Scientific Research Publishing

Open Journal of Statistics, 2020, 10, 154-162
https://www.scirp.org/journal/ojs
ISSN Online: 2161-7198
ISSN Print: 2161-718X

Using Confidence Statements to Ordering Medians: A Simple Microarray Nonparametric Analysis

Carlos A. de B. Pereira[1], Adriano Polpo[2]

[1]Institute of Mathematics and Statistics, University of São Paulo, Brazil
[2]Department of Mathematics and Statistics, The University of Western Australia, Perth, Western Australia, Australia
Email: cpereira@ime.usp.br, adriano.polpo@uwa.edu.au

How to cite this paper: de B. Pereira, C.A. and Polpo, A. (2020) Using Confidence Statements to Ordering Medians: A Simple Microarray Nonparametric Analysis. *Open Journal of Statistics*, **10**, 154-162.
https://doi.org/10.4236/ojs.2020.101012

Received: November 8, 2019
Accepted: February 25, 2020
Published: February 28, 2020

Abstract

Comparing two samples about corresponding parameters of their respective populations is an old and classical statistical problem. In this paper, we present a simple yet effective tool to compare two samples through their medians. We calculate the confidence of the statement "the median of the first population is strictly smaller (larger) than the median of the second." We analyze two real data sets and empirically demonstrate the quality of the confidence for such a statement. This confidence in the order of the medians is to be seen as a pre-analysis tool that can provide useful insights for comparing two or more populations. The method is entirely based on their exact distribution with no need for asymptotic considerations. We also provide the Quor statistical software, an R package that implements the ideas discussed in this work.

Keywords

Significance Test, Comparison of Two Samples, Confidence Interval Based on the Binomial Distribution

1. Introduction

This paper proposes an analysis that can be used as an aid for subsequent more complex statistical data analyses, like classification, clustering, logistic regression, etc. For more details see [1]. We discuss ideas to compare two independent groups and to evaluate a measure that indicates which group has smaller (larger) values than the other one. They are simple and effective without the need for sophisticated techniques. This work was motivated by the following example in

oncology: preoperative Gleason scores, in general, provide valuable prognoses for cases with prostate cancer. However, this is not verified for patients with a high score of Gleason-7. This group of patients is characterized by tumours displaying considerable morphological heterogeneity among affected regions. Microarray data have been collected to search for a gene set that could distinguish between recurrent (R) and non-recurrent (NR) Gleason-7 prostate cancer patients. A possible important gene that is associated with this disease is the *RPS*28 gene. In the study, there are two samples: the first sample has $m=5$ of R patients, and the second sample has $n=20$ of NR patients. Table 1 lists the microarray expression data for the 25 patients, and an illustration is given in Figure 1. As in many medical experiments, there are only a few cases in this study, and most of them are non-recurrent.

Suppose that the expression of a specific important gene is observed for each patient of the two independent samples, let the recurrent and non-recurrent cases, with inter-ordered samples (observations), be, respectively, $\left(x_{(1)}, x_{(2)}, \cdots, x_{(m)}\right)$ and $\left(y_{(1)}, y_{(2)}, \cdots, y_{(n)}\right)$; m and n are the sample sizes. The objective is to find genes that are under (or over) expressed, which is sometimes expressed by the statement that an expected microarray observation of an R case, x, is smaller (larger) than the expected observation of an NR case, y. In other words, it is conjectured that, for x and y being observations of random variables X and Y, one could expect, for under (over) expressed situations that the probability of $\{X<Y\}$ is larger (smaller) than a specified value, for example 0.8 (0.2). One of the statistical hypotheses that could indicate the validity of the conjecture is $M_X < M_Y$ (with M used to indicate median and the subscripts used to separate R and NR cases). Note that uppercase letters are used for random variables and parameters and lowercase letters for observations: probabilities refer to X and Y, and confidence refers to x and y.

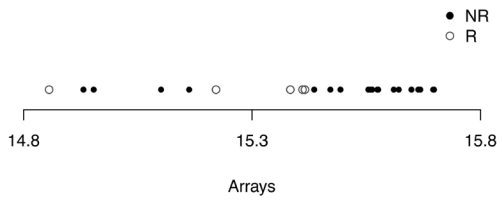

Figure 1. *RPS*28 Arrays for Gleason 7: Non-recurrent and Recurrent Cases.

Table 1. Expression of Gene *RPS*28 for Gleason-7 patients: Recurrent and Non-recurrent Cases.

Recurrent	14.8557	15.2209	15.3839	15.4106	15.4155
4*Non-recurrent	14.9309	14.9535	15.1009	15.1622	15.4361
	15.4716	15.4932	15.5545	15.5584	15.5622
	15.5629	15.5741	15.5759	15.6101	15.6211
	15.6488	15.6638	15.6684	15.6966	15.6984

We propose a measure to evaluate the confidence of the statement $\{M_X < M_Y\}$ (and obviously of $\{M_X > M_Y\}$ as well). We name this measure as *confidence statement*. The proposed confidence statement was developed following the ideas of the non-parametric confidence interval for a population's median based on the binomial distribution. The article is organized as follows: in Section 2, we give a brief review of the confidence interval for the population's median, and then we introduce the confidence statement; in Section 3, we analyze two real data examples, discussing the applicability of the procedure; and in Section 4, we provide conclusions and final remarks.

2. Methods

2.1. Confidence Intervals for Medians

In this section, we present the non-parametric confidence interval for a population's median based on the binomial distribution. For additional details we refer to [2] [3] [Chap. 7].

An event-related to a random variable X is represented by A, while M_X is the median of X. $\Pr(A \mid M_X)$ indicates the probability of the event A when M_X is known. In general, the median M_X of a random variable X is a population parameter that satisfies the following inequalities:

$$\Pr(X \le M_X \mid M_X) \ge \frac{1}{2} \text{ and } \Pr(X \ge M_X \mid M_X) \ge \frac{1}{2}. \quad (1)$$

In the continuous case, these inequalities are tight:

$$\Pr(X \le M_X \mid M_X) = \frac{1}{2} \text{ and } \Pr(X \ge M_X \mid M_X) = \frac{1}{2}. \quad (2)$$

Considering that $(X_1, X_2, \cdots, X_m)$ is a vector of m independent and identically distributed random variables, we have that $(1/2)^m$ is the probability of the event "all observations are smaller than M_X." Hence, the probability that at least $X_{(m)}$ (the sample maximum, the parenthesis in the subscript is used to indicate the order) is larger than M_X is the complementary probability $1-(1/2)^m$. Define $X_{(i)}$ as the *i*-th order statistics. One may consider the interval $(x_{(1)}, x_{(m)})$ as a confidence interval for the median M_X, for which the value of the confidence is obtained as follows: the probability that all observations are in one of the sides of M_X, right or left, should be $2(1/2)^m = (1/2)^{m-1}$. Again, taking the complement, one obtains the probability of the event $\{X_{(1)} \le M_X \le X_{(m)}\}$ as $1-(1/2)^{m-1}$.

After observing the sample, we write that the statement $\{x_{(1)} \le M_X \le x_{(m)}\}$ has a confidence equal to $1-(1/2)^{m-1}$. We call the attention of the reader to the subtle difference between probability and confidence, as presented in [4], which justifies the use of distinct terminology. To clarify, before the observations are obtained and by using the order statistics $X_{(1)}$ and $X_{(m)}$ (minimum and maximum), we write the following expression:

$$\Pr\left(X_{(1)} \le M_X \le X_{(m)}\right) = 1-\left(\frac{1}{2}\right)^{m-1}. \quad (3)$$

After observing the sample, $\left\{x_{(1)} < M_X < x_{(m)}\right\}$ is only a statement: we do not know the value of M_X but we know the sample values of all order statistics, $x_{(1)},\cdots,x_{(m)}$. It can be said that one has a confidence of $1-(1/2)^{m-1}$ that the median is within the sample extreme values: in this case, there are no probabilities any more. Using the sample of recurrent cases in Table 1, and as $1-(1/2)^4 = 0.9375$, we could say with confidence 93.75% that the interval $(14.856, 15.416)$ contains the population's median value. Also, as $1-(1/2)^5 = 0.96875$, we are confident that $\{M_X \le 15.416\}$, with confidence value 96.88%. To be more formal, prior to observations, we use the notation $\Pr\left(M_X \le X_{(5)} \mid M_X\right) = 0.96875$.

As an analogy, one can think of the above method as equivalent to tossing a coin m times, computing the probability of zero successes, which is $(1/2)^m$, and taking its complement, $1-(1/2)^m$. The same arguments can be used to obtain the probability of having two observations in one side and all the remaining on the other side of M_X. The event $\left\{X_{(m-1)} \ge M_X\right\}$ happens if neither $\left\{M_X > X_{(m)}\right\}$ nor $\left\{M_X > X_{(m-1)}\right\}$ occur. Conditional on M_X to be known, the probability of $\left\{M_X > X_{(m)} \text{ or } M_X > X_{(m-1)}\right\}$ is $(1/2)^m + 2(1/2)^m = 3(1/2)^m$. Hence, $\Pr\left(X_{(m-1)} \ge M_X \mid M_X\right) = 1-3(1/2)^m$. Consequently, the confidence of the interval $\left(x_{(2)}, x_{(m-1)}\right)$ is $1-3(1/2)^m$. For instance, considering $m=8$, we obtain the confidence values for the statements $\left\{x_{(m-1)} \ge M_X\right\}$ and $\left\{x_{(m-1)} \ge M_X \ge x_{(2)}\right\}$, which are equal to 0.96484375 and 0.9296875, respectively. Extending now for any order of statistics, we can think of the number of successes in m tosses of a fair coin.

Letting i and j be indices in the set $\{1,2,\cdots,m\}$, the events $\left\{X_{(i)} \le M_X\right\}$ and $\left\{X_{(j)} \ge M_X\right\}$ are those in which we are interested. For $i < j$ and by using the same arguments of the previous discussion, we have the following probabilities:

$$\Pr\left(X_{(i)} \le M_X \mid M_X\right) = \sum_{\ell=i}^{m} \binom{m}{\ell} \left(\frac{1}{2}\right)^m, \tag{4}$$

$$\Pr\left(X_{(j)} \ge M_X \mid M_X\right) = 1-\sum_{\ell=j}^{m} \binom{m}{\ell} \left(\frac{1}{2}\right)^m = \sum_{\ell=0}^{j-1} \binom{m}{\ell} \left(\frac{1}{2}\right)^m. \tag{5}$$

To obtain the confidence of the interval $\left(x_{(i)}, x_{(j)}\right)$, the same argument of tossing a fair coin is used. We then obtain the following:

$$\Pr\left(X_{(i)} \le M_X \le X_{(j)} \mid M_X\right) = \sum_{\ell=i}^{j-1} \binom{m}{\ell} \left(\frac{1}{2}\right)^{m-1} \tag{6}$$

For $m=15$, we have 0.982421875 and 0.96484375 as the confidence values for the statements $\left\{x_{(12)} \ge M_X\right\}$ and $\left\{x_{(4)} \le M_X \le x_{(12)}\right\}$, respectively.

To illustrate the confidence interval, we generate a sample with $n=10$ from a normal distribution with mean 0 and variance 1. The generated data is

$$\begin{aligned} \boldsymbol{x} = (&-1.6293, -0.7927, -0.5913, -0.3776, -0.2004, \\ &-0.1904, -0.0777, 0.0099, 0.1771, 2.5159). \end{aligned} \tag{7}$$

We are interested in the interval with 95% of confidence. Our procedure is

based in an exact discrete distribution, and it will not obtain an exact 95% standard level (or any other level) but a close one: the higher the sample size, the closer it will be. Our simulated data produce the intervals $(-1.6293, 0.0099)$ and $(-0.7927, 0.1771)$ with, respectively, 94.43% and 97.85% of confidence. Since the second, although with smaller amplitude, has larger confidence, we choose it as our confidence interval. From the data, we have that the mean ($\bar{x}$) is −0.1157 and the standard error ($se = sd/\sqrt{n}$) is 0.3370, where sd is the standard deviation. Using now the standard method of the confidence interval we obtain the 95.45% confidence interval as

$$\begin{aligned}(\bar{x} - 2se, \bar{x} + 2se) &= (-0.1157 - 0.6740, -01157 + 0.6740)\\ &= (-0.7897, 0.5583).\end{aligned} \tag{8}$$

The length of our 97.85% interval is 0.9698, smaller than 1.3480, which is the length of the standard one based on the t-student distribution, with 95.45% of confidence. Thus, we obtained a more confident shorter interval.

2.2. Confidence Statement on the Order of Medians

Returning to the problem of two samples that are used to compare two sub-populations, assume they are named case and control, the goal is to analyze the statement that the population median M_X of X is smaller (larger) than the population median M_Y of Y: one of the statements $\{M_X < M_Y\}$ or $\{M_X > M_Y\}$ is true. Recall that we use the notation $(x_{(1)}, x_{(2)}, \cdots, x_{(m)})$ and $(y_{(1)}, y_{(2)}, \cdots, y_{(n)})$ for the ordered sample vectors. In fact, we have independent samples of intra-sample independent and equally distributed observations.

Suppose that there are observations $x_{(i)}$ and $y_{(j)}$, such that $x_{(i)} < y_{(j)}$. We can write the following probabilities:

$$\Pr\left(X_{(i)} \geq M_X \mid M_X\right) = \sum_{\ell=0}^{i-1} \binom{m}{\ell} \left(\frac{1}{2}\right)^m, \tag{9}$$

$$\Pr\left(Y_{(j)} \leq M_Y \mid M_Y\right) = \sum_{\ell=j}^{n} \binom{n}{\ell} \left(\frac{1}{2}\right)^n, \tag{10}$$

and then for the joint probability one obtains

$$\Pr\left(M_X \leq X_{(i)} \,\&\, Y_{(j)} \leq M_Y \mid M_X, M_Y\right) = \left\{\sum_{\ell=0}^{i-1} \binom{m}{\ell} \left(\frac{1}{2}\right)^m\right\} \left\{\sum_{\ell=j}^{n} \binom{n}{\ell} \left(\frac{1}{2}\right)^n\right\}. \tag{11}$$

After observing that $x_{(i)} < y_{(j)}$ for the indices i and j, the confidence of the statement $\{M_X \leq x_{(i)} < y_{(j)} \leq M_Y\}$ is equal to the right side of the previous expression.

We point out that we are looking for the shortest interval with high confidence. Consequently, to evaluate the confidence of the statement $\{M_X < M_Y\}$, we should look for the best pair (i, j) such that $x_{(i)} < y_{(j)}$ that produces a high confidence and a high value of $\Pr\left(M_X \leq X_{(i)} \,\&\, Y_{(j)} \leq M_Y \mid M_X, M_Y\right)$. The consequence is that the statement $\{M_X < M_Y\}$ has a confidence equal to

$$\sup_{i,j:x_{(i)}<y_{(j)}} \Pr\left(M_X \le X_{(i)} \,\&\, Y_{(j)} \le M_Y \mid M_X, M_Y\right). \tag{12}$$

The closer we get to 1, the more confident we are about $M_X < M_Y$. Note that the probability is evaluated in the sample space of the random variables X and Y, given the constraints of $M_X \le X_{(i)}$, $Y_{(j)} \le M_Y$, and $x_{(i)} < y_{(j)}$, which implies the statement $\{M_X < M_Y\}$. Any probability is a number in the interval $(0,1)$. Values close to 1 have a higher chance to occur. However, we are not evaluating the probability of $\{M_X < M_Y\}$. The result comes from a probability of the sample space, and then instead of having a probability, we have confidence in the statement. This procedure is equal to any confidence interval procedure.

3. Examples

3.1. The Prostate Cancer

In the example shown in Table 1, the statement $\{M_X < M_Y\}$ has a confidence equal to

$$\left\{\sum_{\ell=0}^{4}\binom{5}{\ell}\left(\frac{1}{2}\right)^5\right\}\left\{\sum_{\ell=5}^{20}\binom{20}{\ell}\left(\frac{1}{2}\right)^{20}\right\} = 0.9688 \times 0.9941 = 0.9630. \tag{13}$$

This is a consequence of the fact that $15.416 = x_{(5)} < y_{(5)} = 15.436$ and that

$$\begin{aligned} &\Pr\left(M_X < X_{(5)} \mid M_X\right)\Pr\left(Y_{(5)} < M_Y \mid M_Y\right) \\ &= \Pr\left(M_X < X_{(5)} \,\&\, Y_{(5)} < M_Y \mid M_X, M_Y\right) = 0.963. \end{aligned} \tag{14}$$

In other words, we are 96.3% confident about the statement $\{M_X < M_Y\}$.

3.2. The Schizophrenia Data Set

The Schizophrenia data set is from the Altar A study of the Stanley Medical Research Institute's online genomics database (SMRIDB) [5], Higgs 2006 [6]. The data have $m = 32$ patients with schizophrenia and $n = 34$ individuals in the control group. 20,993 probe microarrays were reported. Our interest here is to find the most differentially expressed genes. For the analysis, we evaluate both statements $\{M_X < M_Y\}$ and $\{M_X > M_Y\}$, and keep the highest confidence in each case. Table 2 presents the 10 transcripts with the highest confidence and their respective statements.

3.3. Discussion

In the prostate cancer example, it must be noticed that by using the one side t-test one obtains a p-value of 7.24% (14.48% for the two-sided test). This is used to test $\boldsymbol{H}: \mu_X = \mu_Y$ versus $\boldsymbol{A}: \mu_X < \mu_Y$ ($\boldsymbol{A}: \mu_X \ne \mu_Y$ for the two-sided test). μ here is the notation for the mean, not for medians. Such a particular test has only asymptotic properties if the distributions of X and Y are not normal. On the other hand, the present paper proposes a method that does not use any distribution restriction, is exact and valid for any sample size.

Table 2. Schizophrenia data set: genes with the largest confidence.

Transcripts	Confidence	Status	Median Order*
215003	0.99609	Under	$M_S < M_C$
208581	0.99521	Over	$M_S > M_C$
212854	0.99200	Over	$M_S > M_C$
216336	0.98681	Over	$M_S > M_C$
212294	0.98681	Over	$M_S > M_C$
213626	0.98549	Over	$M_S > M_C$
209847	0.98549	Under	$M_S > M_C$
208399	0.98549	Under	$M_S < M_C$
204326	0.98549	Over	$M_S > M_C$
221011	0.98439	Under	$M_S < M_C$

*M_S: median for schizophrenic patients and M_C: median for control individuals. Under: For the specific transcript, the schizophrenic group is under expressed in comparison to the control individuals. Over: For the specific transcript, the schizophrenic group is over expressed in comparison to the control individuals.

The development of the present method builds on the studies from [7] [8]. The ideas of conditional statements came from [9]. Simplicity and lack of barriers were our main goals in building such a method. Without restrictions and by being simple, a method might not be able to be powerful. Some non-parametric methods, for example, in Noether 1991 [10], Wasserman 2006 [11], do not directly use all the ordered observations. They only use the order statistics $x_{(i)}$ and $y_{(j)}$ of each group.

By using the equivalence of confidence statements and significance testing DeGroot 1975 [12], one could, without great distress, state the significance of testing $\boldsymbol{H}: M_X < M_Y$ versus $\boldsymbol{A}: M_X > M_Y$, for the data in **Table 1**. We are prone to say that the significance favouring $\boldsymbol{H}$ against $\boldsymbol{A}$ could be 96.3%. Interchanging the hypotheses but keeping $\boldsymbol{A}$ as the null hypothesis, the exact *P*-value favouring A would then be 3.7%. That is, under the standard policy, we would reject the hypothesis of equality of medians, and we would expect gene *RPS*28 to be under-expressed for R patients when compared to the same gene in the NR group.

In the schizophrenia example, we analysed all 20993 genes to find those that were most differentially expressed. We found that among the 10 most differentially expressed transcripts, 4 were under, and 6 were over-expressed. Also, all confidence values were higher than 98%, which are good confidence levels in our opinion.

4. Conclusions

This work intends to provide a method that can be employed as a first-step procedure whenever a data set is to be analyzed. The authors believe that this method can be used to eliminate those variables that have no power to help in the

discovery of differentially expressed transcripts, before conducting other more complex/specialized procedures.

The method can be extended to more than two groups. In order to do that, the confidence level to detect a strict order has to be studied in more detail. The larger the number of sample groups, the smaller is the expected confidence. This is because the product of numbers belonging to the interval $(0,1)$ clearly produces numbers that are smaller than any of their factors, for instance, consider 3 random variables, X, Y and Z. The following inequality is obvious:

$$\begin{aligned}&\Pr\left(M_X \le X_{(a)} \mid M_X\right)\Pr\left(Y_{(b)} \le M_Y \le X_{(c)} \mid M_Y\right)\Pr\left(Z_{(d)} \le M_Z \mid M_Z\right)\\ &\le \min\left\{\Pr\left(M_X \le X_{(a)} \mid M_X\right), \Pr\left(Y_{(b)} \le M_Y \le X_{(c)} \mid M_Y\right), \Pr\left(Z_{(d)} \le M_Z \mid M_Z\right)\right\}.\end{aligned} \tag{15}$$

If the observed order of statistics follows the inequality $x_{(a)} < y_{(b)} < y_{(c)} < z_{(d)}$, (for orders a, b, c and d), then the statement $\{M_X < M_Y < M_Z\}$ would have smaller confidence than the confidences obtained when comparing a specific pair of the three medians. Hence, the confidence cut-off point to induce decisions would have to decrease with the increasing number of groups that are to be compared.

de Campos *et al.* [1] present a general theory that may include the statistical aspects of the present paper. Besides, one can find examples showing the superiority of our method compared with other classical solutions. Marques and Pereira, 2014 [13] can be viewed as a Bayesian non-parametric version of the present paper.

The procedure to evaluate the confidence statement is available in the R package Quor at https://code.google.com/archive/p/quor/. The package is distributed as an open-source program under GPLv3 license.

Acknowledgements

Carlos Alberto de Braganca Pereira is CNPq Fellow-Brazil (308776/2014-3).

Conflicts of Interest

The authors declare no conflicts of interest regarding the publication of this paper.

References

[1] de Campos, C., de Pereira, C.A.B., Rancoita, P. and Polpo, A. (2016) Ordering Quantiles through Confidence Statements. *Entropy*, **18**, 357. https://doi.org/10.3390/e18100357

[2] Thompson, W.R. (1936) On Confidence Ranges for the Median and Other Expectation Distributions for Populations of Unknown Distribution Form. *The Annals of Mathematical Statistics*, **7**, 122-128. https://doi.org/10.1214/aoms/1177732502

[3] David, H.A. and Nagaraja, H.N. (2003) Order Statistics. 3rd Edition, Wiley-Interscience, Hoboken. https://doi.org/10.1002/0471722162

[4] Pereira, C.A.D.B. and Castilho, E. (2009) RE: Should Meta-Analyses of Interventions Include Observational Studies in Addition to Randomized Controlled Trials?

A Critical Examination of Underlying Principles. *American Journal of Epidemiology*, **169**, 783. https://doi.org/10.1093/aje/kwp016

[5] The Stanley Medical Research Institute (2012) The Stanley Medical Research Institute Online Genomics Database. http://www.stanleygenomics.org

[6] Higgs, B., Elashoff, M., Richman, S. and Barci, B. (2006) An Online Database for Brain Disease Research. *BMC Genomics*, **7**, 70. https://doi.org/10.1186/1471-2164-7-70

[7] Zellner, A., Keuzenkamp, H. and McAleer, M. (2004) Simplicity, Inference and Modeling: Keeping It Sophisticatedly Simple. Cambridge University Press, Cambridge.

[8] Wasserman, L. (2010) All of Statistics. Springer, New York.

[9] Kiefer, J. (1977) Conditional Confidence Statements and Confidence Estimators. *Journal of American Statistical Association*, **72**, 789-808. https://doi.org/10.1080/01621459.1977.10479956

[10] Noether, G. (1991) Introduction to Statistics, The Nonparametric Way. Springer, New York. https://doi.org/10.1007/978-1-4612-0943-0

[11] Wasserman, L. (2006) All of Nonparametric Statistics. Springer, New York.

[12] DeGroot, M. (1975) Probability and Statistics. 2nd Edition, Addison-Wesley, New York.

[13] Marques, P.C. and de Pereira, C.A.B. (2014) Predictive Analysis of Microarray Data. *Open Journal of Genetics*, **4**, 63-68. https://doi.org/10.4236/ojgen.2014.41009