Non-parametric Tests

Shengping Yang PhD ,Gilbert Berdine MD

I was working on a small study recently to compare drug metabolite concentrations in the blood between two administration regimes. However, the metabolite concentrations for a few patients were so low that they could not be detected by the instrument I was using. I would like to know more about how to analyze data from such a study.

In some studies, the instrument used cannot provide precise measurements of the outcome of interest for some of the samples. In such cases, usually, a value, for example, "undetectable" is assigned to those samples. Statistically, analyzing these data is difficult using parametric methods, such as t test, ANOVA, without making major assumptions or censoring. For example, supposing that we assign two different arbitrary values (beyond the detectable threshold) to the non-detectable observations, we might get very different results because assigning different values to the non-detectables changes the mean and variance of the whole sample. As a simple and easy to implement alternative, a non-parametric method is usually recommended.

Non-parametric tests are also called ***distribution free*** statistics because they do not require that the data fit a known parameterized distribution such as normal. Since they require making fewer assumptions about the data, these tests are widely used in the analysis of many types of data, such as rank data, categorical data, as well as data with "non-detectable" values.

Analog to many of the parametric tests, there are a number of commonly used non-parametric tests for specific types of comparisons.

Corresponding author: Shengping Yang
Contact Information: Shengping.yang@ttuhsc.edu.
DOI: 10.12746/swrccc2014.0208.109

1. Mann-Whitney U Test (also Wilcoxon Rank Sum Test):

This test is commonly used for comparing the median of two independent groups of ordinal or rank data to determine if they are significantly different. It is the non-parametric equivalent of the widely used two-sample t test.

2. Kruskal-Wallis Test:

This test extends the Mann-Whitney U test to more than 2 groups, and it is the non-parametric equivalent of the Analysis of Variance (ANOVA).

3. Wilcoxon Signed-Rank Test:

This test compares two related samples, e.g., paired/matched, or repeated measures on the same samples, to make inferences as to whether the mean ranks of the two related populations differ. It is the non-parametric equivalent of the paired two-sample t test.

4. Friedman's Test:

This test is used to detect differences in treatments with repeated measures on the same samples. It is the non-parametric equivalent of the repeat measures ANOVA.

The principle of a non-parametric test is to make no assumptions about the distribution of the outcome variable, but to use the rank of the data for making statistical inferences. We will use the Mann-Whitney U test to explain how this works. The Mann-Whitney U test has two basic assumptions: the observations are independent of each other, and the data values

are ordinal – that is, one can compare any two data values and objectively state which is greater.

In the study mentioned above, the objective is to compare the drug metabolite concentrations in the blood between two administration regimes. The hypothetical data are presented below. The first row is the metabolite concentrations for patients who took the drug in capsule (n_C = 4), and the second row is the concentrations for patients who took the drug in tablet (n_T = 5). The total number of patients in this study is $N = n_T + n_C$ =9.

Capsule	0.59	0.31	1.22	0.52	
Tablet	0.11	Non-detectable*	0.31	0.05	0.53

* Detection threshold is 0.01*μg/l.*

Since one patient had non-detectable blood metabolite, the commonly used parametric test is not appropriate; we will apply a non-parametric test to this data. Note that patients who took the drug in capsules are independent (not paired/matched) of those who took the drug in tablets, thus a Mann-Whitney U test rather than a Wilcoxon Signed-Rank test should be used.

First, we define the null and alternative hypotheses of the Mann-Whitney test:

H_0*: There is no difference in the ranks of metabolite concentrations between the two regimes;*
H_A*: There is a difference in the ranks of metabolite concentrations between the two regimes.*

The null (H_0) hypothesis can be mathematically stated in two ways. The general meaning is that the probability of drawing larger values from the first population than the second population is equal to the probability of drawing larger values from the second population than the first population. A more strict expression of H_0 is that there is no significant difference between the median values for the ranked data in both populations.

To assign ranks to the data, we order the combined samples of the two administration regimes while keeping track of the two groups (Table below). In other words, the ranks are assigned to individual observations regardless which group they belong to; in the meantime, the grouping information is still kept. Note that when ties are present, we average the ranks. For example, the 4th and 5th ordered values are both 0.31, thus we assign the averaged rank of 4.5 to both of them.

Observations	**Capsule**				0.31	0.52		0.59	1.22
	Tablet	Non-detectable	0.05	0.11	0.31		0.53		

Ranks	**Capsule**				4.5*	6		8	9
	Tablet	1	2	3	4.5*		7		

* The 4th and 5th ordered values are both 0.31, the mean rank of 4.5 was assigned to both of them.

The next step is to calculate the U statistic. The distribution of U under the null hypothesis is known. Tables of this distribution for small samples are available. For samples larger than 20, the distribution is approximated to be normal. The calculation can be done manually or using a formula.

To manually determine U, pick the sample that seems to have the smaller values. The final result is independent of which group is chosen, but one group requires less effort. For our example, pick the Tablet data. For each Tablet data value, count how many Capsule data values are less than the Tablet data value. Add all these counts together. For our example, Non-detectable has 0 Capsule data values less than it, 0.05 has 0 Capsule data values less than it, 0.11 has 0 Capsule data values less than it, 0.31 has 0 Capsule data values less than it and 1 tie, and 0.53 has 2 Capsule data values less than it. Ties are

scored as 0.5. For our example:

$$U_T = 0 + 0 + 0 + 0.5 + 2 = 2.5.$$

If the Capsule data is used as the reference, one gets a different, but predetermined, result:

$$U_C = 3.5 + 4 + 5 + 5 = 17.5.$$

The sum ($U_T + U_C$) must equal the number of possible ways to compare n_T things against n_C things:

$$U_T + U_C = n_T \times n_C = 20.$$

The above algorithm can be automated by calculating the sum of the ranks for both the capsule and tablet groups separately. For the hypothetical data, the rank sums of the Capsule and Tablet groups are $R_T = 27.5$ (4.5+6+8+9) and $R_C = 17.5$ (1+2+3+4.5+7), respectively. Note that it is always a good practice to check whether the total sum of ranks (both groups included) equals to $N(N+1)/2$ to make sure that all the ranks are calculated correctly. In our calculation, we have $N = 9$ and thus $(N(N+1))/2 = 45$, which does equal to 27.5+17.5.

U is the minimum of U_T and U_C, which are calculated below for the Capsule and Tablet groups respectively. We let,

$$U_T = n_T\, n_C + \frac{n_T(n_T+1)}{2} - R_T = 20 + 10 - 27.5 = 2.5,$$

$$U_C = n_T\, n_C + \frac{n_C(n_C+1)}{2} - R_C = 20 + 15 - 17.5 = 17.5,$$

then

$$U = min\,(U_T, U_C) = 2.5$$

Note that in the formulas, the first term is the total number of comparison possibilities, the second term is total sum of the rank sums for both groups, and the *R* term is the rank sum for the chosen group.

The *U* value is converted to a significance or p value using the known distribution of *U* under the null hypothesis. For large samples, a normal approximation can be used:

We define,

$$z = \frac{(U - m_U)}{\sigma_U},$$

where $m_U = (n_T\, n_C)/2$ (the median value for U corresponding to a null assumption),

$$\sigma_U = \sqrt{\frac{n_T\, n_C}{N(N-1)} \times \frac{N^3 - N}{12} - \sum_{j=1}^{J} \frac{(t_j^3 - t_j)}{12}}$$

(the standard deviation for U).

Note that J is the number of groups of ties, and t_j is the number of tied ranks in group j. Also if there are no ties in the data, then the formula reduces to $\sigma_U = \sqrt{\frac{n_T\, n_C (N+1)}{12}}$ The value z is the difference between the observed comparisons vs. the median value (50% greater and 50% less) normalized to the standard deviation of the U statistic for the data. Tables (and computations) of p values from z values are readily available.

Apply this formula to the above example, we have,

$$Z = \frac{U - n_T n_C/2}{\sqrt{\frac{n_T\, n_C}{N(N-1)} \times \frac{N^3 - N}{12} - \sum_{j=1}^{J} \frac{(t_j^3 - t_j)}{12}}} = \frac{2.5 - 10}{4.0654} = -1.8448,$$

Since Z follows a standard normal distribution, the probability of observing a value equal to or more extreme than the observed, given the null hypothesis is true, is $2 \times P(Z \le -|z|)$ for a two sided test. In this example, the p value is $2 \times P(Z \le -|-1.8448|) = 0.0651$.

Since the p value is greater than 0.05, we do not reject H_0, and conclude that there is not sufficient evidence that the ranks of metabolite concentration

differ between the two regimes.

It may, at first glance, seem inappropriate to apply the mathematics of normal distributions to data that are known to not be normally distributed. This is the beauty of using rank methods to analyze the data. Any data point can be greater than, less than, or equal to the independent data point that it is being compared with. There are no other possibilities. Under the null hypothesis, the probability of a given data point having a greater value than the point it is being compared with must be equal to the probability of having a lesser value. The comparison is reduced to a coin flip, so the accumulated comparisons behave exactly as a random walk which does follow a normal distribution for large N.

Since U has a discrete distribution (U is derived from ranks, thus it can take only certain values) and Z follows a normal distribution, which is continuous (can take any value between -∞ and +∞), very often, an adjustment of continuity is performed to correct for the probability of using a continuous distribution to approximate a discrete distribution. In other words, the cumulative probability of a discrete random variable has jumps. To use a continuous distribution to approximate it, a correction is recommended to spread the probability uniformly over an interval, especially when the sample size is small. In this case, the z value after applying continuity correction is,

$$Z = \frac{U - Sign(U - m_U)/2 - n_T n_C / 2}{\sqrt{\frac{n_T n_C}{N(N-1)} \times \frac{N^3 - N}{12} - \sum_{j=1}^{J} \frac{t_j^3 - t_j}{12}}} = \frac{3 - 10}{4.7} = -1.7218,$$

and the corresponding p value for a two sided test is 0.0851.

A number of statistical software can be used to perform a Mann-Whitney U test. For example, the R code for the above Mann-Whitney test is:

Capsule = c (0.59, 0.31, 1.22, 0.52)
Tablet = c (0.11, 0.005, 0.31, 0.05, 0.53)*

Wilcox.test (Tablet, Capsule, correct=TRUE)
if continuity is to be adjusted; or
Wilcox.test (Tablet, Capsule, correct=FALSE)
if continuity is not to be adjusted.

The output from R is (with continuity correction),

Wilcoxon rank sum test with continuity correction

data: Tablet and Capsule
W = 2.5, p-value = 0.0851
alternative hypothesis: true location shift is not equal to 0

Note that the non-detectable observation was assigned a value 0.005, which is equal to the half of the lower detectable threshold*. In fact, assigning any value less than 0.01 would be acceptable since non-parametric test uses the rank of the data to make inferences, thus as long as the assigned value is less than the threshold, the result will be the same. On contrast, assigning different values to the non-detectable observations when using a parametric test can sometimes results in very different results.

The SAS code for a Mann-Whitney test is:

```
proc npar1way data=data Wilcoxon correct=yes;
*(use correct=no if continuity is not to be adjusted)
class regime;
var concentration;
run;
```

The output from SAS is (with continuity correction):

Wilcoxon Two-Sample Test	
Normal Approximation	
Z	1.7218
One-Sided Pr > Z	0.0425
Two-Sided Pr > |Z|	0.0851
Z includes a continuity correction of 0.5.	

In summary, a non-parametric test is a very useful tool for analyzing your data when the sample size is comparatively small and the distribution of the outcome is unknown and cannot be assumed to be approximately normal.

Author affiliation : Shengping Yang is a biostatistician in the Department of Pathology at TTHUSC. Gilbert Berdine is a pulmonary physician in the Department of Internal Medicine at TTUHSC.
Published electronically: 10/15/2014

References

1. Buckle N, Kraft C, van Eeden C. An Approximation to the Wilcoxon-Mann-Whitney Distribution. J Am Stat Assoc 1969; 64(326): 591-599.
2. Mann HB, Whitney DR. (1947). On a Test of Whether one of Two Random Variables is Stochastically Larger than the Other. Annals Math Stat 1947; 18(1): 50–60. doi:10.1214/aoms/1177730491.
3. The NPAR1WAY procedure. http://support.sas.com/documentation/cdl/en/statug/63033/HTML/default/viewer.htm#npar1way_toc.htm (last access: 9/23/2014)
4. Wilcoxon F. Individual comparisons by ranking methods. Biometrics Bulletin 1945; 1(6): 80–83. doi: 10.2307/3001968.
5. Wilcoxon Rank Sum and Signed Rank Tests. http://127.0.0.1:31961/library/stats/html/wilcox.test.html (last accessed: 9/23/2014)