UNIT-V Unit - V

Non-parametric test :- ① [redacted], [redacted]

The tests which do not depend upon the Population parameters such as mean and variance they are also called non-parametric test since these test do not depend on the shape of the distribution their called distribution free test.

/ some of the important non-parametric test are sign test, Rank test, one sample run test, median test, mann whitney 'U' test (one sample and two sample problems), kolmogorov's smirnov one sample test /

Advantages of non-parametric test :-

1. Distribution free that is do not require any assumption to be made about population following normal or any other distribution
2. simple and easy to understand and computed and sample size
3. Applicable to all types of data.
4. It is possible to what one with very small samples particular helpful to the resources collecting to pilot study data. or to the medical resources working with a rare disease
5. Make fewers less stringent assumption.

Limitations or Disadvantages of NP test.

1. If all the assumptions of the parametric test are infact that in the data, if the Measurement is required strength, then NP test are Wastefeell of data

2 There are no NP methods for testing interactions in the analysis of variance.

3. Tables of critical values may not be easily available.

THE SIGN TEST

The sign test is the simplest of the non-parametric tests. Its name comes from the fact that it is based on the direction (or signs for pluses or minuses) of a pair of observations and not on their numerical magnitude.

In any problem in which sign test is used, we count :

Number of + Signs

Number of – signs

Number of 0's (*i.e.*, which cannot be included either as positive or negative).

We take $H_0 : P = 0.5$ (Null hypo.hesis)

If the difference is due to chance effects the probability of a + sign for any particular pair is 1/2, as is the probability of a – sign. If S is the number of times the less frequent sign occurs, then S has the binomial distribution with $p = 1/2$.

The critical value for a two-sided alternative at $\alpha = 0.05$ can be conveniently found by the expression.

$$K = \frac{(n-1)}{2} - (0.98)\sqrt{n}$$

H_0 is rejected if $S \leq K$ for the sign test.

The sign test can be of two types :

(1) The one-sample sign test*

(2) The paired-sample sign test.

In a one-sample sign test we test the null hypothesis $\mu = \mu_0$ against an appropriate alternative on the basis of a random sample of size n, we replace each sample value greater than μ_0 with a plus sign and each sample value less than μ_0 with a minus sign and discard sample value exactly equal to (put O). We then test the null hypothesis that these plus and minus signs are values of a random variable having the binomial distribution with $p = 1/2$.

Mann-whitney U Test:- 4

This test helps us to determine whether 2 samples have come from the same population, The alternative hypothesis is that the means of the population are not equal the test of the null hypothesis is that the 2 samples come from identical populations may either be based on R_1 sum of ranks of the values of 1st sample or on R_2 sum of ranks or on R_2 sum of ranks of the values of the 2nd sample.

If the sample size are n_1 & n_2, sum of R_1 & R_2 is simpty the sum of n_1+n_2 +ve integers consider a new var defind by

$$U = n_1 + n_2 + \frac{n_1(n_1+1)}{2} - R_1$$

Test statistic, $z = \frac{U - E(U)}{SE(U)} \sim N(0,1)$

where, $E(U) = \frac{n_1 n_2}{2}$, $V(U) = \frac{n_1 n_2(n_1+n_2+1)}{12}$

If there are ties in the ascending order value the change only in the variance of U

$$V(U) = \frac{n_1 n_2 [n(n^2-1) - \Sigma T_i]}{12 n(n-1)}$$

where

$$T_i = t_i(t_i^2 - 1)$$

Nald wolfowitz Run test:-

Suppose $x_1, x_2, \dots x_n$ is an ordered sample from a population and $y_1, y_2, \cdots y_n$ be an independent ordered sample from another population. We want to test if the samples have been drawn from the same population.

Let us Combine the two samples and arrange in order of magnitude to give the combined ordered sample.

RUN:-

A run is defined as a sequence of letters of one kind surrounded by a sequence of letters of other kind and the number of element in a run is usually referred to as the length of the Run.

Arrange in ascending order then find the sequence of one observation of one variable bounded by one observation of othe variable test statistic is,

$$z = \frac{R - E(R)}{S.E(R)} \sim N(0,1)$$

where,

R - Total no. of Runs

$$E(R) = 1 + \frac{2n_1n_2}{n} \; ; \; V(R) = \frac{2n_1n_2(2n_1n_2 - n)}{n^2(n-1)}$$

Median test:-

Median test is a statistical procedure for testing if 2 independent ordered samples differ in their central tendencis In other words it gives information of 2 independent samples are likely to have median drawn from the population with the same median.

Let $x_1, x_2, \dots x_{n_1}$ and $y_1, y_2, \dots y_{n_2}$ be 2 independent ordered samples from the population with probability density function's for $f(y)$ respectively the measurement must be atleast ordinal.

Let $z_1, z_2, \dots z_{n_1+n_2}$ be the combinal ordered sample. Let m_1 be the No. of x's and m_2 be the no. of y's exceeding the median value M (say) of the combined sample.

The Null hypothesis H_0: The samples are drawn from the same population (or) from the different population with the same modian (i.e) $H_0 : f(x) = f(y)$.

The joint distribution of m_1 and m_2 is the following hyper-geometric distribution, with probability function is given by, 9 $$p(m_1; m_2) = \frac{\binom{n_1}{m_1}\binom{n_2}{m_2}}{\binom{n_1+n_2}{m_1+m_2}}$$

The distribution is most of the times quite in convenient to use. How ever, for large samples we may define m_1 to be assymptotically normal and use normal test.

$$z = \frac{m_1 - E(m_1)}{\sqrt{var(m_1)}} \sim N(0,1) \text{ assymptotically}$$

$$z = \frac{m_2 - E(m_2)}{\sqrt{var(m_2)}} \sim N(0,1) \text{ assymptotically.}$$

THE KRUSKAL-WALLIS OR H-TEST

If several independent samples are involved, analysis of variance is the usual procedure. Failure to meet the assumptions needed for analysis of variance makes its value doubtful. An alternative technique was developed called the Kruskal-Wallis one-way analysis of variance or the H-test. This test helps in testing the null hypothesis that k independent random samples come from identical populations against the alternative hypothesis that the means of these samples are not all equal.

As is done in the Mann-Whitney U-test all data are ranked as if they were in one sample, from lowest to highest, the rank sums of each sample are calculated. The H-statistic is calculated form the formula:

$$H = \frac{12}{N\,(N+1)}\left(\frac{{R_1}^2}{n_1} + \frac{{R_2}^2}{n_2} + \ldots + \frac{{R_k}^2}{n_k}\right) - 3\,(n+1)$$

9

when $n_1, n_2, \ldots, r_k$ are the number in each of k samples. $N = n_1 + n_2 + \ldots + n_k$, and $R_1, R_2, \ldots R_k$ are the rank sums of each sample. If there are ties, the usual procedure is followed, but H is fairly sensitive to ties, so if there are very many of them a correction should be made. The effect of the correction is to increase slightly the value of H, so its use is not imperative. For small samples, H is approximately distributed as Chi-Square with k-1 degrees of freedom, and table 6 given in the appendix can be used.

If the null hypothesis is true and each sample has at least five observations, the sampling distribution of H can be approximated closely with a chi-square distribution with k-1 degrees of freedom. Consequently, we can reject the null-hypothesis that $\mu = \mu_2 = \mu_k$ and accept the alternative that the μ's are not all equal at the level of significance α(alpha), if $H > \chi\alpha^2$ for k-1 degrees of freedom. If any sample has less than five items, the χ^2 approximation cannot be used, and the test must be based on special tables.

Illustration 6. A company's trainees are randomly assigned to groups which are taught a certain industrial inspection procedure by three different methods. At the end of the instructing period they are tested for inspection performance quality. The following are their scores

Method *A* : 80, 83, 79, 85, 90, 68
Method *B* : 82, 84, 60, 72, 86, 67, 91
Method *C* : 93, 65, 77, 78, 88.

Use the *H* test to determine at the 0.05 level of significance whether the three methods are equally effective.

Solution. Arranging the data jointly according to size and

$R_1 = 61$, $R_2 = 62$, $R_3 = 48$

Applying the formula for H :

$$H = \frac{12}{N(N+1)}\left[\frac{{R_1}^2}{n_1} + \frac{{R_2}^2}{n_2} + \frac{{R_3}^2}{n_3}\right] - 3(n+1)$$

$$= \frac{12}{18 \times 19}\left[\frac{(61)^2}{6} + \frac{(62)^2}{7} + \frac{(48)^2}{5}\right] - 3(19)$$

$$= \frac{12}{342}[620.17 + 549.14 + 460.8] - 57$$

$$= 57.197 - 57 = 0.197$$

$$v = 3 - 1 = 2$$

For $v = 2$, $\chi^2_{0.5} = 5.991$

The calculated value is less than the table value, the null hypothesis is separated and we conclude that the three months are equally effective.

be. This is the oldest nonparametric test still widely used. Another test, suitable for the above problem, is a test based on the empirical distribution function and is known as the Kolmogorov–Smirnov (KS) test.

Kolmogorov–Smirnov Test

11

Let $(X_1, X_2, \ldots, X_n)$ be a sample of n independent random observations on a RV X with distribution function F. The sample distribution function $S_n(x)$ is a consistent and unbiased estimator of $F(x)$. Let us assume that $F(x)$ is absolutely continuous. Then $S_n(x)$ is the UMVUE of $F(x)$ since it is a symmetric function of the sample observations. So the difference $S_n(x) - F(x)$ may be taken as the basis for testing any hypothesis about $F(x)$. Accordingly, we define

$$D_n = \sup_x |S_n(x) - F(x)|$$

which is called the KS test statistic (two-sided). If we define

$$D_n^+ = \sup_x \{S_n(x) - F(x)\} \tag{12.5}$$

and

$$D_n^- = \sup_x \{F(x) - S_n(x)\} \tag{12.6}$$

then

$$D_n = \max \{D_n^+, D_n^-\} \tag{12.7}$$