Non-parametric tests for processing small samples

Arnaud BRINGÉ

Introduction

In standard cases where samples are sufficiently large (n>30), or the distribution corresponds closely to a normal distribution, Fisher and Student tests can be used to test for significant measurement differences between populations. However, these hypotheses require fairly strict rules of application for the data fit to the normal distribution.

Non-parametric methods, on the other hand, do not require many hypotheses about the population studied. In particular they do not assume the population is normally distributed. These tests are the best way of dealing with small samples, where the hypotheses for applying the usual tests do not hold. Below 30 observations it may be hazardous to approximate the sample to a normal distribution.

Tests are called non-parametric when no assumption is made about the distribution of the variables, and consequently no assumption about the underlying data distribution.

Non-parametric tests are based on the use of rank statistics (such as grades allocated or a ranking of individuals). They may be applied to quantitative characters, measurement magnitudes, or classification rankings.

A non-parametric test, then, is a hypothesis test for which the distribution curve of the population studied does not need to be specified.

However, the observations must be independent, i.e., the selection of an individual from the population to form a sample must not influence the selection of other individuals.

Non-parametric tests are only slightly less effective than parametric tests when the distribution of the population studied is specified, e.g., normal. According to Dodge (2007), they may be more effective when the population distribution deviates from normal.

We present below only a few non-parametric tests: Mann-Whitney and Wilcoxon for two-sample cases, Wilcoxon for matched samples, and Kruskal-Wallis for comparing more than two samples. A spreadsheet illustrates the calculation effected when using each of these tests. Our aim is not to be exhaustive and present a catalogue of these legacy tests, but to show an example of their use.

Mann-Whitney *U* test

Principle

Measurements of a variable are available from sample A of size n taken from population P_1, designated (X_1, X_2,…X_n) and from sample B of size p taken from population P_2, designated (Y_1, Y_2,…Y_p).

These samples are combined and ranked in ascending order of the values of this measurement.

For each element X_i of X, we count the total number of elements of Y higher than that element of X, plus half the number of times an element of Y is equal to X_i.

The Mann-Whitney U_{xy} statistic is the sum of this calculation for all elements of X. In the same way U_{yx} is calculated as the sum for all Y_i of the total number of X higher than that value, plus half the number of times they are equal. It can been shown that $U_{yx}=np-U_{xy}$. The values U_{xy} and U_{yx} are calculated for the total of ($n+p$) observations.

If the two groups are completely separate, then $U_{xy}=0$ and $U_{yx}=np$ (or $U_{yx}=0$ and $U_{xy}=np$ as the case may be). Conversely, if the groups are perfectly similar, $U_{xy}=U_{yx}=np/2$.

The smaller of the two statistics (U_{xy}, U_{yx}) is compared with a tabulated Mann-Whitney U. If $\min(U_{xy}, U_{yx})<U$, then the hypothesis of equal distributions for populations P_1 and P_2 will be rejected.

Remarks:

- If $m,n>12$, then the U distribution is approximated by a normal distribution of mean $\mu=\frac{m*n}{2}$ and standard deviation $\sigma=\sqrt{\frac{m*n*(m+n+1)}{12}}$. The statistic $\frac{U-\mu}{\sigma}$ is compared with the value read from the table of the reduced centred normal distribution.
- The values are read from a Mann-Whitney table, to be found at http://math.usask.ca/~laverty/S245/Tables/wmw.pdf. For the test the smaller of the two values U_{xy} and U_{yx} is used. The test is significant if this value is lower than the value in the table.

Wilcoxon test

Principle

The framework for using this test is the same as for the Mann-Whitney *U* test, namely two samples of size *n* and *p*. The two samples are combined in the same procedure.

Next each of these values in this vector of dimension (*n*+*p*) is ranked. If more than one observation present the same value for the measurement, an average rank is given. As in the Mann-Whitney *U* test, this procedure is applied for measurements or variables that are at least ordinal. Next each of these values in this vector of dimension (*n*+*p*) is ranked.

The ranking of observation X_i is denoted as $R(X_i)$ and the statistic $W = \sum_{i=1}^{n} R(X_i)$ is calculated.

It can be shown that there is a relation between the Wilcoxon *U* and Mann-Whitney *T* statistics (Dodge, 2007): $U = np + \frac{n(n+1)}{2} - T$

Wilcoxon test for matched samples

Principle

Here the framework is different. We have two samples, A and B, both of size *n*, representing two values for a measurement taken at two times t_1 and t_2 from the same individual.

First the difference between the two values for each individual is calculated, and then these values are ranked. The sum of rankings of positive differences is denoted *W*. *W* can then be compared with tabulated values (fixed risk) or a software package can be used to calculate a *p* value.

The null hypothesis is that there is no difference between the two groups and therefore the sum of positive rankings will not be significantly different from the sum of negative rankings. If the null hypothesis does not hold, this means that the sum of positive rankings is not equal to the sum of negative rankings and that there is a difference between the two groups compared. The Wilcoxon table provides the lower limit for *T*, the lowest total of positive or negative rankings.

Kruskal-Wallis

Principle

We possess measurements of k samples A_1 of size n_1, A_2 of size n_2, …A_k of size n_k and we seek to measure whether these measurements from the k samples are identical or if one measurement at least from a sample differs from the others.

Let the total size of k samples be *n*. Repeating the Wilcoxon test procedure, this time for *n* samples, the *n* samples are combined and each of the observations is ranked. As before, an average ranking is given if more than one observation has the same value.

As shown in Dodge (2007), if the number of average rankings is limited, the following statistic is calculated: $H = \left(\frac{12}{n(n+1)} \sum_{i=1}^{k} \frac{R_i^2}{n_i} \right) - 3(n+1)$

However, if the number of average rankings is high, statistic H is divided by a term $1 - \frac{\sum_{i=1}^{g} (t_i^3 - t_i)}{n^3 - n}$, where g is the number of groups containing average rankings and t_i the size of each group containing average rankings.

Tests carried out

(H_0): there is no significant difference in measurements between the two samples

(H_1): there is at least one sample in which the values of the measurement are significantly different from those in other samples.

If H is higher than the value $\chi^2_{k-1,1-\alpha}$ read from a table, then the null hypothesis will be rejected.

Example

In the spreadsheet attached, there are two examples of Kruskal-Wallis tests. The various values stated above are calculated, the correction for average rankings is systematically indicated even where it is not applicable in this case, since the number of groups with average rankings is fairly low. The first example does not establish a difference between the samples, and the second example rejects the null hypothesis and indicates that at least one sample differs from the others.

Citations

Dodge Y., *Dictionnaire statistique encyclopédique,* Springer, 2007.

Hollander M. and Wolfe A. D., *Nonparametric Statistical Methods.* New York: John Wiley & Sons, 1973.

Kruskal W.H. and Wallis W.A., Use of ranks in one criterion variance analysis, *Journal of American Statistical Association*, 47, 1952, pp. 583-621 et correction 48 p. 910.

Morice E., Quelques tests non paramétriques, *Revue de Statistique Appliquée*, RSA, tome 4, n°4, 1956, pp. 75-107.

Mann H.B., Whitney D.R., On a test whether one of two random variables is stochastically larger than the other, *Annals of mathematical statistics*, 18, 194, pp. 50-60.

Raison J., Les principaux tests non paramétriques. Quelques généralités et références biblographiques, *Revue de Statistique Appliquée*, RSA, tome 7, n°1, 1959, pp 83-106.

Wilcoxon F., Individual comparaisons by ranking methods, *Biometrics*, 1, pp. 80-84, 1945.