Nonparametric Test for Multiple Populations and Its Application

Yingying Zhang

Shandong University of Management, Jinan ,250300 China

Email: 1831197929@qq.com

Keywords: Overall; Non-parametric test; Application

Abstract: spss single-sample non-parametric test is to analyze the distribution of a single population. There are mainly binomial distribution test, chi-square test, and KS test. However, during the test of various parameters, some data distributions are not easy to discern. In order to better handle this kind of data, non-parametric data of multiple populations can be tested. The purpose of this article is to elaborate a number of nonparametric tests for the population and to explain their practical applications.

Introduction

The parametric test method is a method of testing some of the parameters under the assumption of the overall distribution of the hypothesis. Therefore, the parameter test requirements are more reflected in the distribution of related variables. However, in the actual application process, there are situations in which the overall distribution is unknown or the distribution type is unknown, and the parameter inspection requirements are not met, and the parameter inspection method cannot be adopted. For this reason, the non-parametric inspection method came into being. Due to the relatively large amount of data in multiple populations, all samples can be arranged according to the order of the data from small to large, and a non-parametric test method can be performed on it. This method has wide application prospects in many fields.

1.Application of Non-parametric Inspection

1.1The overall distribution is unknown, and the overall distribution needs to be estimated

During the actual analysis, it was found that the distribution of the research objects is very complicated, and the distribution of the research objects itself is a branch of research. In addition, parametric testing is used to analyze the research object in depth, but it is not clear whether the distribution of each variable meets the requirements of parametric testing. Non-parametric testing can be used in both cases.

1.2The overall distribution is known, and the variables cannot meet the requirements

Since the parameter test is mainly implemented on the basis of variable distribution, some parameters of the distribution of the data and variables are used to obtain a known statistic of a distribution to test its hypothesis. The parameter test requires that the variable be a fixed distance variable, but for a sequence or class variable, the parameter test method is no longer used, and if it is forcibly tested, it may draw wrong conclusions. For this reason, for ordinal and categorical variables, non-parametric tests are usually adopted.

2. Nonparametric Tests for Multiple Populations

The non-parametric test of multiple populations refers to the analysis of the distribution or median difference of multiple populations through multiple independent sample data to determine whether there is a significant difference. Non-parametric tests for multiple populations generally use multiple independent samples to obtain multiple sets of data. Multiple overall non-parametric test methods developed in the spss software, including the Jonckheere-Terpstra test, the

DOI: 10.25236/icme.2019.057

Kruskal-Wallis test, and the median test. In contrast, three independent samples can be obtained through independent sampling, that is, three populations.

2.1Jonckheere-Terpstra test

The Jonckheere-Terpstra test was proposed by two scholars, Jonckheere and Terpstra. Just as there are unilateral and bilateral tests in the hypothesis test, multiple alternative hypotheses may also be directional. For example, if there is a downward or upward trend in the sample position, from a statistical perspective, is the trend significant.

Suppose k populations come from continuous distribution functions of the same shape, and their position parameters are β1, β2 ... βk .

Suppose the sample size of k populations is ni, i = 1,2 ..., k. Let xij be the j-th independent observation of the i-th population (i = 1,2, ... k, j = 1, 2, ... ni)

Then xij can represent a linear model such as Equation 1:

xij = μ + β1 + εij, i = 1,2, ... k, j = 1, 2, ... ni (Equation 1)

Among them, the errors are independent and identically distributed. Under the independent conditions of multiple populations, the differences of the mean values of the independent data of each population are compared.

2.2Kruskal-Wallis test

The Kruskal-Wallis test is also called the H test. The essence of the test method is to test whether there are significant differences in the distribution of multiple independent samples. The basic idea of the Kruskal-Wallis test is to first sort the number of samples of multiple populations in order and obtain the rank of the variables, and then compare the differences in the mean of the ranks of each population. The population data is fully mixed. If there is a significant difference, the data representing multiple populations is not mixed. If the data of some populations is large, and the data of some populations is small, there is a difference in the distribution of multiple populations. The sample is different from other population samples.

In order to analyze the differences in ranks of multiple populations, an analysis of variance method can be introduced. The method analysis method considers that the total variation of the ranks of multiple populations must be derived from the differences between the populations or from sampling errors within the populations. The former is the group error, the latter is the difference within the group. If the total variation of the ranks of multiple populations is the difference between groups, it means that there is a difference in the distribution within the population. If the total variation of the ranks of multiple populations cannot be explained by the difference between the groups, there is no difference in the distribution of the multiple populations. According to the above description, it is not difficult to know that the purpose of non-parametric testing of multiple populations is to determine whether there is a difference in the distribution of multiple independent samples by using certain independent sample data.

Based on the above description, the Kruskal-Wallis statistic can be constructed, that is, KW = the average of the sum of squares between the groups / the sum of the total squares of the ranks. The original assumption is that there is no difference between the populations. If the assumption is true, the rank average of each population sample and the rank average of all populations is closer, the sum of squares between groups is shown in Equation 2:

$$\text{Sum of squares between groups} = \sum_{i=1}^{k} n_i \left(\frac{R_i}{n_i} - \frac{n+1}{2}\right)^2 \qquad \text{(Eq. 2)}$$

Dividing the statistic of closeness by the average of multiple population rank equations can eliminate the dimensional effect, and the degree of freedom of the variance is expressed as n-1, for this purpose:

$$\text{Average of the sum of the squares of the ranks} = \frac{1}{n-1}\sum_{i=1}^{k}\sum_{j=1}^{n_i}\left(\mathrm{R}_{ij} - \frac{n+1}{2}\right)^2 = \frac{1}{n-1}\sum_{i=1}^{n}\left(\mathrm{i} - \frac{n+1}{2}\right)^2 = \frac{1}{n-1}\left(\sum_{i=1}^{n} i^2 - \frac{n(\mathrm{n}+1)^2}{2}\right) = \frac{1}{n-1}\left(\frac{n(\mathrm{n}+1)(2\mathrm{n}+1)}{6} - \frac{n(\mathrm{n}+1)^2}{4}\right) = \frac{n(\mathrm{n}+1)}{12} \qquad \text{(Eq. 3)}$$

To this end, K-W is:

W = Sum of sum of squares between ranks / average of total sum of ranks of rank

$$= \frac{12}{n(\mathrm{n}+1)}\sum_{i=1}^{k} n_i\left(\frac{R_i}{n_i}-\frac{n+1}{2}\right)^2 = \frac{12}{n(\mathrm{n}+1)}\sum_{i=1}^{k}\frac{{R_i}^2}{n_i}-3(\mathrm{n}+1) \qquad \text{(Eq.4)}$$

Among them, n is the total number, k is the total number of groups, Ri is the sum of the rank of the i-th sample, and ni is the size of the i-th sample, which is the rank value.

If the sample has a knot value, Equation 4 can be adjusted. The correction coefficient C is expressed as follows:

$$\mathrm{C}=1-\frac{\sum(t_j^3-t_j)}{n^3-n} \qquad \text{(Eq. 5)}$$

Among them, is $t_j$the number of j-junction values, and the adjusted statistic is $\mathrm{KW_c}$ = KW / C.

2.3Median test

The median test is to use a sample analysis of multiple populations to test whether there is a significant difference in the median of the population. The basic idea is that if there is no difference in the median of multiple populations, or if there is a common among the medians of multiple populations, The common median should be in the middle of multiple population samples. For this reason, the number of samples that are less than or greater than the median in multiple populations is basically the same.

3. Application Example Analysis

For example, in the comparative analysis of the height of children in the three cities of Changchun, Harbin and Shenyang, three independent samples were obtained. The sample values are mixed to rank, and 15 children from three cities are ranked for height, as shown in Table 1:

Table 1 Children's height and uniform rank values in different cities

Children	A	Rank value	B	Rank value	C	Rank value
1	71	7	70	4	74	12
2	74	12	71	7	75	14
3	75	14	72	9	76	16
4	77	18	73	10	77	18
5	78	20	73	10	77	18
Sum of Rank	-	71	-	40	-	78
Average Rank	-	14.4	-	8.2	-	15.8

Calculate the sum of the ranks of the samples, n1 = 5, n2 = 5, n3 = 5, n = 15, and the analysis can obtain R1 = 71, R2 = 40, R3 = 78, k = 3, H0: three population height distributions the same.

Substituting the data into (Eq. 4) for calculation, the K-W result is obtained as follows:

$$\text{K-W}=\frac{12}{15(21)}\left[\frac{(71)^2}{5}+\frac{(40)^2}{5}+\frac{(78)^2}{5}\right]-3(15+1)=104.7$$

Calculate the correction coefficient C by using Formula 5. From Table 1, it is known that the number of equal ranks of height is 3, 3, 2, 2, 2, and 3 respectively.

C = 1- (33-3 + 33-3 + 23-2 + 23-2 + 23-2 + 33-3) / (153-15) = 0.9733

After adjustment KWc = KW / C = 104.7 / 0.9733 = 107.5721

The test results show that the chi-square distribution with degrees of freedom k-1 = 3, the critical value (3) = 7.815, because 107.5721 ＞ 7.815, and the null hypothesis is rejected, it can be considered that the overall distribution of children's heights in the three cities is significantly different.

The Kruskal-Wallis test can only test whether there are differences in multiple populations. If there are differences in the overall test, but it is not clear what the specific differences are, a pairwise comparison needs to be carried out in this respect. .

Conclusion

The non-parametric test method is used as a type of hypothesis test. The overall characteristics can be judged by sample data. However, in the actual inference process, the following two errors are easy to occur. One is rejection when the null hypothesis is established, and the other is original. Acceptance at the time of the assumption, no matter which one, will misjudge the result. In a sense, the higher the sample size, the more sufficient the sample information will be, and the fewer errors there will be. Due to the obstacles of parameter testing, multiple overall non-parametric testing methods can be adopted to obtain the conclusion will be more reliable.

References

[1] Shi Haiyan, Wei Chun, Li Qingqing. Non-parametric test analysis of student exercise [J]. Mathematics Learning and Research, 2016 (9): 133-134.

[2] Cui Hongfang. Nonparametric Test Using SPSS Software [J]. Sci-Tech Innovation and Application, 2015 (33): 96-96.

[3] Zhang Shenghu, Zhu Jiazheng, Zhang Sanguo. An Efficient Method for Multiple Response Comparison with Covariate Adjustment and Its Application in Genomic Data [J]. Journal of the University of Chinese Academy of Sciences, 2019,36 (2): 155-161.

[4] Yu Lihuang, Lu Li, Zhang Shi, etal. T-wave alternating joint detection method based on particle swarm optimization and non-parametric tests [J]. Journal of Northeastern University (Natural Science), 2013,34 (3): 326-329.

[5] Hu Wenjing, Nie Bin. Partitioning based on non-parametric tests: agglomeration hierarchical clustering change point recognition method [J]. Standard Science, 2013 (12).