International Journal of Education Humanities and Social Science

ISSN: 2582-0745
Vol. 3, No. 04; 2020

RESEARCH ON NONPARAMETRIC STATISTICAL METHOD BASED ON MULTI-SAMPLE PROBLEMS

Xiaotong Cai[1] and Xia Liu[2*]
[1,2] Department of Statistics, School of Mathematics and Statistics, Shandong University of Technology

***Corresponding author**
E-mail addresse:lxia010@163.com

ABSTRACT
As a branch of mathematical statistics, nonparametric statistical method is the main method to solve many data problems with unknown distribution. The statistical model of things can be established and inferred using nonparametric statistical method. In this paper, we mainly introduce the nonparametric statistical methods for two samples and multi-sample problems, point out their test statistics and the gradual distribution. Finally we analyze the advantages of nonparametric statistics.

Keywords: nonparametric statistics, multi-sample problems

1. INTRODUCTION

Parametric statistical method and non-parametric statistical method together constitute statistical analysis method, which is the basic content of statistical analysis. Parameter statistical method means to estimate unknown parameters through samples assuming that the theoretical distribution of the overall is known and some parameters in the distribution are unknown, and to test the rationality of these unknown parameters by means of hypothesis testing. However, in the process of data analysis, due to various reasons, people are often unable to make specific assumptions about the overall, so the parametric statistical method is no longer applicable, and the non-parametric statistical method emerges as the times require. Nonparametric estimation analysis method does not make assumptions about the overall distribution, or only gives very general assumptions, such as continuous distribution, symmetric distribution and so on. Such method does not involve overall parameters or does not depend on strict assumptions about the overall distribution is called nonparametric method.

To observe the development status of non-parametric statistical analysis in the world, non-parametric statistical methods have been widely used in some fields with development potential. Pinto Carolina Cristiane、Calazans Giovanna Moura and Oliveira Sílvia Corrêa[1] assessed the spatial variations in the surface water quality of the Velhas River Basin, Brazil, using multivariate statistical analysis and nonparametric statistics. They think this association between multivariate statistical techniques and nonparametric tests was effective for the classification and processing of large water quality datasets and the identification of major differences between water pollution sources in the basin.P. Yu. Kostenko, V. V. Slobodyanyuk,K. S. Vasiuta and V.

ISSN: 2582-0745
Vol. 3, No. 04; 2020

I. Vasylyshyn[2] used nonparametric Statistic to assess measure of Filtering Quality of Image Noise. In the paper[3], nonparametric statistic method is used in the spent fuel rack model to analyze the uncertainty of manufacturing parameters, nuclide composition from assembly depletion calculation and reaction cross sections in criticality calculation code.Nonparametric statistics method is also used in physics. J. Situmorang and S. Santoso[4]studiedsafety indicators in nuclear installations using Mann Withney nonparametric statistic technique. The study was conducted using nonparametric statistical techniques, especially Mann Whitney test techniques that emphasize the presence or absence of differences between two examples of independent or dependent populations. In the paper[5],authors used genetic diversity and nonparametric statistics to identify possible ISSR marker association with fiber quality of pineapple. Studies demonstrate the potential use of pineapple fibers in composites. A. V. Morgunova and O. S. Sazhina[6] studied the use of nonparametric methods of mathematical statistics to search for cosmic strings. Nonparametric rank tests are used to handle small samples under an unknown distribution law.

Nonparametric statistical methods based on two sample problems

2. MOOD MEDIAN TEST

We suppose sample $x_1, x_2, ..., x_m$ and $y_1, y_2, ..., y_n$ are derived from independent continuous random variables X and Y, and the medians of population X and Y are denoted as me_x and me_y .Table 1 is the fourfold table of the mood median test.

Table 1 The fourfold table of the mood median test

	$<me$	$>me$	合计
Sample of X	N_{11}	N_{12}	N_{1+}
Sample of Y	N_{21}	N_{22}	N_{2+}
Total	N_{+1}	N_{+2}	N

In the table, N_{11} and N_{12} is the number of observations in sample X that is greater or less than the median of the combined samples. N_{21} and N_{22} is the number of observations in sample Y that is greater or less than the median of the combined samples. Here, we take N_{11} as the test statistic. N_{11} follows the hypergeometric distribution and the distribution law is $P(N_{11}, N_{1+}, N_{+1}, N)$ ：

$$P(N_{11}, N_{1+}, N_{+1}, N) = \frac{\binom{N_{+1}}{N_{11}}\binom{N_{+2}}{N_{12}}}{\binom{N}{N_{1+}}}.$$

ISSN: 2582-0745
Vol. 3, No. 04; 2020

P-Value of hypothesis test:

(1) H_0: $me_x = me_y$, H_1: $me_x > me_y$, we reject H_0 when N_{11} is small, P-Value is

$$\sum_{k \le N_{11}} P(N_{11}, N_{1+}, N_{+1}, N).$$

(2) H_0: $me_x = me_y$, H_1: $me_x < me_y$, we reject H_0 when N_{11} is large, P-Value is

$$\sum_{k \ge N_{11}} P(N_{11}, N_{1+}, N_{+1}, N).$$

(3) H_0: $me_x = me_y$, H_1: $me_x \ne me_y$, we reject H_0 when N_{11} is small or large. Now P-Value should be discussed in two ways：

- When N_{11} is small, $\sum_{k \le N_{11}} P(N_{11}, N_{1+}, N_{+1}, N) \le 0.05$, P-Value is:

$$\sum_{k \le N_{11}} P(N_{11}, N_{1+}, N_{+1}, N) + \sum_{k \ge a} P(N_{11}, N_{1+}, N_{+1}, N), \quad (1)$$

$$a = \inf\{a : \sum_{k \le N_{11}} P(N_{11}, N_{1+}, N_{+1}, N) \ge \sum_{k \ge a} P(N_{11}, N_{1+}, N_{+1}, N)\}.$$

- When N_{11} is large, $\sum_{k \ge N_{11}} P(N_{11}, N_{1+}, N_{+1}, N) < 0.05$, P-Value is:

$$\sum_{k \ge N_{11}} P(N_{11}, N_{1+}, N_{+1}, N) + \sum_{k \le a} P(N_{11}, N_{1+}, N_{+1}, N), \quad (2)$$

$$a = \inf\{a : \sum_{k \ge N_{11}} P(N_{11}, N_{1+}, N_{+1}, N) \ge \sum_{k \le a} P(N_{11}, N_{1+}, N_{+1}, N)\}.$$

Wilcoxon rank sum test

We suppose sample $x_1, x_2, ..., x_m$ and $y_1, y_2, ..., y_n$ are derived from independent continuous random variables X and Y,and the distribution function is denoted as $F(x)$ and $G(y)$.

The process of Wilcoxon rank sum test is as follows:

Step1:Determine the original hypothesis H_0: $P(X > Y) = 0.5$.

Step2: Determine the tests statistics and their distributions. The basic idea of the rank sum test is sorting N(N=m+n) samples after mixing them, R_y represents the rank of y_j in the combined samples. The rank sum of sample $y_1, y_2, ..., y_n$ is W_y ：

ISSN: 2582-0745
Vol. 3, No. 04; 2020

$$W_y = \sum_{j=1}^{n} R_j. \qquad (3)$$

When n and m are large, we consider the limit distribution of W_y .When H_0 is true, the exception and variance of W_y are $E(W_y) = n(N+1)/2$, $D(W_y) = nm(N+1)/12$.

In this case, the gradual distribution of W_y is normal, denoted as

$$W_y \sim AN(\frac{n(N+1)}{2}, \frac{nm(N+1)}{12}). \qquad (4)$$

Step3: Determine significance level α , so rejection region of H_0 is

$\{W_y > \mu_{\frac{\alpha}{2}}$ 或 $W_y < -\mu_{\frac{\alpha}{2}}\}$,

$\mu_{\frac{\alpha}{2}}$ is the upper quartile of normal distribution.

Step4: Calculate W_y from the sample. If W_y falls in the acceptance domain, then H_0 is true, X and Y can be thought of as identically distribute. If W_y falls in the rejection region, then H_1 is true. In fact, if $F(x) > G(y)$,then $P(X > Y) > 0.5$,so W_y has a larger trend；conversely, if $F(x) \le G(y)$, then $P(X > Y) < 0.5$, so W_y has a smaller trend.

Wilcoxon rank sum test was also used mean value method when there were equal values in the observed values. According to mean value method, W_y is

$$W_y = \sum_{i=1}^{n} a(R_i). \qquad (5)$$

Function $a(r)$, $r = 1,2,...,n$ is scoring function, when the length of the node is 1, $a(R_i) = R_i$ and $a(R_i)$ is equal to the average of the ranks when the length of the node exceed 1. We can know the exception and variance of W_y by mean value method:

$$\begin{aligned} E(W_y) &= \frac{n(N+1)}{2}, \\ D(W_y) &= nm(N+1)/12 - nm\sum_{t=1}^{g}(\tau_t^3 - \tau_t)/(12N(N-1)). \end{aligned} \qquad (6)$$

In this case, the gradual distribution of W_y is also normal, denoted as

$$W_y \sim AN(\frac{n(N+1)}{2}, D(W_y) = nm(N+1)/12 - nm\sum_{t=1}^{g}(\tau_t^3 - \tau_t)/(12N(N-1))). \qquad (7)$$

ISSN: 2582-0745
Vol. 3, No. 04; 2020

Case1 To compare the mileage of two types of cars per gallon, we choose 12 cars from each model at random, and ask each car to travel 1000 kilometers at high speed. Sample data foreach car's miles per gallon and their rank in the combined samples are recorded in table 2.

Table 2 miles per gallon and their rank

Type I			TypeII		
The car	Mileage(miles)	Rank	The car	Mileage(miles)	Rank
1	20.6	21	1	21.3	24
2	19.9	16	2	17.6	4
3	18.6	8	3	17.4	3
4	18.9	11	4	18.5	7
5	18.8	9.5	5	19.7	13
6	20.2	18	6	21.1	23
7	21.0	22	7	17.3	2
8	20.5	19.5	8	18.8	9.5
9	19.8	14.5	9	17.8	5
10	19.8	14.5	10	16.9	1
11	19.2	12	11	18.0	6
12	20.5	19.5	12	20.1	17

In this example, m=n=12, N=24, there are three knots in the sample data, each of which has a length of 2. After calculating the rank sum of the data of a car'smiles per gallon W_1=185.5 .According to formula(7), the expectation and variance of the test statistic W_1 is calculated : $E(W_1)=150$, $D(W_1)=299.61$.

Alternative hypothesis: There is a difference in miles per gallon between the two types.We used statistical software Minitab, and according to thegradual distribution of W_1, We can get P-Value:

$2P(\mathrm{N}(150,299.61)\geq 185.5)=0.0432$.

The P-value is small, so we can reject the null hypothesis and assume that there is a difference in miles per gallon between the two types. As you can see from the sample data, type Igot a lot of miles per gallon.

Wilcoxon rank sum test is an improvement on the symbol test. The symbol test only considers the symbol information without considering the size information.But the rank sum test takes into account the size information of the sample on the basis of the total sample and improves the accuracy rate. And the method is simple, easy to operate, don't need to know the specific distribution. The disadvantage is that information is lost, and the testing effect will be reduced if the data suitable for parameter testing is analyzed by non-parametric method.

Rank test method for two samples scale parameters

ISSN: 2582-0745
Vol. 3, No. 04; 2020

Now we consider whether the two populations have the same measures of dispersion when there are no differences in location parameters. We define $bX \overset{d}{=\!=\!=} Y$, and we call b the scale parameter. For example, H_0: $b=1$ H_1: $b>1$.

The main method we adopt is Siegel- Turkey test. The scoring function $a(r)$ is a single valley function, and when the rank of y_i is R_i, the score of y_i is defined as $a(R_i)=R_i$. $a(1)=N$, $a(N)=N-1$, $a(N-1)=N-2$, $a(2)=N-3$, $a(3)=N-4$, $a(N-2)=N-5$,.... For example as N=9,

r	1	2	3	4	5	6	7	8	9
$a(r)$	9	6	5	2	1	3	4	7	8

The test statistic is denoted as $S_y=\sum_{i=1}^{n} a(R_i)$,when S_y is large, we reject the nullhypothesis.

It is worth noting that when $x_1, x_2, ..., x_m$ and $y_1, y_2, ..., y_n$ have the same distribution,

the Siegel-Turkey test statistics and Wilcoxon rank sum test statistics have the same

distribution, so they have the same expectation and variance as well as the same

progressive distribution. Therefore, when H, is true, that is, when X and Y are the same the gradual distribution. Therefore, when H_0 is true, that is, when X and Y are the same continuous distribution

$$E(S_y)=\frac{n(N+1)}{2},$$

$$D(S_y)=\frac{nm(N+1)}{12},$$

$$\frac{S_y - n(N+1)/2}{\sqrt{mn(N+1)/12}} \xrightarrow{L} N(0,1) \qquad m,n\to\infty. \quad (8)$$

When applying the Siegel Turkey test statistics to solve the scale-parameter test problem, we can consult the critical value table of Wilcoxon rank sum test to obtain the P-value and critical value for the Siegel Turkey test. .

Nonparametric statistical methods based on multiple sample problems

Kruskal-Wallis test

ISSN: 2582-0745
Vol. 3, No. 04; 2020

Set k continuous random variable populations: $X_1, X_2, ..., X_k$, the capacity is $n_i, i = 1,2,...,k$. The total sample size is $N = \sum_{i=1}^{k} n_i$, and let's say that these k populations are just different location parameters.

Kruskal-Wallis test is used to test whether all the k position parameters $\theta_1, \theta_2, ..., \theta_k$ are equal.

$H_0 : \theta_1 = \theta_2 = ... = \theta_k, H_1 : \theta_1, \theta_2, ..., \theta_k$ are not all equal.

We sum the k samples together and calculate that the rank of x_{ij} in the sum samples called R_{ij}. The basic idea of kruskal-Wallis test is to replace x_{ij} with the rank R_{ij} of x_{ij}, and then do statistical analysis using ANOVAmethod. The selected test statistic is

$$H = \frac{12}{N(\mathrm{N}+1)} SSB = \frac{12}{N(\mathrm{N}+1)} \sum_{i=1}^{k} n_i (\overline{R_i} - (\mathrm{N}+1)/2)^2, \quad (9)$$

$\overline{R_i} = \sum_{j=1}^{n_i} R_{ij} / n_i$ is the mean of the rank of a sample from population i. When H is high, we assume that these k position parameters are not all equal.

The above test statistics are sometimes complicated to calculate. Now we consider

the gradual distribution of the H,

$$\begin{aligned} E(SSB) &= \sum_{i=1}^{k} n_i E(\overline{R_i} - (N+1)/2)^2 \\ &= \sum_{i=1}^{k} n_i \frac{(N-n_i)(N+1)}{12 n_i} \\ &= \frac{N+1}{12} \sum_{i=1}^{k} (N-n_i) \\ &= \frac{N(N+1)(K-1)}{12} \end{aligned} ,$$

then

$$E(H) = \frac{12}{N(\mathrm{N}+1)} E(SSB) = k-1. \quad (10)$$

It can be proved that when H is true,if $\min\{n_1, n_2, ..., n_k\} \to \infty$, the gradual distribution of H is $\chi^2(k-1)$.

Case2 The managers of a company are from three universities, including n_1=7 managers fromuniversity A, n_2=6 managers from the universityB, n_3=7 managers from university C.The company $N = n_1 + n_2 + n_3 = 20$ managers' annual performance scoreare listed in table 3. The annual

ISSN: 2582-0745
Vol. 3, No. 04; 2020

performance scoreare from 0 to 100, 100 of them are the highest. Question:Are there any differences in the performance of managers' performance from the three universities?

Table 3 Annual performance scores of 20 managers

university A	universityB	university C	university A	universityB	university C
84	75	58	72	95	65
72	65	78	90	69	72
75	80	80	75		42
95	55	62			

We used statistical software Minitab to calculate test statistics H=4.06.Becausethe sample size is relatively large, we use the gradual distribution of H is $\chi^2(2)$. We can get P-Value:

$P(\chi^2(2) \geq 4.06) = 0.129$.

P-value is greater than 0.05, so based on the available data, we cannot assume that the performance of managers from the three universities is different.

The rank test method of trend

Set k continuous random variable populations: $X_1, X_2, ..., X_k$, the capacity is $n_i, i = 1,2,...,k$. The total sample size is $N = \sum_{i=1}^{k} n_i$, and let's say that the distribution function of X_i is $F(x-\theta_i)$.

The null hypothesis and the alternative hypothesis of the monotonically rising tendency test are

$H_0 : \theta_1 = \theta_2 = ... = \theta_k, H_1 : \theta_1 \leq \theta_2 \leq ... \leq \theta_k$且$\theta_1 < \theta_k$.

The null hypothesis and the alternative hypothesis of the monotonically decreasing tendency test are

$H_0 : \theta_1 = \theta_2 = ... = \theta_k, H_1 : \theta_1 \geq \theta_2 \geq ... \geq \theta_k$且$\theta_1 > \theta_k$.

The test statistic of the trend test is

$$T = \sum_{i=1}^{k} w_i R_{i+}, \quad w_i = 2\sum_{t=1}^{i} n_t - n_i, i - 1,2,...,k. \tag{11}$$

ISSN: 2582-0745
Vol. 3, No. 04; 2020

Moreover, when T is relatively large, it is considered to have a monotonically increasing trend, and when T is relatively small, it is considered to have a monotonically decreasing trend.

When H_1 is true, k populations obey the same continuous distribution, we have

$$E(T) = N^2(N+1)/2,$$

$$D(T) = N(N+1)(\sum_{i=1}^{k} n_t w_i^2 - N^3)/12.$$

Similarly, the gradual distribution of T is normal distribution

$$T \sim AN(N^2(N+1)/2, N(N+1)(\sum_{i=1}^{k} n_t w_i^2 - N^3)/12). \quad (12)$$

3. CONCLUSIONS

Non-parametric statistical methods can be used to deal with Interval Data like parametric statistical methods, and are more suitable to deal with classified data and Ordinal Data. The starting point of non-parametric statistical methods is to assume that the theoretical distribution of the population is unknown, which is very common in practical applications. The non-parametric statistical method reduces the dependence on the hypothesis in practical application, thus making the research on the multiple problems more objective, the possibility of model error is less and the model is more stable.

Non-parametric methods are more effective than parametric methods when distributions are unknown, and their advantages are summarized as follows:

(1) The non-parametric statistical method does not need to assume what the distribution of the population is, and it can only make judgments on the issues of concern through samples。

(2)Non-parametric statistics has good robustness. The non-parametric statistical method estimates characteristics of the population based on the "general" information in the sample. It can be seen that when the population model is slightly changed, there is no significant impact on the estimation.

(3)It’s easy to understand and calculate. It’s a good choice when you need results badly.

REFERENCES

[1]Pinto Carolina Cristiane, Calazans Giovanna Moura, Oliveira Sílvia Corrêa,

Assessment of spatial variations in the surface water quality of the Velhas River Basin, Brazil, using multivariate statistical analysis and nonparametric statistics, Environmental monitoring and assessment, 10.1007/s10661-019-7281-y, 164

ISSN: 2582-0745
Vol. 3, No. 04; 2020

[2]P. Yu. Kostenko, V. V. Slobodyanyuk, K. S. Vasiuta, V. I. Vasylyshyn, Measure of Filtering Quality Assessment of Image Noise Using Nonparametric Statistic, Radioelectronics and Communications Systems, 2020, Vol.63 (7), pp.201-212, Springer Nature Journal, 10.3103/S0735272720040032

[3] Tian Chen, Ji Xing, Xiao-dong Huo. Study on nonparametric statistic method applied to nuclear criticality safety analysis of spent fuel rack. Annals of Nuclear Energy,10.1016/j.anucene.2019.107065

[4]J. Situmorang,S. Santoso, Perception Study of Safety Indicators in Nuclear Installations using Mann Withney Nonparametric Statistic Technique,10.1088/1742-6596/1198/2/022058

[5] Silva Julianna M, Lima Paulo R L, Souza Fernanda V D, Ledo Carlos A S; Souza Everton H, Pestana Katia N, Ferreira Cláudia F, Genetic diversity and nonparametric statistics to identify possible ISSR marker association with fiber quality of pineapple, Anais da Academia Brasileira de Ciencias,10.1590/0001-3765201920180749, e20180749

[6] A.V.Morgunova, O.S.Sazhina,The Use of Nonparametric Methods of Mathematical Statistics to Search for Cosmic Strings, Moscow University Physics Bulletin, 2019, Vol.74 (5), pp.529-536,10.3103/S0027134919050102