Nonparametric Statistical Test Approaches in Genetics Data

Author, Rakesh Saroj

A Data Science Foundation White Paper

March 2019

---

www.datascience.foundation

Data Science Foundation
Data Science Foundation, Atlantic Business Centre, Atlantic Street, Altrincham, WA14 5NQ
Tel: 0161 926 3641 Email: contact@datascience.foundation Web: www.datascience.foundation
Registered in England and Wales 4th June 2015, Registered Number 9624670

ABSTRACT

The biggest challenge of genetic research lies in significant and intellectual analysis of the large and complex data sets generated by the cutting edge techniques like massively parallel DNA sequencing and genome wide analysis. Statistical analyses are the most important of such experimental data. When the data are not normally distributed and using non numerical (rank, categorical) data then use the nonparametric test for exact result of research hypothesis. Order statistics are among the most fundamental tools in non-parametric statistics and inference. Non parametric test does not depend upon parameters of the population from which the samples are drawn, no strict assumption about the distribution of the population. Nonparametric tests are known as distribution free test also because their assumptions are less and weaker than those connected with parametric test. Nonparametric test does not follow probability distribution. To analyze microarrays and genomics data several non-parametric statistical techniques are used like Wilcoxon's signed rank test (pre-post group),Mann-Whitney U test (two groups) or Kruskal-Wallis test (two or more groups).Importance of this paper is to look at the nonparametric test how to use in genetic research and provide the understanding of these test. (9 pt).

1. INTRODUCTION

In some situations, the assumption that data are realizations of Gaussian random variables is not suitable. In the non-parametric context, no assumption is made on the distribution of the differential score, and theoretical quintiles and p-values are not calculable in a close form. Nonparametric methods require assumption like symmetry of distribution and continuity. These test applied if the measurements are nominal, ordinal as well as continuous score, N.P. test cannot estimate the parameters its use only for testing the hypothesis. NP tests are based on the order Statistics. Order statitics are not independent even if original variate values are independent. During measuring of the quantity of mRNA bound to each site of the array, they can determine how genes are expressed under various situations, in different tissues, and in different organisms. Then have become significant technique because several thousand genes can be expressed at one time in one experiment. This facilitates the procedure of gene study clearly. According to The International HapMap Consortium (2003), the statistical analysis and modeling of the links between DNA sequence variants and phenotypes will play a pivotal role in the characterization of specific genes for various diseases and, ultimately, the design of personalized medications that are optimal for inpidual patient. When analyzing the many thousands of genes on a microarray, we would need to check the normality of every gene in order to ensure that appropriate statistical test. There are many sources of variability in microarray experiment and outliers are frequent. The distribution of intensities of many genes may not be normal then apply the nonparametric test. There are a number of nonparametric test used for test one sample nonparametric test are sign test, kolomogorov-smirnov test, Wilcoxon's signed-rank test. Two or more samples nonparametric test are like wald-wolfowiz run test,mann –whiteny U test,kruskal-walis test, Wilcoxon's paired signed-rank test, sign test for paired sample, spearman's rank correlation test,mcnemar's test. Order sample is desirable for the NP test x1 □ x2 □, ...,□ xn . The distribution of the area under the density

function between any two ordered observations in independent of the form of the density function.

Order statistic: If the observations are arranged in any order that is known as order tatistics .All the observations are dependent in the ordered statistics; probability function of ordered statistics is not the same as that of original variables. In statistics, the nth order statistic of a statistical sample is equal to its nth-smallest value. Let X_1 , X_2 ,..., X_r be random variable sample. If x_r is the highest X value in X_1 , X_2 ,..., X_r . The next value is x_r-1 which is less than x_r and x1 is lowest value, then set of values X_1 , X_2 ,..., x_r is descending order and this is known as ordered statistics. This observation can show in ascending order also. Together with rank statistics, order statistics are among the most fundamental tools in non-parametric statistics and inference.

When using probability theory to analyze order statistics of random samples from a continuous distribution, the cumulative distribution function is used to reduce the analysis to the case of order statistics of the uniform distribution. Important special cases of the order statistics are the minimum and maximum value of a sample, and the sample median and other sample quintiles. Nonparametric tests make no or very minimal assumptions about the probability density from which the data are derived. They are used when the sample size is small, when the data are not normally distributed and cannot be approximated as normal, and when using non numerical (rank, categorical) data. The Nonparametric tests are often a good option for small sample sizes (n < 30).

Objective:
The objective of this paper to apply of nonparametric tests and its approach in genetics data on the sampled example and better understanding of these test included statistical hypothesis. This paper will show how the nonparametric tests are very useful if the data is not follow the normality specific in the various genetic researches.

Model Specification of tests:
If study data do not follow the distributional assumptions of parametric methods, even after transformation, or data involve non-interval scale measurements, then non-parametric test is reliable. Thumb rule apply the nonparametric test is SD>1/2 MEAN. There are various one sample, two or more sample non parametric test which are using in the various biological, public health and genetic research. Here I am going to explain only those tests who is mostly use in genetic research and important for the research.

Mann Whitney U/Wilcoxon's Ranked Sum test: When normality assumptions are not satisfied for any one or both of the groups, the equivalent nonparametric. It is alternative of the parametric independent t –test. This test is applied for the find the difference between two independent groups have been drawn from same population.

Assumption:

- Variable of interest is continuous.
- Measurement scale is at least ordinal
- The both sample should be independent.

Let $x_1 \leq x_2 \leq, \ldots, \leq x_n$ and $y_1 \leq y_2 \leq, \ldots, \leq y_n$. be independent ordered samples of size from population. Then the null hypothesis Ho: $f_1(.) = f_2(.)$ and alternative Hypothesis H_1: $f_1(.) \neq f_2(.)$. Where $f_1(.)$ and $f_2(.)$ is p.d.f. of the population. This test is based on the two independent x's and y's combined ordered sample.

The test statistics is given as $U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$, U_1 value can find through this formula. Where n1 = no. of observation in the sample x, R_1= sum of the ranks of the values in sample x.

$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$, U_2 value can find through this formula Where n2 = no. of observation in the sample y, R_2= sum of the ranks of the values in sample y.

For the test statistics we consider the U= smaller value of U_1&U_2 and based on that conclude our null or research hypothesis at the 0.05 significant level. Determine a critical value of U such that if the observed value of U is less than or equal to the critical value, we reject Ho: f_1%(.)= $f_2(.)$ in favor of H_1 and if the observed value of U exceeds the critical value we do not reject Ho: $f_1(.) = f_2(.)$.

Example: In a genetic inheritance study, we want compare the groups X and group Y with respect to the variable MSCE (mean sister chromatid exchange).

The data is as follows:

Group (X): 7.5 8.48 7.65 7.16 8.83 8.76 8.63

Group (Y): 7.32 8.20 7.25 8.14 9.00 7.10 7.20 8.32 8.70

Set up hypotheses and determine level of significance.

H_0: $F_X(x) = F_Y(x)$ (H_0: MSCE distribution is the same as Group X and Group Y)

H_1: $F_X(x) \neq F_Y(x)$ (H_1: MSCE distribution is not same as Group X and Group Y)

The ordered of the group x and group y as follows

		Smallest to Largest ordered of sample		**Rank**	
Group (X)	Group (Y)	Group (X)	Group (Y)	Group (X)	Group (Y)
7.50	7.32		7.10		1
8.48	8.20	7.16		2	
7.65	7.25		7.20		3

7.16	8.14		7.25		4
8.83	9.00		7.32		5
8.76	7.10	7.50		6	
8.63	7.20	7.65		7	
	8.32		8.14		8
	8.70		8.20		9
			8.32		10
		8.48		11	
		8.63		12	
			8.70		13
		8.76		14	
		8.83		15	
			9.00		16
			R_1=67		R_2=69

Then calculate the value

$U_1 = n_1 n_2 + \frac{n_1(n_1+1)}{2} - R_1$ = 7*9 +7*(7+1)/2 -67 = 63+7*8/2-67=24

$U_2 = n_1 n_2 + \frac{n_2(n_2+1)}{2} - R_2$ = 7*9 +9*(9+1)/2 -69 = 63+9*10/2-69=39 Thus test statistics U = smaller value of U_1 & U_2 = 2.

Tabulated value of U for $n_1 = 7, n_2$ = 9 at 0.05 significance level is 12 which is less than calculated value of U (22).Then we reject the H0. It means that MSCE distribution is not same in the Group X and Group Y.

In every test, U_1+ U_2 is always equal to n1*n2. (U_1+ U_2= 22+41=63 and n1*n2 =7*9 = 63)

Wilcoxon's signed-rank test: This test is useful for testing the significance of differences in paired observations. This test is an alternative of the paired Student's t-test for matched pairs, when the population not follow the normality then use this test. In this test we measure a variable in each subject pre and post an intervention.

Assumption:

- Sample must be pair and should be same population.
- Measurement scale is at least ordinal
- Pairs are chosen randomly and independently

Let xi and yi be the pre and post sample size of the population. State the null hypothesis H_0: M_d =

M^0d , H_1: $M_d \neq M^0d$

Calculate each paired differnces, $d_i = x_i - y$ i, where x i,y i are the pairs of observations. Rank the d_i value, ignoring their negative signs and make the rank according to their sign value. Then calculate the Test Statistics W.

$$W = \sum_{i=1}^{Nr} sgn\ (x_i - y_i)$$

Where N is the number of pairs of observations in the sample. Compute the sum of the ranks of the positive d_i, which is W+ and W−, the sum of the ranks of the negative d_i Then compare the calculated value to tabulated value of W at 0.5 level of significance. Based on that we can find the hypothesis. The two-sided test consists in rejecting H_0, if $|W| \geq W_{tab}$. In the test total,

W+ + W−, would be equal to n (n+1)/2.

Example: The genetic disorder autism patients taken for the study, this study measure the behaviors of children affected with autism, before and after a 4 weeks course of meditation. (Order 10-100).

Before	95	80	50	50	80	57	55	20
After	70	60	60	75	40	65	44	25

Set the Hypothesis

H_0: There is no significant effect off meditation on autism after 4 weeks
H_1: There is significant effect off meditation on autism after 4 weeks

Calculate the value in table in given below:

Children	Before Treatment	After 4 Week of Treatment	di =Difference(Before-After)	Ignore sign	Order	Positive rank	Negative rank
1	95	70	15	15	5		1
2	80	60	20	20	8		2
3	50	60	-10	10	10		3
4	50	75	-25	25	11	4	
5	80	40	40	40	15	5	
6	57	65	-8	8	20	6	
7	55	44	11	11	25		7

8	20	25	-5	5	40	8

Then calculate the value of W+ = $W = \sum_{i=1}^{Nr} \text{sgn}\ (x_i - y_i)$ = 23
W- = $W = \sum_{i=1}^{Nr} \text{sgn}\ (x_i - y_i)$ = 13
Test statistics: smaller value of (W+ and W-) = 13

Tabulated value of 3 when n is 8 at 0.05 significance level which is less than calculated value of W (13).Then we reject the H0. It means that there is significant effect off meditation on autism after 4 weeks. We have n (n+1)/2 = 8(8+1))/2 = 8*9/2 = 36, which is equal to W- + W+) 12+ 23 =36.

Kruskal–Wallis test: Kruskal–Wallis is a non-parametric method for testing to compare medians among j comparison groups (j > 2) and this is like the one-way analysis of variance (ANOVA) with the data replaced by their ranks. Kruskal–Wallis test does not follow the Normal distribution, unlike ANOVA.

Assumption:

- Sample should be independent
- Variable of the study is continuous
- Populations are equal except maybe in value of median

We set hypothesis H_0: The j population medians are equal.
H_1: The j population medians are not equal.

Let there be j independent samples from j population with sizes n1, n2,..., nj .

Allocate the rank of each group together from 1 to N = $\sum_{i=1}^{n} ni$, for the ith sample of size ni, then probable sum of rank is = $\frac{ni.N}{N}\frac{(N+1)}{2} = \frac{ni\,(N+1)}{2}$, Ri sum of ranks of observations in sample i
Then the Kruskal-Wallis test H_0 and defined as follows:

$$H = \frac{12}{N(N+1)} * \sum_{i=1}^{J} 1/ni\ \left[R_i - \frac{ni\,(N+1)}{2}\right]^2$$

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

Where N= the total sample size
The statistics H is approximate distributed as x_2 with (j-1) degree of freedom.

Example: To evaluate protein secondary structure through CF AVG, GOR, and PHD three different

methods. We want to test whether all three methods is differ to each other.
Let the data is given below way

CF AVG	0.477	0.467	0.405	0.449
GOR	0.664	0.840	0.604	0.772
PHD	0.898	0.679	0.857	0.790

The above problem defines the hypothesis is given below:
H_0: All three methods are equal
H_1: All three methods are not equal

Calculate the value of the problem and ordered

CF AVG	GOR	PHD	CF AVG	GOR	PHD	CF AVG	GOR	PHD
0.477	0.664	0.898	0.405	0.604	0.679	1	5	
0.467	0.840	0.679	0.449	0.664	0.790	2	6	
0.405	0.604	0.857	0.467	0.772	0.857	3		7
0.449	0.772	0.790	0.477	0.840	0.898	4	8	
								9
							10	
								11
								12
Sum of ranks						10	29	39

In this example total sample size N = 12, R_1 = 10, R_2 = 29, and R_3 = 39.Remember that the sum of the ranks will always equal n (n+1)/2. As a check in our assignment of ranks, we have n (n+1)/2 = 12(13)/2=78 which is equal to 10+29+39 = 78. The H statistic for this example is computed as follows:

$$H = \frac{12}{N(N+1)} * \sum_{i=1}^{j} \frac{R_i^2}{n_i} - 3(N+1) \qquad = \frac{12}{12(12+1)}\left[\frac{10^2}{2} + \frac{29^2}{2} + \frac{39^2}{2}\right] - 3(12+1)$$

H = 1/13*[25+210.25+380.25] – 39 = 615.5/13 -39 = 47.35 – 39 H = 8.35

Calculated value x^2 = 8.35 greater than tabulated value which 5.99 at the 0.5 % significance with 2 d.f (j-1=3-1=2). .Then we reject the H0, means that all three methods are not equal.

Fisher's Exact Test: Fisher's exact test is more accurate than the Chi-Square test ore when the expected numbers are small. This test is calculating the probability of the "R×C" table. Where R is the number of rows and C is the number of columns. Mostly 2x2 table use in Fisher's exact test. This test hypothesis of independence to a hyper geometric distribution of the numbers in the cells of the table.

Assumption:

- The binary data should be independent
- Out of any expected numbers are less than 5

	B		
A	B1	B2	Total
A1	A	B	a+b
A2	C	D	c+d
Total	a+c	b+d	a+b+c+d

Fisher's exact test the calculate the probability of getting any set of values was given by hyper geometric distribution formula: = $(a+b)!_{(c+d)}\ !\,(a+c)!_{(b+d)}\ !\,/n!\,a!\,b!\,c!\,d!$

Example: Doing a genetic study and studying the effect on which of two alleles for a gene a person has and the presence of a disease. We perform a genetic test to determine which allele the test subjects have and a disease test to determine whether the person has a disease. The data for a 2 x 2 contingency analysis should be entered in the format below, which apply the tests. The tests we want to perform with this contingency table are whether or not the two factors, disease and gene allele, are independent or whether there is a significant relationship between the factors.
H_0: The null hypothesis is that there is no relation and the factors are independent
H_1 : The null hypothesis is that there is relation and the factors are independent

Calculate the table data

	Disease	**Total**	
Gene	Yes	No	
Allele1	2	10	12
Allele2	12	3	15
Total	14	13	27

The above data we calculate the p = (a + b)!(c + d) ! (a + c)!(b + d)!/n!a!b!c = 12!15!14!13!/27!2!10!12!13
p = 0.0018

Based on the p value we can reject the hypothesis that the factors gene allele and disease are independent and conclude that there is a significant relation between the disease and which allele of the gene a person has

2. **SUMMARY**

Non-parametric methods have fewer assumptions than parametric tests so useful when these assumptions not met. The NP tests are often a good option for small sample sizes (n < 30).Non-parametric methods are a mixture of tests. Ordered statistics play very important role in the nonparametric test. These are the entire test is very useful the genetics study and microarray data analysis. In this paper define the idea and how calculate the nonparametric test with example of genetics data. Overall conclude of this paper is that nonparametric methods play very important role if the genetics data not follow the normality and these test can give the appropriate result of the hypothesis.

REFERENCES

1. Mount, D.W. (2004). Bioinformatics: Sequence and Genome Analysis. Second edition. Cold Spring Harbor Laboratory Press, New York, USA.
2. The International HapMap Consortium, 2003 the International HapMap Project. Nature 426: 789-94.
3. David, H. A.; Nagaraja, H. N. (2003). "Order Statistics". Wiley Series in Probability and Statistics. doi:10.1002/0471722162. ISBN 9780471722168.
4. Gentle, James E. (2009), Computational Statistics, Springer, p. 63, ISBN 9780387981444
5. Conover WJ. Practical Nonparametric Statistics, 2nd edition, New York: John Wiley and Sons.
6. Ikewelugo Cyprian Anaene Oyeka (Apr 2012). "Modified Wilcoxon's Signed-Rank Test". Open Journal of Statistics: 172–176.
7. Wilcoxon's, Frank (Dec 1945). "Inpidual comparisons by ranking methods" (PDF). Biometrics Bulletin 1 (6): 80–83
8. Siegel and Castellan. (1988). "Nonparametric Statistics for the Behavioral Sciences," 2nd edition, New York: McGraw-Hill.
9. 9.Kruskal; Wallis (1952). "Use of ranks in one-criterion variance analysis". Journal of the American Statistical Association 47 (260): 583–621. doi:10.1080/01621459.1952.10483441.
10. Corder, Gregory W.; Foreman, Dale I. (2009). Nonparametric Statistics for Non-Statisticians. Hoboken: John Wiley & Sons. pp. 99–105. ISBN 9780470454619.
11. Statistical Analysis of Gene Expression Microarray Data. T. P. Speed (Ed). Chapman & Hall. Collection of essays by microarray authorities.
12. The Analysis of Gene Expression Data. G. Parmigiani (Ed) et al. Springer. Covers various statistical tools for microarray analysis, including R.
13. Fisher, R. A. (1922). "On the interpretation of χ2 from contingency tables, and the calculation of P". Journal of the Royal Statistical Society 85 (1): 87–94. Doi: 10.2307/2340521. JSTOR 2340521.

About the Data Science Foundation

The Data Science Foundation is a professional body representing the interests of the Data Science Industry. Its membership consists of suppliers who offer a range of big data analytical and technical services and companies and individuals with an interest in the commercial advantages that can be gained from big data. The organisation aims to raise the profile of this developing industry, to educate people about the benefits of knowledge based decision making and to encourage firms to start using big data techniques.

Contact Data Science Foundation

Email:contact@datascience.foundation
Telephone: 0161 926 3641
Atlantic Business Centre
Atlantic Street
Altrincham
WA14 5NQ
web: www.datascience.foundation