UNIT - III

Applications of F - distribution

F - distribution has the following applicants in statistical theory.

F - test for Equality of two population variances:

Suppose we want to test.

i) Whether two independent samples x_i, (i=0,1,2... n_1) and y_i (i: 1,2,... n_2) have been drawn from the normal population with the same variance σ^2

ii) whether the two independent estimates of the population variance are homogeneous or not.

under the Null hypothesis (H_0) that i) $\sigma_x^2 = \sigma_y^2 = \sigma^2$. the population variances are equal or (ii) Two independent esitimates of the population variance are homogeneous the Statistic F is given by.

$$F = \frac{S_x^2}{S_y^2}$$

where $S_x^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} (x_i - \bar{x})^2$ and $S_y^2 = \frac{1}{n_2 - 1} \sum_{i=1}^{n_2} (y_i - \bar{y})^2$

are unbiased estimates of the common population variance σ^2 obtained from two independent samples and it follows Snedecor's F-distribution with (ν_1, ν_2) d.f. where $\nu_1 = n_1 - 1$ and $\nu_2 = n_2 - 1$.

Proof: $$F = \frac{s_x^2}{s_y^2} = \left[\frac{n_1}{n_1 - 1} s_x^2\right] \Big/ \left[\frac{n_2}{n_2 - 1} s_y^2\right]$$

$$= \left[\frac{n_1 s_x^2}{\sigma_x^2} \cdot \frac{1}{(n_1 - 1)}\right] \Big/ \left[\frac{n_2 s_y^2}{\sigma_y^2} \cdot \frac{1}{(n_2 - 1)}\right]$$

$(\because \sigma_x^2 = \sigma_y^2 = \sigma^2$, under $H_0)$.

Since $\frac{n_1 s_x^2}{\sigma_x^2}$ & $\frac{n_2 s_y^2}{\sigma_y^2}$ are independent chi-square variates with $(n_1 - 1)$ and $(n_2 - 1)$ d.f. respectively. F follows snedecor's F-distribution with $(n_1 - 1, n_2 - 1)$ d.f

F-test for testing the significance of an observed Multiple correlation coefficient:

If R is the observed multiple correlation coefficient of a variate with a K other variates in a random sample of the size n from a $(K+1)$ variate population, then Prof R. A. Fisher proved that

under the null hypothesis (H_0) that the Multiple correlation coefficient in the population is zero, the statistic.

$$F = \frac{R^2}{1-R^2} \cdot \frac{n-k-1}{k}$$

conforms to F - distribution with $(k, n-k-1)$ d.f.

F - test for testing the significance of an observed sample correlation ratio. η_{yx}.

Under the null hypothesis that population correlation ratio is zero, the test statistic is:

$$F = \frac{\eta^2}{1-\eta^2} \cdot \frac{N-h}{h-1} \sim F(h-1, N-h)$$

where N is the size of the sample (from a bi-variate normal population) arranged in h arrays.

F - test for testing the linearity of Regression:

For a sample of size N arranged in h arrays, from a bi-variate normal population, the test statistic for testing the hypothesis of linearity of regression is:

$$F = \frac{\eta^2 - r^2}{1-\eta^2} \cdot \frac{N-h}{h-2} \sim F(h-2, N-h).$$

Example 01:- In one sample of 8 observation, the sum of the squares of deviations of the sample values from the sample mean was 84.4 and in the other sample of 10 observation it was 102.6. Test whether this difference is significant at 5 percent level, given that the 5 percent point of F for $n_1 = 7$ and $n_2 = 9$ degrees of freedom is 3.29.

Solution: Here $n_1 = 8$, $n_2 = 10$ and $\Sigma(x-\bar{x})^2 = 84.4$,

$$\Sigma(x-\bar{y})^2 = 102.6$$

$$s_x^2 = \frac{1}{n_1-1}\Sigma(x-\bar{x})^2 = \frac{84.4}{7} = 12.057$$

$$s_y^2 = \frac{1}{n_2-1}\Sigma(y-\bar{y})^2 = \frac{102.6}{9} = 11.4$$

Under H_0: $\sigma_x^2 = \sigma_y^2 = \sigma^2$ the estimates of σ^2 given by the samples are homogeneous, the test statistic is:

$$F = \frac{s_x^2}{s_y^2} = \frac{12.057}{11.4} = 1.057.$$

Tabulated $F_{0.05}$ for (7, 9) d.f is 3.29.

Since calculated $F < F_{0.05}$, H_0 may be accepted at 5% level of significance.

Example 02 :- Two random samples gave the following results

Sample	Size	Sample mean	Sum of squares of deviation from the mean
1	10	15	90
2	12	14	108

Test whether the samples come from the same normal population at 5% level of significance.

[Given : $F_{0.05}(9, 11) = 2.90$, $F_{0.05}(11, 9) = 3.10$ (approx) and $t_{0.05}(20) = 2.086$, $t_{0.05}(22) = 2.07$]

Solution: A normal population has two parameter, mean μ and variance σ^2. To test if two independent samples have been drawn from the same normal population, we have to test i) the equality of population means. ii) the equality of population variances.

Null hypothesis : The two samples have been drawn from the same normal population.

$H_0 : \mu_1 = \mu_2$ and $\sigma_1^2 = \sigma_2^2$

$n_1 = 10$, $n_2 = 12$; $\bar{x}_1 = 15$, $\bar{x}_2 = 14$, $\Sigma(x_1 - \bar{x}_1)^2 = 90$,

$\Sigma(x_2 - \bar{x}_2)^2 = 108$.

F-test : Here

$$s_1^2 = \frac{1}{n_1-1}\sum(x_1-\bar{x}_1)^2 = \frac{90}{9} = 10,$$

$$s_2^2 = \frac{1}{n_2-1}\sum(x_2-\bar{x}_2)^2 = \frac{108}{11} = 9.82$$

Since $s_1^2 > s_2^2$, under $H_0 : \sigma_1^2 = \sigma_2^2$, the test statistic is

$$F = \frac{s_1^2}{s_2^2} \sim F(n_1-1, n_2-1) = F(9,11)$$

$$\therefore F = \frac{s_1^2}{s_2^2} = \frac{10}{9.82} = 1.018$$

Tabulated $F_{0.05}(9,11) = 2.90$. Since calculated F is less than tabulated F, it is not significant. Hence null Hypothesis of equality of population variances may be accepted

Since $\sigma_1^2 = \sigma_2^2$, we can now apply t test for testing $H_0 : \mu_1 = \mu_2$.

t-test : under $H_0' : \mu_1 = \mu_2$ against alternative hypothesis, $H_1' : \mu_1 \neq \mu_2$ the test statistic is :

$$t = \frac{\bar{x}_1 - \bar{x}_2}{\sqrt{s^2\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \sim t_{n_1+n_2-2} : t_{20}.$$

where, $s^2 = \frac{1}{n_1+n_2-2}\left[\sum(x_1-\bar{x}_1)^2 + \sum(x_2-\bar{x}_2)^2\right] = \frac{1}{20}(90+108)$
$= 9.9$

$s^2 = 9.9.$

$$t = \frac{15.14}{\sqrt{9.9\left(\frac{1}{10}+\frac{1}{12}\right)}} \cdot \frac{1}{\sqrt{9.9 \times \frac{11}{60}}} = \frac{1}{\sqrt{1.815}} = 0.743.$$

Tabulated $t_{0.05}$ for 20 d.f : 2.086. Since $|t| < t_{0.05}$, it is not significant. Hence the hypothesis H_0' : $\mu_1 = \mu_2$ may be accepted. Since both the hypothesis, H_0' : $\mu_1 = \mu_2$ and H_0 : $\sigma_1^2 = \sigma_2^2$ are accepted, we may regard that the given samples have been drawn from the same normal population.

Example 03: Pumpkins were grown under two experimental conditions. Two random samples of 11 and 9 pumpkins showed the sample standard deviations of their weight as 0.8 and 0.5 respectively. Assuming that the weight distribution are normal, test the hypothesis that the two variance are equal, against the alternative that they are not, at the 10% level.

[Assume that $P(F_{10,8} \geq 3.35) = 0.05$ and $P(F_{8,10} \geq 3.07) = 0.05$]

Solution: We want to test Null hypothesis H_0: $\sigma_x^2 = \sigma_y^2$

Alternative hypothesis, H_1 : $\sigma_x^2 \neq \sigma_y^2$ (Two-tailed)

We are given : $n_1 = 11$, $n_2 = 9$, $s_x = 0.8$ and $s_y = 0.5$

Under the null hypothesis, $H_0: \sigma_x^2 = \sigma_y^2$ the statistic

$$F = \frac{s_x^2}{s_y^2} \text{ follows } F \text{ distribution with } (n_1 - 1, n_2 - 1) \text{ d.f}$$

$$n_1 s_x^2 = (n_1 - 1) S_x^2 \Rightarrow S_x^2 = \left(\frac{n_1}{n_1 - 1}\right) s_x^2 = \left(\frac{11}{10}\right) \times (0.8)^2 = 0.704$$

Similarly, $S_y^2 = \left(\frac{n_2}{n_2 - 1}\right) s_y^2 = \left(\frac{9}{8}\right) \times (0.5)^2 = 0.28125$

$$F = \frac{0.704}{0.28125} = 2.5$$

The significant values of F for two-tailed test at level of significance $\alpha = 0.10$

unit - IV

Applications of chi-square Distribution :- χ^2- distribution has a large number of applications in statistics some of which are enumerical below :-

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

(ii) To test the "goodness of fit".

(iii) To test the independance of attributes.

(iv) To test the homogeneity of indipendence estimates of the population variance.

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

(vi) To test the homogenity of indipendent estimates of the population correlation coefficient.

Example 1 :- It is believed that the precision (as measured by the variance) of an instrument is no more than 0.16. write down the null and alternative hypothesis for testing this belief. carry out the test at 1% level given 11 measurements of the same subject on the instrument.

2.5 , 2.3, 2.4, 2.3, 2.5, 2.7, 2.5 , 2.6 , 2.6, 2.7, 2.5.

Sol.

COMPUTATION OF SAMPLE VARIANCE

x	$x - \bar{x}$	$(x-\bar{x})^2$
2.5	- 0.01	0.0001
2.3	- 0.21	0.0441
2.4	- 0.11	0.0121
2.3	- 0.21	0.0441
2.5	- 0.01	0.0001
2.7	+ 0.19	0.0361
2.5	- 0.01	0.0001
2.6	+ 0.09	0.0081
2.6	+ 0.09	0.0081
2.7	+ 0.19	0.0361
2.5	- 0.01	0.0001 .
$\bar{x} = \frac{27.6}{11} = 2.51$		$\Sigma (x-\bar{x})^2 = 0.1891.$

Null hypothesis, $H_0 : \sigma^2 = 0.16$. Alternative Hypothesis: $H_1 : \sigma^2 > 0.16$.

under the null hypothesis, $H_0 : \sigma^2 = 0.16$, the test statistic is :

$$\chi^2 = \frac{ns^2}{\sigma^2} = \frac{\Sigma (x-\bar{x})^2}{\sigma^2} = \frac{0.1891}{0.16} = 1.182 ,$$

which follows χ^2- distribution with d.f $n-1 = (11-1) = 10$.

Since the calculated value of χ^2 is less than the tabulated value 23.2 of χ^2 for 10 d.f at 1% level of significance, it is not significant. Here H_0 may be accepted and we conclude that the data are consistent with the hypothesis that the precision of the instrument is 0.16.

Example - 2:

Test the hypothesis that $\sigma = 10$, given that $s = 15$ for a random sample of size 50 from a normal population.

Sol

Null hypothesis, $H_0 : \sigma = 10$.

We are given $n = 50$, $s = 15$. Now $\chi^2 = \frac{ns^2}{\sigma^2} = \frac{50 \times 225}{100} = 112.5$.

Since n is large, using (15.4a), the test statistic

$$z = \sqrt{2\chi^2} - \sqrt{2n-1} \sim N(0,1).$$

$$\therefore z = \sqrt{225} - \sqrt{99}$$

$$= 15 - 9.95$$

$$= 5.05$$

Since $|z| > 3$, it is significant at all levels of significance and hence H_0 is rejected and we conclude that $\sigma \neq 10$.

Goodness of Fit Test:

A very powerful test for testing the significance of the discrepancy between theory and experiment was given prof. Karl. pearson in 1900 and is known as "chi-square test of goodness of fit". It enables us to find if the deviation of the experiment from theory is just by chance or is it really due to inadequacy of the theory of fit

If $f_i (i = 1, 2, .., n)$ is a set of observed (experimental) frequencies and $e_i (i = 1, 2, .., n)$ is corresponding set of expected (theoretical or hypothetical) frequencies, then karl pearson's chi-square, given by.

$$\chi^2 = \sum_{i=1}^{n} \left[\frac{(f_i - e_i)^2}{e_i}\right], \quad \left(\sum_{i=1}^{n} f_i = \sum_{i=1}^{n} e_i\right).$$

follows chi-square distribution with (n-1) d.f.

Example - 3 :

The fg figures show the distributions of digits in numbers chosen at random from a telephone directory :

Digits :	0	1	2	3	4	5	6	7	8	9	Total.
Frequency :	1026	1107	997	996	1075	933	1107	972	853		10,000.

Sol

Here we set up the null hypothesis that the digits occur equally frequently in the directory.

under the null hypothesis, the expected frequency for each of the digits $0, 1, 2, \ldots, 9$ is $10,000/10 = 1000$. The value of χ^2 is computed as follows :

CALCULATION FOR χ^2

Digits	Fruquency		$(f_i - e_i)^2$	$\frac{(f_i - e_i)^2}{e_i}$
	Observed (f_i)	Expected (e_i)		
0	1026	1000	676	0.676
1	1107	1000	11449	11.449
2	997	1000	9	0.009
3	966	1000	1156	1.156
4	1075	1000	5625	5.625
5	933	1000	4489	4.489
6	1107	1000	11449	11.449
7	972	1000	784	0.784
8	964	1000	1296	1.296
9	853	1000	21609	21.609
Total	10,000	10,000		58.542.

$$\chi^2 = \sum \frac{(f_i - e_i)^2}{e_i}$$

$$= 58.542.$$

The number of degrees of freedom

= Number of observations − Number of observations constraints.

Since the calculated value of χ^2 is much greater than the tabulated value, it is highly significant and we reject the null hypothesis. Thus we conclude that the digits are not uniformly distributed in the directory.

Test of independence of Attributes – contingency tables:

Let us consider two attributes A, B and A Bs. Such a classification in which attributes are divided into more than two classes is known as manifold classification. The various cell frequencies can be expressed in the fg table.

Also $\sum_{i=1}^{r} (A_i) = \sum_{j=1}^{s} (B_j) = N$, where N is the total frequency.

rXs CONTINGIENCY TABLE.

A / B	A_1	A_2	....	A_i	....	A_r	Total.
B_1	(A_1B_1)	(A_2B_1)	....	(A_iB_1)	....	(A_rB_1)	(B_1)
B_2	(A_1B_2)	(A_2B_2)	....	(A_iB_2)	....	(A_rB_2)	(B_2)
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
B_j	(A_1B_j)	(A_2B_j)	....	(A_iB_j)	....	(A_rB_j)	(B_j)
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
B_s	(A_1B_s)	(A_2B_s)	....	(A_iB_s)	....	(A_rB_s)	(B_s)
Total	(A_1)	(A_2)	....	(A_i)	....	(A_r)	N

UNIT - V

Non parametric Methods:-

Most of the statistical tests that we discussed so far had the following two features in common.

i) The form of the frequency function of the parent population from which the samples have been drawn is assumed to be known, and

ii) They were concerned with testing statistical hypothesis about the parameters of this frequency function or estimating its parameters.

For example, almost all the exact sample tests of significance are based on the fundamental assumption that the parent population is normal and are concerned with testing or estimating the means and variance of these population. Such tests, which deal with the parameters of the population are known as parametric tests. Thus the parametric statistical test is a test whose model specifies certain conditions about the parameter of the population from which the samples are drawn.

on the other hand, a Non-parametric (NP) test is a test does not depend on the particular form of this 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.

However certain assumptions associated with N.P. Tests are.

Sample observations are independent

The variable under study is continuous

P.d.f is continuous

Lower order moments exist.

Obviously these assumptions are fewer and much weaker than those associated with parametric tests.

Advantages and Disadvantages of N.P. Methods over parametric methods:-

Below we shall give briefly the comparative study of parametric and non-parametric methods and their relative merites and demerits.

Advantages of N.P. Methods:-

i) N.P. methods are readily comprehensible, very simple and easy to apply and do not require

complicated sample theory.

ii) No assumptions is made about the form of the frequency function of the parent population from which sampling is done.

iii) No parametric technique will apply to the data which are mere classification, while Np methods exist to deal with such data.

iv) Since the socio-economic data are not in general normally distributed. Np tests have found applications in psycometry, sociology and Educational statistics.

v) Np. 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, non parametric test can be applied of the scores are given in grades such as A, B ..etc.

Disadvantages of N.p. Tests:

(i) Np tests can be used only in the measurements are nominal methods. Even in that case, if a parametric test can exist it is more powerful than the parametric test. In other words. If all the assumptions of a statistical model are ~~eatle~~ satisfied the data and if the measurements are

of required strength than the N.P. test wasteful of time and data

ii) So far no N.P. methods exist for testing interactions in Analysing variance model unless special assumptions about the additivity of the model model.

iii) N.P. test are designed to test statistical hypothesis only and estimating the parameters.

Basic distribution :-

Let z be a continuous random variable with a p.d.f $f(\cdot)$. Let $z_1, z_2 \ldots z_n$ be a random sample of size n from $f(\cdot)$ and let $x_1, x_2, \ldots x_n$ be the corresponding ordered sample. Then the Joint density of $x_1, x_2, \ldots x_n$ is given by

$$g(x_1, x_2, \ldots x_n) = n!\ f(x_1)\ f(x_2) \ldots f(x_n), \quad -\infty < x_1 < x_2 < \ldots < x_n < \infty$$

$$U_i = \int_{-\infty}^{x_i} f(z)\, dz = F(x_i) \quad (i = 1, 2 \ldots n)$$

when $F(\cdot)$ is the distribution function of z. Since $F(x_i)$ is a uniform random variable on $[0, 1]$ U_i $(i = 1, 2, \ldots n)$ are random variables following uniform distribution on $[0, 1]$.

Thus the joint density $K(\cdot)$ of the random variables U_i $(i=1,2,\ldots,n)$

$$K(u_1, u_2 \ldots u_n) = n! \quad 0 \le u_1 \le u_2 < \ldots < u_k \le 1$$

and does not depend on $f(\cdot)$

$$E(U_i) = \int_0^1 \ldots \int_0^{u_1} \int_0^{u_2} u_1 u_1 \, du_1 du_2 \ldots du_n$$

$$= \frac{i}{n+1}$$

$$E(U_i) - E(U_{i+1}) = \frac{i}{n+1} - \frac{i-1}{n+1} = \frac{1}{n+1}$$

which is independent of $f(\cdot)$.

Wald-Wolfowitz Run Test:-

Definition:-

A run is defined as a sequence of letters of one kind surrounded by a sequence of letters of the other kind, and the number of elements in a run is usually referred to as the length(l) of the run.

We have in order a run of x ($l=2$), a run of Y ($l=3$) a run of x ($l=1$), a run of Y ($l=1$) etc..

Test for Randomness:-

Another application of the 'run' theory is in testing the randomness of a given set of observations. Let $x_1, x_2, \ldots x_n$ be the set of observation arranged in the order in which they occur. i.e x_i is the ith observations in the outcome of experiment. Then for each of the observations, and write A if the observation is above and B if it is below the median value. Thus we get a sequence of A's and B's of the type.

A B B A AA B A B B

under the null hypothesis H_0 that the set of observation is random. the number of run U in r.v. with

$$E(u) = \frac{n+2}{2} \text{ and}$$

$$\text{var}(u) = \frac{n}{4}\left(\frac{n-2}{n-1}\right)$$

Median Test:-

Median test is a statistical procedure for testing if two independent ordered sample differ in their central tendencies. In other words, it gives information if two independent samples are likely to have been drawn from the population with same median.

As in 'run' test. Let $x_1, x_2, \ldots x_{n_1}$ and $y_1, y_2, \ldots y_{n_2}$ be two independent ordered samples from the populations with pdf's $f_1(\cdot)$ and $f_2(\cdot)$ respectively The measurements must be at least ordinal. Let $z_1, z_2, \ldots z_{n_j} \cdot n_2$ be the combined ordered sample. Let m_1 be the number of x's and m_2 the number of y.

under $H_0: f_1(\cdot) = f_2(\cdot)$, the joint distribution of m_1 and m_2 is the hypergeometric distribution with probability function.

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

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

The distribution of m_1 under H_0 is also Hyper geometric

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

$$= \frac{n_1}{2} \cdot \frac{N-1}{N} \text{ if } N \text{ is odd}$$

$$\text{var}(m_1) = \frac{n_1 n_2}{4(N-1)} \text{ if } N \text{ is even}$$

$$= \frac{n_1 n_2 (N+1)}{4N^2} \text{ if } N \text{ is odd.}$$

$$Z = \frac{m_1 - E(m_1)}{\sqrt{\text{var}(m_1)}} \sim N(0,1) \text{ asymptotically.}$$

Sign Test :-

consider a situation where it is desired to compare two things or materials under various sets of conditions. An experiment is thus conducted under the following circumstances.

(i) where there are pairs of observations on two things being compared.

ii) For any given period, each of the two observation is made under similar extraneous conditions.

iii) Different pairs are observed under different conditions.

This implies that the differences $d_i = x_i - y_i$; $i = 1, 2, \ldots, n$ have different variances and thus renders the paired t-test invalid which would have otherwise been unless was obvious non-normality.

The only assumptions are;

(i) Measurements are such that the deviations $d_i = x_i - y_i$, can be expressed in terms of positive or negative signs.

ii) variables have continuous distribution

iii) d_i's are independent.

Procedure :-

Let (x_i, y_i) $i = 1, 2, \ldots n$ be n paired sample observations drawn from the two populations with pdf's $f_1(\cdot)$ and $f_2(\cdot)$.

Null Hypothesis $H_0 : f_1(\cdot) = f_2(\cdot)$

To test H_0 consider $d_i = x_i - y_i$ $(i = 1, 2, \ldots n)$

$$P[X - Y > 0] = 1/2 \text{ and } P(X - Y < 0) = 1/2$$

$$u_i = \begin{cases} 1, & \text{if } x_i - y_i > 0 \\ 0, & \text{if } x_i - y_i < 0 \end{cases}$$

u_i is Bernoulli variate with $p = P(x_i - y_i) > 0 = 1/2$

since u_i's $i = 1, 2, \ldots n$ are independent

$U = \sum_{i=1}^{n} u_i$, the total number of positive deviations. is a Binomial variate with parameters n and p $(= 1/2)$. the Number of positive deviations be k.

$$P(U \leq k) = \sum_{r=0}^{k} \binom{n}{r} p^r q^{n-r}, \quad (p = q = 1/2 \text{ under } H_0)$$

$$= (1/2)^n \sum_{r=0}^{k} \binom{n}{r} = p'$$

If $p' \leq 0.05$, we reject H_0 at 5% level of significance and if $p' > 0.05$ we conclude that the data do not provide any evidence against the null hypothesis.

For large samples $n \geq 30$

$$E(U) = np = n/2$$

$$\text{Var}(U) = npq = n/4$$

$$Z = \frac{U - E(U)}{\sqrt{\text{var}(U)}}$$

$$Z = \frac{U - n/2}{\sqrt{n/4}} \quad \text{is asymptotically } N(0,1)$$