Analysis of Variance and Design of Experiments-II

MODULE IX

LECTURE - 37

NONPARAMETRIC ANALYSIS OF VARIANCE

Dr. Shalabh
Department of Mathematics & Statistics
Indian Institute of Technology Kanpur

The usual methods of analysis of variance depend on the basic assumption that all the observations are originating from a normal distribution with mean μ and variance σ^2. When the observations do not follow the normal distribution or even if it is difficult to ascertain the form of the distribution, then the parametric procedures like analysis of variance based on the normal distribution cannot be used as such. In such cases, the nonparametric procedures are helpful. The nonparametric statistical procedures are concerned with the development of statistical tools without making any assumption about the form of the probability distribution of the sample observations. So now our aim is to develop the techniques of analysis of variance without making any assumption about the distribution of sample observations like the normal distribution.

The nonparametric procedures are not generally as efficient as their counterpart parametric procedures. The small sample properties of nonparametric procedures are also difficult to carry out due to complicated computational procedures. So the relative performance of the nonparametric methods is generally investigated in the large samples. We consider here only the development of analysis of variance in nonparametric set up and describe the theory as well as the fundamentals underlying it and skip the proofs.

There are two type of parameters – location parameters and scale parameters. For example, in $N\left(\mu,\sigma^2\right)$, μ is the location parameter and σ^2 is the scale location parameter. We consider the set up of nonparametric estimation and testing of hypothesis of location parameter in one sample and two-ample location problems.

Let $F\left(x-\theta\right)$ be a symmetric distribution with centre of symmetry θ and assume it to be absolutely continuous.

Estimation of one sample location parameter

Let $X_1, X_2, ..., X_n$ be a sample from the distribution $F(x-\theta)$. Since the distribution of , $X_1-\theta, X_2-\theta, ..., X_n-\theta$ is symmetric about the origin, so the best estimator of θ is that value of θ for which the values $X_1-\theta, X_2-\theta, ..., X_n-\theta$ give the best balance relative to origin.

Consider a test statistic $h(X_1, X_2, ..., X_n)$ for testing the hypothesis that the common distribution of $X_1-\theta, X_2-\theta, ..., X_n-\theta$ is symmetric about the origin. Suppose *h* is evaluated for $X_1-\theta, X_2-\theta, ..., X_n-\theta$. Now $X_1-\theta, X_2-\theta, ..., X_n-\theta$ will be best balanced around the origin when $h\left(X_1-\theta, X_2-\theta, ..., X_n-\theta\right)$ take on the value which gives strongest support to the hypothesis. This will be the case when *h* takes on its central value. One may thus expect the resulting estimator to share some of the properties of the test from which it is derived..

Sign test statistic and Wilcoxon signed rank test statistic are some commonly used statistic for testing the hypothesis about the location parameter. We first discuss these two procedure.

Sign-test statistic

The procedure is as follows:

- Evaluate $X_1-\theta, X_2-\theta, ..., X_n-\theta.$
- Count the number of $(X_i-\theta)'s$ which are positive.
- The greatest support to the hypothesis is given if half of $(X_i-\theta)$ are positive and half of $\left(X_i-\theta\right)$ are negative.
- It happens if θ is the median of X_i.
- So the sign test leads to estimate $\tilde{\theta}=\text{median}(X_i)$ where $\text{median}(X_i)$ denotes the median $X_1, X_2, ..., X_n$.

Wilcoxon signed rank test statistics

The procedure is as follows:

- Evaluate $(X_i-\theta)+(X_j-\theta)$ with $i \le j, i, j = 1, 2, ..., n.$
- Count the number of $(X_i-\theta)+(X_j-\theta)$ which are positive.
- Count the number of such pairs $(i, j).$
- Such count takes its central value if half of the sums $(X_i-\theta)+(X_j-\theta)$ for all $i \le j$ are positive and half are negative.
- Thus the corresponding estimator is the median of $\frac{n(n+1)}{2}$ values $\left(\frac{X_i+X_j}{2}\right)$ for $1 \le i \le j \le n,$ i.e.,

$$\tilde{\tilde{\theta}} = \underset{i<j}{Median}\left(\frac{X_i+X_j}{2}\right).$$

Estimation and testing of two-sample location parameters

Let $F(x)$ and $G(x)=F(x-\theta)$ are absolutely continuous distribution functions. A sample $X_1, X_2, ..., X_m$ of size n is drawn from $F(x)$. Another sample $Y_1, Y_2, ..., Y_n$ of size n is drawn from $G(x)$. Both the samples are independently drawn.

Since the variables $X_1, X_2, .., X_m, Y_1-\theta, Y_2-\theta, ..., Y_n-\theta$ have the same distribution, we may estimate θ by $\hat{\theta}$ which is the amount by which the Y values have to be shifted to give the Y values the best possible agreement with the X - values. To achieve this, consider a test statistic $h(X_1, X_2, ..., X_m; Y_1, Y_2, ..., Y_n) = h(X, Y)$, say for testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0$.

Suppose under H_0, $h(X, Y)$ is symmetrically distributed about some value μ. Let the two sided test rejects H_0 whenever

$$h(X, Y) \geq \mu + C \text{ or } h(X, Y) \leq \mu - C.$$

Then the most supportive value of H_0 is the value which is closest to the centre μ of the distribution.

The sets

$$X_1, X_2, ..., X_m \text{ and } Y_1-\hat{\theta}, Y_2-\hat{\theta}, ..., Y_n-\hat{\theta}$$

will therefore be considered as being in the best agreement if they assign to the statistic $h(X_1, X_2, ..., X_m, Y_1-\hat{\theta}, Y_2-\hat{\theta}, ..., Y_n-\hat{\theta})$ this closest value.

The Wilcoxon statistic is based on such $h(X,Y)$ statistic and is denoted by $W_{X,Y}$. By definition, $W_{X,Y-\theta}$ denotes the number of pairs (*i, j*) for which $Y_j - \theta > X_i$ or $Y_j - X_i > \theta$. So $W_{X,Y-\theta}$ will take its central value if half of the differences $Y_j - X_i$ are greater than θ and rest half are less than θ. Thus the value $\hat{\theta}$ of θ for which this is achieved is the median of *mn* differences $Y_j - X_i$. Let $d_{(1)}, d_{(2)}, ..., d_{(mn)}$ denote the ordered set of *mn* differences $Y_j - X_i$. Now *mn* can be an even number or an odd number.

Let *mn* be an even number, say *mn* = 2*k* (*k* is integer) then $k = \frac{mn}{2}$ and then

$$\hat{\theta} = \frac{d_{(k)} + d_{(k+1)}}{2}$$
$$= \text{median } (Y_j - X_i).$$

If *mn* is an odd number, say *mn* = 2*k*+1 then

$$\hat{\theta} = d_{(k+1)}$$
$$= \text{median}\,(Y_j - X_i).$$

Alternative expression of $W_{X,Y}$ is given by **Mann-Whitney statistic** as follows:

- Rank the (*m* + *n*) observations $X_1, X_2, ..., X_m, Y_1, Y_2, .., Y_n$ in the ascending order of magnitude.
- Let $R_1, R_2, ..., R_n$ be the ranks of $Y_1, Y_2, ..., Y_n$ in the combined sample.
- Let $R = \sum_{i=1}^{n} R_i$ then $W_{X,Y} = R - \frac{n(n+1)}{2}$

is the **Mann-Whitney statistic.**

For testing $H_0 : \theta = 0$ versus $H_1 : \theta \neq 0,$ the decision rule based on Wilcoxon test based on *R* rejects H_0 whenever

$$R \leq C_1 \text{ or } R \geq C_2 \quad (C_1 < C_2)$$

where C_1 and C_2 are constants such that

$$P(R \leq C_1) + P(R \geq C_2) = \alpha$$

when H_0 is true and α is the level of significance.

If the roles of the two treatments are symmetric, then C_1 and C_2 are chosen under H_0 as

$$P(R \leq C_1) = P(R \geq C_2) = \frac{\alpha}{2}.$$

Since the distribution of *R* is symmetric about $\frac{n(m+n+1)}{2}$, so Wilcoxon test based on *R* rejects H_0 whenever

$$\left| R - \frac{n(m+n+1}{2} \right| \geq C$$

or equivalently

$$\left| W_{XY} - \frac{mn}{2} \right| \geq C$$

where $P\left[\left| W_{X,Y} - \frac{mn}{2} \right| \geq C \right] = \alpha$ under H_0

and *C* depends on the size α of the test.

Now we consider the analysis of variance in the set up of one way classification.

One way classification (Location model)

Consider $p(\geq 2)$ independent set of random variables $Y_{ij}, i = 1,2,..., p,\ j = 1,2,..., n_i.$

Let $Y_{i1}, Y_{i2},..., Y_{in_i}$ be a random sample of size n_i from the continuous distributions $F_i(y),\ i = 1,2,.., p$ and let

$$F_i(y) = F(y - \theta_i) \quad \text{for all } y \text{ where } \theta_i \text{ is the location parameter.}$$

First consider the problem of estimating the differences

$$\Delta_{ij} = \theta_i - \theta_j$$

Using Wilcoxon statistic, we have

$$\hat{\Delta}_{ij} = \text{median}(y_{i\alpha} - y_{j\beta})$$

is the median of $n_i n_j$ differences

$$Y_{i\alpha} - Y_{j\beta};\ \alpha = 1,2,..., n_i;\ \beta = 1,2,..., n_j$$

We may consider $\hat{\Delta}_{ij}$ as an estimator of Δ_{ij}.

These estimators are not compatible in the following sense. Suppose we directly estimate $\theta_2 - \theta_1$ by $\hat{\Delta}_{21}$. Since $\theta_2 - \theta_1 = (\theta_3 - \theta_1) + (\theta_2 - \theta_3)$, so $\Delta_{21} = \Delta_{31} + \Delta_{23}$ but $\hat{\Delta}_{21}, \hat{\Delta}_{31}$ and $\hat{\Delta}_{23}$ may differ.

To avoid such situations, replace $\hat{\Delta}_{ij}$ as

$$\hat{\Delta}_{ij} = \hat{\Delta}_{io} - \hat{\Delta}_{oj}$$

where

$$\hat{\Delta}_{ko} = \frac{1}{p}\sum_j \hat{\Delta}_{kj}.$$

If sample sizes n_i are not equal, then

$$\hat{\hat{\Delta}}_{ij} = \hat{\hat{\Delta}}_{io} - \hat{\hat{\Delta}}_{oj}$$

where $$\hat{\hat{\Delta}}_{ko} = \frac{1}{\sum_j n_j} \sum_j n_j \hat{\Delta}_{kj}.$$

To estimate a contrast $\varphi = \sum_{j=1}^{p} \ell_j \theta_j$ with $\sum_{j=1}^{p} \ell_j = 0$, first express it is terms of elementary contrast $\varphi = \sum_i \sum_j d_{ij} \Delta_{ij}$, say.

An estimate of φ is

$$\hat{\varphi} = \sum_i \sum_j d_{ij} \left(\hat{\Delta}_{io} - \hat{\Delta}_{jo} \right).$$

In case of unequal sample sizes, use the following estimator of ϕ as

$$\hat{\hat{\varphi}} = \sum_i \sum_j d_{ij} \left(\hat{\hat{\Delta}}_{io} - \hat{\hat{\Delta}}_{jo} \right).$$

Based on this, we proceed as follows for the test of hypothesis.

Test of hypothesis

- Let $N = \sum_{i=1}^{p} n_i$.
- Rank all the *N* observations y_{ij}, $i = 1, 2, .., p,\ j = 1, 2, ..., n_i$ in ascending order of magnitude.
- Let $rank(Y_{ij}) = R_{ij}$.
- Null hypothesis $H_0 : \theta_1 = \theta_2 = ... = \theta_p$.

Let θ_i represents the response of *p* different treatments under consideration in the analysis of variance model. We assume that the treatments mainly affect the response level under this model and there is a natural order among the treatments in the sense that one tends to give the lowest response, another gives the next lowest and so on.

The average rank provides an indication of the position of a particular treatment in this ordering. The average rank of i^{th} treatment is given by

$$R_{io} = \frac{1}{n_i} \sum_{j} R_{ij}.$$

When H_0 is true, then R_{i0} will be close to $R_{oo} = \frac{1}{N}\sum_i \sum_j R_{ij}$. The validity of the hypothesis H_0 can be checked by judging the closeness of R_{i0} to R_{00}. The **Kruskal-Wallis test statistic** on R_{i0} is given by

$$Q = \frac{12}{N(N+1)} \sum_{i=1}^{p} \left(R_{io} - \frac{N+1}{2} \right)^2$$

$$= \frac{12}{N(N+1)} \sum_{i=1}^{p} \frac{R_i^2}{n_i} - 3(N+1); \quad R_i = n_i R_{io}$$

which rejects H_0 whenever

$$Q \geq C$$

where C depends on size α of the test.

For $p = 2, \; Q = W_{X,Y}$ where $W_{x,y}^2$ is the Wilcoxon statistic under H_0,

$$P(R_{ij} = r_{ij}; \; j = 1, 2, ..., n_i, \; i = 1, 2, ..., p) = \frac{1}{\binom{n}{n_1 n_2 ... n_p}}.$$

The value of C can be determined by using the available tables for the distribution of Q.

For a large N, the distribution Q can be approximated by Chi-square distribution with (p - 1) degrees of freedom.

Case of tied observations

Suppose the observations are tied and out of N observations X_{ij} take on n distinct values.

Suppose

- k_1 observations are equal to the smallest value,
- k_2 observations are equal to the second smallest value

and

- k_n observations are equal to the largest value.

R_{ij}^{*} : midrank of X_{ij}

$$R_i^{*} = \sum_i R_{ij}.$$

The Kruskel-Wallis test statistic for the tied observations is

$$Q^{*} = \frac{\frac{12}{N(N+1)}\sum_i \frac{R_i^{*2}}{n_i} - 3(N+1)}{1 - \sum_i \frac{(k_i^3 - k_i)}{N^3 - N}}$$

which rejects H_0 whenever

$$Q^{*} \geq C$$

where the constant C depends on the size α of the test. Except for very small values of N, the distribution of Q^{*} under H_0 can be approximated by the Chi-square distribution with (p - 1) degrees of freedom.