STUDY MATERIALS

BY KUNAL DAS

class: BSc 6th Semester (Major)

Paper: 6.1(M)

NON-PARAMETRIC TEST

A non-parametric (N.P.) test is a test that doesnot depend on the particular form the basic distribution (eg: Normal) from which the samples are drawn. In many cases an experimenter doesnot know the form of basic frequency function and needs statistical test/techniques which are applicable regardless of the form of the density function. Another term which is often used interchangeably with non-parametric is distribution free. In such procedures assumptions regarding the population arenot necessary. The term "Non-Parametric" is due to J. Wolfowitz who used it to indicate that the population couldnot be specified by finite number of parameters. A non-parametric test is concerned with the form of the population and not with any parametric values. The term "distribution free" is used to denote the distribution of the test statistic and the term "non-parametric" is used to denote the type of hypothesis problem to be investigated.

According to Gibbons, a statistical technique is said to be non-parametric if it satisfies one of the following five criteria.

(a) The data are count data of number of observations in each catagory.

(b) The data are nominal in scale.

(c) The data are ordinal in scale.

(d) The inference doesnot concern a parameter.

(e) The assumptions are general rather than specific.

Assumptions of Non-Parametric Test:

(a) Sample observations are independent.
(b) The variable under study is continuous.
(c) The probability density function of a random variable is continuous.
(d) Lower order moment exist.

Advantages and drawbacks of Non-Parametric tests over parametric tests: (See Fundamentals of MS by Gupta & Kapoor)

Advantages of N.P. test

(a) N.P. tests are readily comprehensible, very simple and easy to apply and don't require complicated sample theory.

(b) No assumption is made about the form of the frequency distribution of the parent population from which the sample is taken.

(c) N.P. technique will apply to the data which are mere classification (i.e which are measured in nominal scale), while NP test exist to deal with such data.

(d) Since the socio-economic data are not, in general, normally distributed, NP tests have found applications in Psychometry, Sociology and Educational Statistics.

(e) NP tests are available to deal with the data which are given in ranks or whose seemingly numerical scores have the strength or ranks. eg. NP test can be applied if the scores are given in grades such as A+, A-, B, A, B+ etc.

Drawbacks of NP test.

(a) NP tests can be used only if the measurements are nominal or ordinal. Even in that case, if a parametric test exists it is more powerful than NP test.

2. So far, no NP test exists for testing interactions in "analysis of variance" models unless special assumptions about the additivity of the model are made.
3. NP tests are designed to test statistical hypothesis only and not for estimating the parameters.

Single Sample Problem:

Problem of Location: Let $X_1, X_2, \ldots, X_n$ be a sample of size n from some unknown distribution function (DF) $F_X(x)$. We assume that $F_X(x)$ is absolutely continuous. Let p be a positive real number, $0<p<1$. Here an appropriate measure of location is median or the pth quartile i.e. M_p (say) the pth quartile of $F_X(x)$. One may wish to know whether the given M_0 is median of the distribution, or whether $f_X(x)$ is symmetric probability distribution. To test $H_0: M_p = M_0$ we first consider sign test.

SIGN TEST/ORDINARY SIGN TEST/SIGN TEST FOR LOCATION OF UNIVARIATE POPULATION

A random sample of N observations $X_1, X_2, \ldots, X_N$ is drawn from a population with unknown median M. Here $F_X(x)$ is assumed to be continuous around M. In other words, the assumptions are independent observations and $P_r(X=M)=0$ or $P_r(X-M=0)=0$. The hypothesis to be tested concerns the value of the population median

$$H_0: M = M_0$$

against $H_1: M \neq M_0$ (two-tailed)

It's corresponding alternative hypothesis can be one sided or two sided on the value of M.

For any distribution which satisfies $P_r[X=M]=0$, by the definition of M, we have

$$P_r[X>M] = P_r[X<M] = 1/2$$

Since the hypothesis here states that M_0 is the value of X which divides the area under the frequency distribution into two equal parts and equivalent symbolic representation of H_0 is

$$H_0: \theta = P_r[X>M_0] = P_r[X<M_0]$$

If the sample data are consistent with the hypothesized median value, on the average of half of the sample observation will lie above M_0 and half below. Thus the number of observations above M_0 which will be denoted by "K" can be used to test the validity of the null hypothesis. The observation will constitute a set of N independent random variables from the Bernoulli population with parameter $\theta = P_r[X>M_0]$ regardless of the population $f_X(x)$. The sampling of the random variable K then is the Binomial probability defn with parameter θ. The value of $\theta = 1/2 = 0.5$ of the null hypothesis is true. Since K is actually the number of positive signs (+) among the N differences $X_i - M_0$; $i=1,2,\ldots,N$ the non-parametric test based on K is called sign test.

To test sign test, theoritically no problem of zero differences exists. The population was assumed to be continuous around median. But in reality zero difference can occure the usual procedure follows here is simply to ignore zero differences and reduce N.

Test Procedure:

So the null hypothesis becomes equivalent to testing

$H_0: p = 1/2$ since $r \sim B(n, \theta = 1/2)$

Against $H_1: p \neq 1/2$ (whatever the case may be)

or $H_1: p < 1/2$

The test criterion is reject H_0 if $r \geq r_{\alpha/2}$ where $r_{\alpha/2}$ is the critical value at significance α. $r_{\alpha/2}$ is the smallest integer which satisfies the conditions.

$$\sum_{r=r_{\alpha/2}}^{N} {}^{N}C_r \left(\tfrac{1}{2}\right)^r \left(\tfrac{1}{2}\right)^{N-r} \leq \alpha/2$$

or $r \leq r_{\alpha/2}$ is the smallest integer such that

$$\sum_{r=0}^{r_{\alpha/2}} \left({}^{N}C_r\right) \left(\tfrac{1}{2}\right)^r \left(\tfrac{1}{2}\right)^{N-r} \leq \alpha/2$$

(for small samples)

For large sample: If $n \geq 25$, the normal test can be used to decide H_0.

The Z statistic is given by

$$Z = \frac{(r+0.5) - n \cdot \frac{1}{2}}{\sqrt{n \cdot \frac{1}{2} \cdot \frac{1}{2}}} = \frac{(r+0.5) - n/2}{\sqrt{n}/2} \quad \text{where } r < n/2$$

$$= \frac{(r-0.5) - n/2}{\frac{\sqrt{n}}{2}} \quad \text{, where } r > n/2$$

PAIRED SIGN TEST / SIGN TEST FOR LOCATION OF BIVARIATE POPULATION.

Let $X_1, X_2, \ldots, X_n$ and $Y_1, Y_2, \ldots, Y_n$ be two random samples of same size "n" drawn from two populations with pdf $f_1(x)$ and $f_2(y)$. We want to test the null hypothesis,

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

The observations are arranged in pairs i.e. (x_i, y_i); $i = 1, 2, \ldots, n$ and it is assumed that

each pair is observed under identical conditions. Now $d_i = (x_i - y_i)$ is measured and only the sign (+ or −) is noted in lieu of the actual deviations. Now, under the null hypothesis the probability that the first sample exceeds the first observations of the second sample is equal to the probability that the first observation of the second sample exceeds the first observation of the first sample, and the probability of a tie is zero. So it can be written as —

$$H_0: \Pr[X-Y>0] = \tfrac{1}{2} \text{ and } \Pr[X-Y<0] = \tfrac{1}{2}$$

Let us define

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

So, U_i is a bernoulli variate with $p = \Pr(x_i - y_i > 0) = \tfrac{1}{2}$. Since U_i's $i = 1, 2, \ldots, n$ are independent, $U = \sum_i^n U_i$, the total number of positive deviations, is a binomial variate with parameters n and $p = \tfrac{1}{2}$ (under H_0). Let K be the no. of positive deviations, so

$$\Pr[U \le K] = \sum_{r=0}^{K} {}^nC_r\, p^r q^{n-r} = \left(\tfrac{1}{2}\right)^n \sum_{r=0}^{K} {}^nC_r = p' \text{ (say)}$$

$\because p = \tfrac{1}{2} = q$

If $p' \le 0.05$, then we reject H_0 at 5% level of significance, If $p' > 0.05$ then we conclude that the data doesnot go against H_0 and so H_0 is accepted.

For large samples ($n \geq 30$), we may regard U to be asymptotically normal with $E[U] = np = n/2$ & $V[U] = npq = n/4$

$$\therefore \quad Z = \frac{U - E[U]}{\sqrt{Var(U)}} = \frac{U - n/2}{\frac{\sqrt{n}}{2}} \sim N(0,1)$$

and the decision will be taken on the basis of normality test for null hypothesis H_0.

—x—

WILCOXON SIGNED RANK TEST / WILCOXON SIGNED RANKSUM TEST

The ordinary sign test utilizes only the signs of difference between each observation and the hypothesized median M_0. Here the magnitude of these observations relative to M_0 are ignored. Assuming that such information is available, a test statistic which takes into account this individual relative magnitude might be expected to give better performance. If we are willing to make assumption that the parent population is symmetric, the Wilcoxon signed rank test provides an alternative test of location which is affected by both the magnitude and signs of these differences.

Test procedure:

Let the sample $x_1, x_2, \ldots, x_n$ be drawn from a population with CDF $F_X(x)$. For testing $H_0: F(m) = \frac{1}{2}$

against $H_1: F(m) \neq \frac{1}{2}$

or $H_1: F(m) > \frac{1}{2}$

or $H_1: F(m) < \frac{1}{2}$

Case 1: For sample size $n > 25$.

Let us suppose that $m = 0$. So the distn is symmetric about the origin. We have $F(-x) = 1 - F(x)$ and $f(-x) = f(x)$

Then test statistic $W \sim N\left(0, \frac{n(n+1)(2n+1)}{6}\right)$ for testing $H_0: m = 0$ for which $T = \frac{W}{\sqrt{n(n+1)(2n+1)/6}} \sim N(0,1)$

Case 2: For sample size $n > 25$

For testing $H_0: m = m_0$ ie $H_0: F(m = m_0) = \frac{1}{2}$

we have $W \sim N\left(\frac{n(n+1)}{4}, \frac{n(n+1)(2n+1)}{24}\right)$

for which $T = \frac{W - n(n+1)/4}{\sqrt{\frac{n(n+1)(2n+1)}{6}}} \sim N(0,1)$

Case III Sample size $n<25$

For testing $H_0: m = m_0$ we first find out the differences $d_i = x_i - \theta_0$ under H_0. We can assume that the values of d_i are independent and come from a population symmetrical about zero. We then find $|d_i|$, the absolute difference. These absolute differences are then arranged in ascending order and are accordingly ranked. The absolute difference with zero values are ignored. Let the number of observation be now $n_1 < n$ after the tied ranks are eliminated. In case of a tie in rank, each of the tied values are given the average value of the rank. Let T^+ be the sum of ranks for positive d_i's and T^- be the sum of ranks for negative d_i's, then $T^+ + T^- = \frac{n_1(n_1+1)}{2}$

The null distribution of T^+ and T^- are identical and symmetrical about the value $\frac{n_1(n_1+1)}{2}$. The smaller (minimum) of the two values T^+ and T^- is compared with the table value (critical value for T in the Wilcoxon signed rank test) for a given level of significance for n_1 observations and accordingly about the null hypothesis is taken.

of Alt. Hyp. $H_1: m > m_0$ Then we reject H_0 if $T^- \leq T_\alpha$

$H_1: m < m_0$ Then we reject H_0 if $T^+ \leq T_\alpha$

$H_1: m \neq m_0$ Then we reject H_0 if $T^- \leq T_\alpha$ or $T^+ \leq T_\alpha$

—x—

Example 1: A drug was injected to a fresh group of 10 rats every day. The scientist -in-charge of the experiment made a claim that not more than 3 rats showed an increase in blood pressure on an average. The following increase in blood pressure was noticed in the following number of rats in the last 10 days after the drug was administered:

2, 4, 5, 1, 6, 3, 2, 1, 7 and 8.

Solution : Here the null hypothesis is $H_0 : m = 3$ tested against $H_1 : m > 3$. To perform the calculations we construct the following table:

x_i	$d_i = x_i - 3$	*Signs*
2	–1	–
4	1	+
5	2	+
1	-2	-
6	3	+
3	0	ignored
2	-1	-
1	-2	-
7	4	+
8	5	+

Thus, $n = 9$, and the number of plus signs $= x = 5$. Under the null hypothesis $X \sim B\left(9, \frac{1}{2}\right)$.

$$\text{So, } P(X \geq 5) = 1 - P(X < 5) = 1 - \sum_{x=0}^{4} {}^9C_x \left(\frac{1}{2}\right)^x \left(\frac{1}{2}\right)^{9-x}$$

$$= 1 - [\left(\frac{1}{2}\right)^9 + 9 \times \left(\frac{1}{2}\right)^9 + 36 \times \left(\frac{1}{2}\right)^9 + 84 \times \left(\frac{1}{2}\right)^9 + 126 \times \left(\frac{1}{2}\right)^9]$$

$$= 1 - \left(\frac{1}{2}\right)^9 (1 + 9 + 36 + 84 + 126) = \frac{1}{2} = 0.5$$

Thus, $P(X \geq 5) > 0.05$ (level of significance) α. So, H_0 is accepted and thus the claim made by the scientist is true.

Example 2: Suppose that we want to test the hypothesis that the median body length (θ) of frogs of a particular variety is $\theta_0 = 6.9$ cms against the alternative hypothesis $\theta_0 \neq 6.9$ cms with $\alpha = 0.05$ on the basis of the following measurements.

6.3, 5.8, 7.7, 8.5, 5.2, 6.7, 7.3, 5.6, 8.3, 7.7, 8.2, 6.0, 6.8, 6.9, 7.3, 7.0, 7.1, 6.6, 7.4

Solution : We set up the following null hypothesis $H_0 : \theta = 6.9$ to be tested against the alternative hypothesis $H_1 : \theta \neq 6.9$.

Let us put '+' for values in the series greater than 6.9 for values in the series less than 6.9 and 0 for values in the series equal to 6.9 Thus we get

$-, -, +, +, -, -, +, -, +, +, +, -, -, 0, -, +, +, +, -, +$

Thus, number of positive sign = 10, number of negative sign = 9 and so $n = 19$.

x_i	$d_i = x_i - 6.9$	$\lvert d_i \rvert$	R_i
6.3	–0.6	0.6	9.5
5.8	–1.1	1.1	14
7.7	0.8	0.8	11.5
8.5	1.7	1.7	18
5.2	–1.7	1.7	19
6.7	–0.2	2	3.5
7.3	0.4	0.4	6.5
5.6	–1.3	1.3	15.5
8.3	1.4	1.4	17
7.7	0.8	0.8	11.5
8.2	1.3	1.3	15.5
6.0	–0.9	0.9	13
6.8	–0.1	0.1	1.5
6.3	–0.6	0.6	9.4
6.9	0	0	ignored
7.3	0.4	0.4	6.5
7.0	0.1	0.1	1.5
7.1	0.2	0.2	3.5
6.6	–0.3	0.3	5
7.4	0.5	0.5	8

Thus, the sum of ranks with positive values of d_i's is $T^+ = 99.5$. Similarly, the sum of ranks with negative values of d_i's $T^- = 90.5$. Thus, $T = \min(T^+, T^-) = \min(99.5, 90.5) = 90.5$.

Here, since one of the observation is ignored so we have $n = 19$ and thus for a two sided alternative we find the table value of $T_\alpha = 46$ for $n = 19$ for $\alpha = 0.05$ level of significance. Since $T_n < T$ so we accept the null hypothesis and conclude that $H_0 : \theta = 6.9$ is true.

Example 3: A medical representative visited 12 doctors in a town. In order to meet the doctor he had to wait for 25, 10, 15, 20, 17, 11, 30, 27, 36, 40, 5 and 26 minutes respectively. However, the senior sales representative earlier claimed that the doctor kept him waiting for more than 20 minutes on an average. Using Wilcoxon Signed rank test verify the claim made by the senior sales representative at 5% level of significance.

Solution : Here the null hypothesis is

$H_0 : \mu = 20$ minutes

tested against the alternative hypothesis

$H_1 : \mu > 20$ minutes.

To calculate the test statistic we construct the following table:

x_i	$x_i - 20$	$d_i = \lvert x_i - 20 \rvert$	*Ranks*
25	5	5	2.5
10	10	10	7.5
15	-5	5	2.5
20	0	0	ignored
17	-3	3	1
11	-9	9	6
30	10	10	7.5
27	7	7	5
36	16	16	10
40	20	20	11
5	-15	15	9
26	6	6	4

Here, the sum of positive ranks is equal to $T^+ = 40$ and the sum of negative ranks is equal to $T^- = 26$. Also, $T = \min(T^+, T^-) = 26$. The effective sample size is $n = 11$.

Since the alternative sample size is of the form $H_1 : m > m_0$ so the test statistic is T^-. Here $T^- = 26 > T_{0.05} = 14$. Thus H_0 is accepted and so it can be concluded that the average waiting time of the sales representative for the doctor is 20 minutes.

****