - E(a) = \frac{\cancel{a^2} + \cancel{ab} + \cancel{ac} + ad - \cancel{a^2} - \cancel{ac} - \cancel{ab} - bc}{a+b+c+d}$$

$$\text{or, } a - E(a) = \frac{ad - bc}{N} \qquad [\because N = a+b+c+d$$

similarly,

$$b - E(b) = -\frac{ad-bc}{N}$$

$$c - E(c) = -\frac{ad-bc}{N}$$

$$d - E(d) = \frac{ad-bc}{N}$$

Now,

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

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

$$\text{or, } \chi^2 = \frac{(ad-bc)^2}{N}\left[\frac{b+d+a+c}{(a+b)(a+c)(b+d)} + \frac{b+d+a+c}{(a+c)(c+d)(b+d)}\right]$$

$$\text{or, } \chi^2 = \frac{(ad-bc)^2}{N} \times N\left[\frac{1}{(a+b)(a+c)(b+d)} + \frac{1}{(a+c)(c+d)(b+d)}\right]$$

$$\text{or, } \chi^2 = (ad-bc)^2 \cdot \left[\frac{c+d+a+b}{(a+b)(a+c)(b+d)(c+d)}\right]$$

$$\text{or, } \chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)} \quad [\because N = a+b+c+d]$$

[Proved].

An important special case of the two-way contingency table emerges when both variables are dichotomous. The resulting table is a 2 by 2 or a four-fold table. The accompanying table is a typical one of this form.

	Categories of Y		
Categories of X	1	2	Row total
1	n_{11}	n_{12}	m_1
2	n_{21}	n_{22}	m_2
Column total	n_1	n_2	n

The formula for χ^2 for such a table can be written as

$$\chi^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{m_1 m_2 n_1 n_2}$$

Example:

Beautification	Mood status: Moody	Not Moody	Total
Beauty	13	8	$m_1 = 21$
Not beauty	5	14	$m_2 = 19$
Total	$n_1 = 18$	$n_2 = 22$	$n = 40$

Test whether the mood of females is independent of their beauty.

Solution:

H_0 = The mood of females is independent of their beauty.

$H_1 = H_0$ is not true

We, know,

$$\chi^2 = \frac{n(n_{11}n_{22} - n_{12}n_{21})^2}{m_1 m_2 n_1 n_2}$$

$$\text{or, } \chi^2 = \frac{40(182-40)^2}{21 \times 19 \times 18 \times 22} = 5.1$$

At 5% level of significance with $(2-1)(2-1) = 1$ df the tabulated value of χ^2 is 3.84.

Here cal. $\chi^2 >$ tab. χ^2

<u>Comment:</u> Since the calculated value is greater than the tabulated value at 5% level of significance with 1 df.

Hence we reject H_0.

That means there is a relation between mood of females and their beauty.

We know that,

$$\chi^2 = \sum \frac{(O-E)^2}{E} \quad \ldots \text{ (a)}$$

where, O = observed cell frequency

E = Expected cell frequency

The folmula can be further simplified as below:

$$\chi^2 = \left(\sum \frac{O^2}{E}\right) - n \quad \ldots \text{ (b)}$$

sample size

Let's apply formula (b)

<u>Example:</u> Is 1st year credit independent of student's financial status?

	1st yr. credit		
Financial status	<3.5	≥3.5	total
Very good	18	26	44
Average	15	20	35
Not goot	10	8	18
Total	43	54	n=97

$E(18) = \frac{44 \times 43}{97} = 19.5$, $E(26) = \frac{44 \times 54}{97} = 24.5$

$E(15) = \frac{35 \times 43}{97}$ 15.5, $E(20) = \frac{35 \times 54}{97} = 19.48$

$E(10) = \frac{18 \times 43}{97} = 7.98$, $E(8) = \frac{18 \times 54}{97} = 10.02$

Now,

$$\chi^2 = \left(\sum \frac{O^2}{E}\right) - n$$

or, $\chi^2 = \frac{18^2}{19.5} + \frac{26^2}{24.5} + \frac{15^2}{15.5} + \frac{20^2}{19.48} + \frac{10^2}{7.98} + \frac{8^2}{10.02} - 97$

or, $\chi^2 = 1.18$

Here, H_0 = 1s year credit is independent of student's financial status

H_1 = H_0 is not true

At 5% level of significance with $(3-1)(2-1) = 2$ df the tabulated value of χ^2 is 5.99

Here cal. $\chi^2 <$ tab. χ^2

<u>comment:</u> Since the calculated value is less than the tabulated value at 5% level of significance with 2 df.

Hence, we fail to reject H_0

That means, 1st year credit is independent of student's financial condition.

Yate's Correction:

If any cell frequency is less than 5, we need Yate's correction.

i.e. a or b or c or d < 5

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

Non-parametric Test

Question: Define non-parametric test with example.

Definition: Non-parametric test is a test that does not depend on the particular form of the basic frequency function from which the samples are drawn. In other words, non-parametric test does not make any assumption regarding the form of the population.

Some non-parametric tests are:

1. Wald-Wolfowitz Run Test
2. Median Test
3. Sign Test
4. Mann-Whitney-Wilcoxon U-Test

Question: Write down the advantages and disadvantages of non-parametric test over parametric test.

Answer:

Advantages of non-parametric tests:

① N.P. tests are readily comprehensible, very simple and easy to apply and do not require complicated sample theory.

② No assumption is made about the form of the frequency function of the parent population from which sampling is done.

③ No parametric technique will apply to the data which are mere classification, while N.P. tests exist to deal with such data.

④ Since the socio-economic data are not, in general, normally distributed, N.P. tests have found applications in Psychometry, Sociology and Educational statistics.

⑤ N.P. tests are available to deal with the data which are given in ranks or whose seemingly numerical scores have the strength of ranks. For instance, no parametric test can be applied if the scores are given in grades such as A^+, A, A^-, B etc.

Disadvantages of N.P. tests

① N.P. tests can be used only if the measurements nominal or ordinal. Even in that case, if a parametric test exists it is more powerful than the N.P. test.

② So far, no N.P. tests exist for testing interactions in 'Analysis of Variance' model unless special assumptions about the additivity of the model are made.

③ N.P. tests are designed to test statistical hypothesis only and not for estimating the parameters.

Median Test

Definition: Median test is a statistical procedure of non-parametric test which compare whether two independent samples are same or two independent samples are drawn from same population.

https://www.youtube.com/watch?v=cpoRCpKMq-w&t=737s&ab_channel=StatisticsLearning

Procedure: Let $x_1, x_2, \ldots, x_n$, be a random sample comes from a population with pdf $f_1(x)$ and $y_1, y_2, \ldots, y_{n_2}$ be another random sample comes from a population with pdf $f_2(y)$.

Test whether $f_1(x) = f_2(y)$.

Solution:

$$H_0 : f_1(x) = f_2(y)$$

$$H_1 : f_1(x) \neq f_2(y)$$

Let $z_1, z_2, \ldots, z_{n_1+n_2}$ be the combined ordered sample.

Let m_1 be the number of x's and m_2 be the number of y's exceeding the median of the combined sample.

Then under $H_0: f_1(x) = f_2(y)$, the joint distribution of m_1 and m_2 is the hypergeometric distribution with probability function:

$$p(m_1, m_2) = \frac{\binom{n_1}{m_1}\binom{n_2}{m_2}}{\binom{n_1+n_2}{m_1+m_2}}$$

If $m_1 < \frac{n_1}{2}$, then the critical region corresponding to the size of type-1 error α, is give by $m_1 < m_1'$, where m_1' is is computed from the equation

$$\sum_{m_1=1}^{m_1'} p(m_1, m_2) \propto$$

The distribution of m_1 under H_0 is also hyper-geometric with

$$E(m_1) = \frac{n_1}{2}, \text{ if } N = n_1 + n_2 \text{ is even}$$
$$= \frac{n_1}{2} \cdot \frac{N-1}{N}, \text{ if } N \text{ is odd}$$

$$V(m_1) = \frac{n_1 n_2}{4(N-1)}, \text{ if } N \text{ is even}$$
$$= \frac{n_1 n_2 (N+1)}{4N^2}, \text{ if } N \text{ is odd}$$

But for large sample size, the distribution follows asymptotically normal distribution with mean 0 and variance 1.

$$\therefore z = \frac{m_1 - E(m_1)}{\sqrt{V(m_1)}} \sim N(0,1)$$

If the calculated value lies between -1.96 to 1.96 at 5% level of significance, we may accept H_0 otherwise reject.

Random Examples

2016: 4.b

b) The average IQ of university female students in Bangladesh is suspected to be less than the average 110 for all students. A random sample of 64 female students yielded a sample average IQ of 105 and standard deviation of 20. Can you conclude that the average score of the female students is really less than 110? [$Z_{0.05}$=1.64]

Solution:

Given that

$\mu = 110$

$n = 64$

$\bar{x} = 105$

$s = 20$

Let

$H_0 : \mu = \mu_0 = 110$

$H_1 : \mu < 110$

As $n \geq 30$, we use Z-test

$$\therefore Z = \frac{\bar{x} - \mu}{s/\sqrt{n}} = \frac{105 - 110}{20/\sqrt{64}} = -2$$

Critical value, $Z_\alpha = -1.64$

Since $|Z| > |Z_\alpha|$, we reject null hypothesis.

Hence, the average score of the female students is really less than 110.

2019: 4(c): A sample of 900 items taken from population with standard deviation 2.61 cms. The mean of the sample is 3.4 cms. Test whether the sample come from the population with mean 3.25 cms.

Solution:

The null and alternative hypotheses are

$H_0: \mu = \mu_0 = 3.25$

$H_1: \mu \neq 3.25$ (two tail test)

The sample data give, $\bar{x} = 3.4$, $s = 2.61$ and $n = 900$

Since $n \geq 30$, we use z test.

$$z = \frac{\sqrt{n}(\bar{x} - \mu_0)}{s} = \frac{\sqrt{900}\,(3.4 - 3.25)}{2.61} = 1.724$$

$\therefore$ calculated value of z is 1.724

The tabulated value of z at 5% level of significance is -1.96 to $+1.96$.

Since, the calculated value of z lies between -1.96 to $+1.96$ at 5% level of significance.

We fail to reject H_0 null hypotheses.

That means, the sample come from the population with mean 3.25 cms.

CS CamScanner

2018: 4(c): A bulb manufacturing company claims that the average longevity of their bulb is 3.65 years. with a standard deviation of 0.16 years. A random sample of 36 bulbs gave a mean longevity of 4.45 years. Does the sample mean justify the claim of the manufacturer?

Solution:

$$H_0: \mu = \mu_0 = 3.65$$

$$H_1: \mu \neq 3.65$$

The sample data give $\bar{x} = 4.45$, $\sigma^{\bar{x}} = 0.16$ and $n = 36$

Since, $\sigma^{\bar{x}}$ is known, we use z-test.

$$Z = \frac{\sqrt{n}(\bar{x} - \mu_0)}{\sigma^{\bar{x}}} = \frac{\sqrt{36}\,(4.45 - 3.65)}{0.16} = 30$$

Tab.(Z) is -1.96 to $+1.96$ at 5% level of significance

Since, cal.(Z) lies in the critical region.

We reject H_0 (null hypotheses).

That means, the claim of longevity of their ~~bulb~~ bulb is 3.65 years is not true.

2017: 5(a): Researchers are interested in the mean level of some enzyme in a certain population. They take a sample of 10 individuals, determine the level of enzyme in each and compute a sample mean 22. It is known that the variable of interest is approximately normally distributed with a variance of 45. Can you conclude that the mean enzyme level in this population is different from 25 at the 5% level of sig.?

Solution:

$$H_0: \mu = \mu_0 = 25$$

$$H_1: \mu \neq 25$$

The sample data give $\bar{x} = 22$, $\sigma^2 = 45$ and $n = 10$

Since, σ^2 is known, we use z-test

$$z = \frac{(\bar{x} - \mu_0)}{\sqrt{\sigma^2 / n}} = \frac{(22 - 25)}{\sqrt{45/10}} = -1.41$$

Tab.(Z) is -1.96 to $+1.96$ at 5% level of significance

Since, cal.(Z) lies between the -1.96 to $+1.96$ interval we fail to reject H_0 (null hypotheses).

That means, the mean enzyme level in this population is not different from 25 at 5% level of significance.

CamScanner

2015: 4(b): A random sample of 10 persons is selected as follows: 5, 2, 0, 4, 16, 14, 10, 11, 6, 8. Do you think that the average schooling year of the persons in population is 5? (Tabulated value at 5% with 9 d.f. is 2.26)

Solution: $\bar{X} = \frac{5+2+0+4+16+14+10+11+6+8}{10} = 7.6$

$$s^2 = \frac{(5-7.6)^2+(2-7.6)^2+(0-7.6)^2+(4-7.6)^2+(8.4)^2+(6.4)^2+(2.4)^2+(3.4)^2+(-1.6)^2+(.4)^2}{10-1}$$

or, $s^2 = 20.29$

$H_0: \mu = \mu_0 = 5$

$H_1: \mu \neq 5$

We have, $n = 10$, $s^2 = 20.29$ and $\bar{x} = 7.6$

Since, $n < 30$ and σ^2 is unknown, we use t-test.

$$t = \frac{(\bar{x} - \mu_0)}{\sqrt{s^2/n}} = \frac{7.6-5}{\sqrt{20.29/10}} = 1.83$$

Tab. (t) is -2.26 to 2.26 at 5% level of significance with 9 d.f.

Since, cal. (t) lies between the -2.26 to 2.26 interval, we fail to reject H_0.

That means, the average schooling year of the persons in population is 5

Class test: 2(c): Suppose a baker claims that his bread height is more than 15 cm, on the average. several of his customers do not beleive him. To persuade his customers that he is right, the baker decides to do a hypothesis test. He bakes 10 loaves of bread. The average height of the sample loaves is 17 cm. The baker knowns from baking hundreds of loaves of bread that the standard deviation for the height is 0.5 cm.

Solution:

$$H_0: \mu = \mu_0 = 15$$

$$H_1: \mu > 15$$

The sample data give $\bar{x} = 17$, $\sigma = 0.5$ and $n = 10$

since, standard deviation (σ) is known, we use Z-test

$$Z = \frac{\sqrt{n}(\bar{x} - \mu)}{\sigma} = \frac{\sqrt{10}(17 - 15)}{0.5} = 12.65$$

Tab. (Z) is 1.64 at 5% level of significance

since, cal. (Z) > Tab. (Z)

Hence, we reject H_0 (null hypothesis)

That means, the claim of the baker is right.

CamScanner

(c) Two sample polls of votes for two candidates A and B for a public office are taken, one from among the residents of rural areas. The results are given in the adjoining table. Examine whether the nature of the area is related to voting preference in the election.

Area	Votes for		Total
	A	B	
Rural	620	380	1000
Urban	550	450	1000
Total	1170	830	2000

[Value of the cumulative distribution function of Chi-square at 10.09 with 1 d.f is 0.0015]

Solution:

$E(620) = \frac{1170 \times 1000}{2000} = 585$ $E(380) = \frac{830 \times 1000}{2000} = 415$

Solution:

Here, $a = 620$, $b = 380$, $c = 550$, $d = 450$, $N = 2000$

Let,
H_0 = The nature of the area is not related to voting preference in the election.
$H_1 = H_0$ is not true.

We know,
$$\chi^2 = \frac{N(ad-bc)^2}{(a+b)(a+c)(b+d)(c+d)}$$
$$= \frac{2000(620 \cdot 450 - 380 \cdot 550)^2}{(620+380)(620+550)(380+450)(550+450)}$$
$$= 10.09$$

Tab. (χ^2) is 3.84 at 5% with $(2-1)(2-1) = 1$ d.f.

Here, cal. (χ^2) > Tab. (χ^2)

Comment: From the above calculation,

we Reject H_0.

That means the nature of the area is ~~te~~ related to voting preference in the election.

CamScanner

2016: 5(b): In a psychological test, out of 100 boys came out successful while 60 out of 100 girls of the same age group as the boys passed the test. Do the data provide any evidence of difference in respect of abilities between the genders?

Solution:

First we form the table

Test	Gender		Total
	Boys	Girls	
Passed	80	60	140
Not Passed	20	40	60
Total	100	100	200

Here, $a = 80$, $b = 60$, $c = 20$, $d = 40$, $N = 200$

Let, H_0 = There is no difference in respect of abilities between the genders.

$H_1 = H_0$ is not true.

$$\chi^2 = \frac{200(80 \times 40 - 60 \times 20)}{140 \times 60 \times 100 \times 100} = 9.52$$

tab.(χ^2) is 3.84 at 5% with $(2-1)(2-1) = 1$ d.f.

Hence, cal.(χ^2) > tab.(χ^2).

Comment: From above calculation, we reject H_0. that means, there is difference in respect of abilities between the genders.

(c) In the contingency table below, the two categorical variables are gender and ice cream flavor preference. This is a two-way table (2 X 3) where each cell represents the number of times males and females prefer a particular ice cream flavor. 3

Gender	Chocolate	Strawberry	Vanilla	Total
Female	37	17	12	66
Male	21	18	32	71
Total	58	35	44	137

Is there any relationship between gender and flavor preference?
(Chi-square value at $\alpha = 0.05$ is 3.99)

Solution:

$E(37) = \frac{66 \times 58}{137} = 27.94$

$E(21) = \frac{71 \times 58}{137} = 30.06$

$E(17) = \frac{66 \times 35}{137} = 16.86$

$E(18) = \frac{71 \times 35}{137} = 18.14$

$E(12) = \frac{66 \times 44}{137} = 21.19$

$E(32) = \frac{71 \times 44}{137} = 22.8$

Let, H_0 = There is no relationship between gender and flavor preference
H_1 = H_0 is not true

We know,

$$\chi^2 = \left(\Sigma \frac{O^2}{E}\right) - n$$

or, $\chi^2 = \frac{37^2}{27.94} + \frac{21^2}{30.06} + \frac{17^2}{16.86} + \frac{18^2}{18.14} + \frac{12^2}{21.19} + \frac{32^2}{22.8} - 137$

or, $\chi^2 = 13.38$

tab. (χ^2) is 3.99 at 5% with $(3-1)(2-1) = 2 \times 1 = 2$ d.f

Hence. cal. (χ^2) > tab. (χ^2)

comment: From the above calculation, we reject null hypothesis.

That means, there is a relationship between gender and flavor preference

CamScanner

(b) For given information in the following table, test at level of significance 0.05 that whether level of education affects the job performance. [$\chi^2_{0.05,4} = 13.3$]

Job Performance	Level of Education			Total
	Below primary	College	University	
Excellent	10	40	10	60
Good	30	30	20	80
Fair	10	30	20	60
Total	50	100	50	200

Solution:

Let, H_0 = Level of education doesn't affect the job performance

$H_1 = H_0$ is not true.

$E(10) = \frac{60\times50}{200} = 15$, $E(30) = \frac{80\times50}{200} = 20$, $E(10) = \frac{60\times50}{200} = 15$

$E(40) = \frac{60\times100}{200} = 30$, $E(30) = \frac{80\times100}{200} = 40$, $E(30) = \frac{60\times100}{200} = 30$

$E(10) = \frac{60\times50}{200} = 15$, $E(20) = \frac{80\times50}{200} = 20$, $E(20) = \frac{60\times50}{200} = 15$

We know,

$$\chi^2 = \left(\sum \frac{O^2}{E}\right) - n$$

or, $\chi^2 = \frac{10^2}{15} + \frac{30^2}{20} + \frac{10^2}{15} + \frac{40^2}{30} + \frac{30^2}{40} + \frac{30^2}{30} + \frac{10^2}{15} + \frac{20^2}{20} + \frac{20^2}{15} - 200$

or, $\chi^2 = 17.5$

tab. (χ^2) is 13.3 at 5% with $(3-1)(3-1) = 2\times2 = 4$ d.f.

Comment: Since cal. (χ^2) > tab. (χ^2). We reject H_0.

That means, level of of education affects the job performance.

CamScanner