Unit IV

Non - parametric test

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

Some of the important non- parametric test are sign test, rank test, one sample run test, median test, mann whitney 'U' test (1 sample and 2 sample problems) kolmogorov's smirnov one sample test.

Advantages of NP test

1) Distn free that is do not require any assumption to be made about popln following normal or any other distn

2) simple and easy to understand and computed sample size

3) Applicable to all types of data

4) It is possible to whatone with very small samples particular shelpful to the resources collecting to pilot study data or to the medical resources working with a rare disease

5) Make fewer less stringent assumption

Disadvantages of NP test

1) If all the assumptions of the parametric test are infact that in the data, if the measurement is of the required strength the NP-test are wasteful of data.

2) There are no non-parametric ~~m~~methods testing interactions in the ANOVA.

3) Tables of critical values may not be easily available

Run test

Null hypothesis H_0: The samples have been drawn from the same popln. $f_1(.) = f_2(.)$

H_1: $f_1(.) \neq f_2(.)$

Definition:

Run: (i) The run is defined as sequence of letters of same kind by the sequence of letters of other kind.

(ii) The no. of elements in the run is called a length of the run 'l'

Now let the combined sample be ordered

$$x_1\ x_2\ y_1\ y_2\ y_3\ x_3\ y_4\ x_4\ x_5\ x_6 \ldots$$

Test statistics:-

To test the null hypothesis we have the statistics 'U', where U is the no. of ~~rows~~ runs

The test statistics $z = \dfrac{U - E(U)}{\sqrt{V(U)}} \sim N(0,1)$

where u is the no. of runs

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

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

n_1 = 1st sample size n_2 = 2nd sample size

Inference:

If $Z_{cal} \leq Z_{exp}$, then accept the null hypothesis H_0 otherwise reject H_0.

Median test:-

H_0: The two samples have been from the popln with same median $H_0: f_1(\cdot) = f_2(\cdot)$

H_1: There is a significant difference between the samples. $f_1(\cdot) \neq f_2(\cdot)$.

WKT if m is the median to test the null hypothesis H_0.

$$z = \frac{m - E(m)}{\sqrt{V(m)}}$$

where $E(m) = \dfrac{n_1}{2}$ if $N = n_1 + n_2$ is even

$= \dfrac{n_1}{2}\left(\dfrac{N-1}{N}\right)$ if N is odd.

$$V(m) = \frac{n_1 n_2}{4(N-1)} \quad ; \text{if } N \text{ is even}$$

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

where n_1 = 1st sample size n_2 = 2nd sample size

Inference:

If $Z_{cal} \le Z_{exp}$, then accept null hypothesis. Otherwise reject the H_0.

Sign test

H_0: The samples are taken from the same population $f_1(\cdot) = f_2(\cdot)$

H_1: The samples are significantly different $f_1(\cdot) \neq f_2(\cdot)$

in other words $H_0 : P[(X-Y) > 0] = 1/2$ similarly $H_1 : P[(X-Y) < 0] = 1/2$

Derivation of test statistics:-

Let (x_i, y_i) $i = 1, 2 \ldots n$ be the paired observations, x_i represent the first group, y_i represent the second group.

Let $d_i = x_i - y_i$ represent +ve (or) -ve value know defined

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

$$V = \Sigma u_i \sim B(n, p = 1/2)$$

Let 'V' be number of positive sign

$E(U) = np \qquad V(U) = npq$

when $p = 1/2$, $q = 1/2$ $\quad E(U) = \frac{n}{2} \quad V(U) = \frac{n}{4}$

The test statistic is $\quad z = \frac{U - E(U)}{\sqrt{V(U)}} \sim N(0,1)$

$$z = \frac{U - \frac{n}{2}}{\sqrt{\frac{n}{4}}} \sim N(0,1)$$

Inference:

If $z_{cal} \leq z_{exp}$, we accept the null hypothesis at 5% level of significance otherwise reject H_0.

Mann-Whitney wilcoxen U test:-

Null hypothesis H_0 : There is no significant difference between the two groups $f_1(\cdot) = f_2(\cdot)$

Alternative hypothesis H_1: $f_1(\cdot) \neq f_2(\cdot)$ The 2 groupsare not significantly difference.

Derivation of test statestics:-

Let (x_i, y_i) $i = 1, 2 \ldots n$ be the ordered samples of size n_1 and n_2 with $f_1(\cdot)$ and $f_2(\cdot)$ respectively.

Now find the combined ordered statestics with corresponding ranks.

Let $T =$ sum of ranks of the variable y in the combined order.

By using the first and second sample size. Let the statestic U we define as

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

W.K.T $\quad E(U) = \frac{n_1 n_2}{2} \qquad V(U) = \frac{n_1 n_2 (n_1+n_2-1)}{n12}$

The test statistic is $z = \frac{U - E(U)}{\sqrt{V(U)}} \sim N(0,1)$

$$Z_{cal} = \frac{\left(n_1 n_2 + \frac{n_2(n_2+1)}{2} - T\right) - \left(\frac{n_1 n_2}{2}\right)}{\sqrt{\frac{n_1 n_2 (n_1+n_2+1)}{12}}} \sim N(0,1)$$

Inference:

If $Z_{cal} \leq Z_{exp}$ we accept the null hypothesis H_0 at 5% level of significance otherwise reject H_0.

The Kolmogorov - smirnov one - sample statistic

A random sample $x_1, x_2 \ldots x_n$ is drawn from a popln with unknown cumulative distn fun $F_X(x)$. For any value of x, the empirical dbn fn of the sample. $S_n(x)$, provides a consistent point estimate for $\sqrt{x}(x)$. The values of the order statistics $x_{(1)}, x_{(2)} \ldots x_{(n)}$. For the sample approaches the true distn fn for all x. Therefore, for large n

Scanned with CamScanner

Comparison made b/w the empirical distn fn of the 2 samples.

For two r.s of size m and n from continuous popln F_x and F_y, their order statistics are

$x_{(1)}, x_{(2)} \ldots x_{(m)}$ and $y_{(1)}, y_{(2)} \ldots y_{(n)}$

their respective empirical distn fn's denoted by $S_m(x)$ & $S_n(x)$ are defined as

$$S_m(x) = \begin{cases} 0 & \text{if } x < x_{(1)} \\ k/m & \text{if } x_{(k)} \leq x \leq x_{(k+1)} \text{ for } k=1,2..m \\ 1 & \text{if } x > x_{(m)} \end{cases}$$

$$S_n(x) = \begin{cases} 0 & \text{if } x < y_{(1)} \\ k/n & \text{if } y_{(k)} < x < y_{(k+1)} \text{ for } k=1,2..m-1 \\ 1 & \text{if } x > y_{(n)} \end{cases}$$

The empirical distn fn for the x and y sample should be reasonable estimate of their respective popln distn

if the null hypothesis $H_0 : F_Y(x) = F_X(x) \ \forall x$ is true, the popln distn are identical and we have 2 samples from the same popln. The two-sided K-S two sample test criterion, denoted by $D_{m,n}$ is the maximum absolute difference b/w the two empirical distn is

$$D_{mn} = \max_x |S_m(x) - S_n(x)|$$

Since here only the magnitudes, and not the direction of the deviations are considered, $D_{m,n}$ is appropriate for a general two sided alternative. $H_1 : F_Y(x) \neq F_X(x)$ for some x

The test statistic is consistent here with the rejection region defined by $D_{m,n} > C_\alpha$

Define a run in a sequence of symbols

A run is a sequence of symbols followed and preceded by other type of symbols or no symbols.

For eg:- a sequence FMMF ~~F~~ I MMMFF of symbols F & M has 5 runs.

Test for Randomness (single sample)

Another application of the 'run' given set of observation.

H_0: The sample is drawn randomly H_1: The sample is bias.

Step 1:-

Let $x_1, x_2 \ldots x_n$ be a set of observation arranged in which they occur x_i is the ith observation in the outcome of an experiment.

step 2:- Let M be the median

step 3:- We see if it is above or below the median of the the observation 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 AB B A AA BA BB — ①

Step4:- Under the H_0 that the set of observation is random, the no. of runs u in ① is r.v with

$$U = \text{no. of rows} \qquad E(U) = \frac{n+2}{2} \qquad V(U) = \frac{n}{4}\left(\frac{n-2}{n-1}\right)$$

step 5:- For large n (>25) U may be regard as asymptotic normal and we may use the normal test.

$$Z_{cal} = \frac{U - E(U)}{\sqrt{V(U)}} \sim N(0,1)$$

If $Z_{cal} \leq Z_{exp}$ value, we accept H_0 otherwise reject H_0.

wilcoxon's signed rank test

Ordinary sign test was based only on the direction of difference ignoring their magnitudes. But wilcoxon's signed rank test takes into consideration the both. This test is more sensitive and powerful than ordinary sign test.

To perform the test for $H_0: M = M_0$ Vs $H_1: M \neq M_0$. Find the difference $d_i = x_{(i)} - M_0$ for $i = 1, 2 \ldots n$ d_i will be distributed symmetrically about the median zero so that +ve and -ve difference of equal absolute magnitude have equal prob. of occurences. The steps of the test are as follows:-

step 1:-

Arrange the difference in ascending order ignoring the sign and rank them from 1 to n

step 2:-

now assign the signs to the ranks which the original difference passess

step 3:-

suppose the sum of ranks of +ve d_i's then

$$T^+ = \sum_{i=1}^{n} Z_{(i)}$$

$Z_{(i)}$ are independent Bernoulli variables but are not identically distributed $Z_{(i)}$ has mean P_i and variance $P_i q_i$ and $cov(Z_{(i)}, Z_{(j)}) = 0$ for $i \neq j$ T^+ has mean

$\sum_{i=1}^{n} i P_i$ and variance $\sum_{i=1}^{n} i P_i (1 - P_i)$ under H_0.

$P_i = 1/2$ and hence $E(T^+) = \frac{1}{2} \sum_{i=1}^{n} i = \frac{n(n+1)}{4}$

and $Var(T^+) = \frac{1}{4} \sum_{i=1}^{n} i^2 = \frac{n(n+1)(2n+1)}{24}$

If T^- is smaller, some treatment can be given.

Let $T = \min(T^+, T^-)$ if T_α is a number such that $P(T < T_\alpha) = \alpha$.

The test criterion for testing $H_0 : M = M_0$ Vs $H_1 : M \neq M_0$ is, find the critical value of T from the table for the sample size n and prefixed level of significance α.

If $T^+ < T_\alpha$, reject H_0, otherwise accept H_0. If the alternative hypothesis leads to one tailed test, the critical value of T from the table

Kruskal Wallis test.

Kruskal wallis' test is one of the most frequently used method in nonparametric statistics for analysing data in one way classification it is equivalent to 1 way ANOVA in parametric methods.

We test the identi of K popln (in respect of medians) from which the independent samples have been drawn. There is no restrictions on sample sizes.

Assumptions:-

The observations are independent within and between samples. The variable under study is continuous.

The popln are identical except possibly in respect of median.

H_0 : All the popln are identical

H_1 : At least one pair of poplns do not have the same median

Let there are k ind. samples from k popln of sizes $n_1, n_2 \cdots n_k$

The observations in k sample can always be presented in the tabular form as given below.

Sample Numbers

1	2	$\cdots$	i	$\cdots$	K
x_{11}	x_{21}	$\cdots$	x_{i1}	$\cdots$	x_{K1}
x_{12}	x_{22}	$\cdots$	x_{i2}	$\cdots$	x_{K2}
$\vdots$	$\vdots$		$\vdots$		$\vdots$
x_{1n_1}	x_{1n_2}		x_{in_i}		x_{Kn_K}

Assign rank to each observation from 1 to $N = \sum_{i=1}^{n} x_i$ by pooling all the sample observation and writing them in assending order. The sum of rank is obviously equal to $\frac{N(N+1)}{2}$ under H_0, the sum of the ranks would be divided in proportion to sample size amoung K

Samples, for the ith sample of size n_i, the expected sum of rank is

$$\frac{n_i}{N}\frac{N(N+1)}{2} = \frac{n_i(N+1)}{2}$$

Suppose R_i is the actual sum of ranks of observations in sample i. To test H_0, KW test statistic is a weighted sum of squares of deviations of the sum of ranks of treatments from the expected sum of ranks using reciprocals of sample size as the weights. The Kruskal Wallies statistic in notational form is,

$$H = \frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{1}{n_i}\left[R_i - \frac{n_i(N+1)}{2}\right]^2$$

$$= \frac{12}{N(N+1)}\sum_{i=1}^{k}\frac{R_i^2}{n_i} - 3(N+1)$$

The statistic it is approximately distributed as χ^2 with $(K-1)$ df. Subject to the condition that n_i should be large, (ie) each n_i should not be less than 5. The decision about H_0 can be taken in the usual manner.