The Stata Journal

The *Stata Journal* publishes reviewed papers together with shorter notes or comments, regular columns, book reviews, and other material of interest to Stata users. Examples of the types of papers include 1) expository papers that link the use of Stata commands or programs to associated principles, such as those that will serve as tutorials for users first encountering a new field of statistics or a major new technique; 2) papers that go "beyond the Stata manual" in explaining key features or uses of Stata that are of interest to intermediate or advanced users of Stata; 3) papers that discuss new commands or Stata programs of interest either to a wide spectrum of users (e.g., in data management or graphics) or to some large segment of Stata users (e.g., in survey statistics, survival analysis, panel analysis, or limited dependent variable modeling); 4) papers analyzing the statistical properties of new or existing estimators and tests in Stata; 5) papers that could be of interest or usefulness to researchers, especially in fields that are of practical importance but are not often included in texts or other journals, such as the use of Stata in managing datasets, especially large datasets, with advice from hard-won experience; and 6) papers of interest to those who teach, including Stata with topics such as extended examples of techniques and interpretation of results, simulations of statistical concepts, and overviews of subject areas.

For more information on the *Stata Journal*, including information for authors, see the webpage

http://www.stata-journal.com

The *Stata Journal* is indexed and abstracted in the following:

- CompuMath Citation Index®
- Current Contents/Social and Behavioral Sciences®
- RePEc: Research Papers in Economics
- Science Citation Index Expanded (also known as SciSearch®)
- Scopus™
- Social Sciences Citation Index®

Users of any of the software, ideas, data, or other materials published in the *Stata Journal* or the supporting files understand that such use is made without warranty of any kind, by either the *Stata Journal*, the author, or StataCorp. In particular, there is no warranty of fitness of purpose or merchantability, nor for special, incidental, or consequential damages such as loss of profits. The purpose of the *Stata Journal* is to promote free communication among Stata users.

The *Stata Journal*, electronic version (ISSN 1536-8734) is a publication of Stata Press. Stata, STATA, Stata Press, Mata, MATA, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2012)
12, Number 2, pp. 182–190

What hypotheses do "nonparametric" two-group tests actually test?

Ronán M. Conroy
Royal College of Surgeons in Ireland
Dublin, Ireland
rconroy@rcsi.ie

Abstract. In this article, I discuss measures of effect size for two-group comparisons where data are not appropriately analyzed by least-squares methods. The Mann–Whitney test calculates a statistic that is a very useful measure of effect size, particularly suited to situations in which differences are measured on scales that either are ordinal or use arbitrary scale units. Both the difference in medians and the median difference between groups are also useful measures of effect size.

Keywords: st0253, ranksum, Wilcoxon rank-sum test, Mann–Whitney statistic, Hodges–Lehman median shift, effect size, qreg

1 Introduction

It is a common fallacy that the Mann–Whitney test, more properly known as the Wilcoxon rank-sum test and also known as the Mann–Whitney–Wilcoxon test, is a test for equality of medians. Many of its users are probably unaware that the test calculates a useful parameter (and therefore should not be called "nonparametric") that is often of more practical interest than the difference between two means.

I will use an extreme case to illustrate the tests available to compare two groups and, in particular, the procedures that examine differences in medians.

I will use a dataset that is deliberately constructed so that the medians of two groups are equal but with distributions skewed in opposite directions. Although this is an extreme case, you should bear in mind that differences between two groups in the shape of underlying distributions will have consequences in the same direction, albeit smaller than the ones illustrated here.

The dataset is as follows:

```
. list

     +---------------+
     | group   value |
     |---------------|
  1. |     0       5 |
  2. |     0       5 |
  3. |     0       5 |
  4. |     0       5 |
  5. |     0       5 |
     |---------------|
  6. |     0       5 |
  7. |     0       7 |
  8. |     0       8 |
  9. |     0       9 |
 10. |     0      10 |
     |---------------|
 11. |     1       1 |
 12. |     1       2 |
 13. |     1       3 |
 14. |     1       4 |
 15. |     1       5 |
     |---------------|
 16. |     1       5 |
 17. |     1       5 |
 18. |     1       5 |
 19. |     1       5 |
 20. |     1       5 |
     +---------------+
```

2 The Mann–Whitney test

Both groups have a median of 5, but group 0 has no values less than 5 and group 1 has no values greater than 5. We can confirm that the medians are the same by using the `table` command:

```
. table group, contents(p50 value)

----------------------
    group | med(value)
----------+-----------
        0 |          5
        1 |          5
----------------------
```

Next we run the Wilcoxon rank-sum test (Mann–Whitney test):

```
. ranksum value, by(group)

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

       group |      obs    rank sum    expected
-------------+---------------------------------
           0 |       10         137         105
           1 |       10          73         105
-------------+---------------------------------
    combined |       20         210         210

unadjusted variance      175.00
adjustment for ties      -37.63
                     ----------
adjusted variance        137.37

Ho: value(group==0) = value(group==1)
             z =   2.730
    Prob > |z| =   0.0063
```

The test gives a highly significant difference between the two groups. Clearly, the test cannot be testing the hypothesis of equal medians, so what hypothesis does it test? We can see the answer by adding the `porder` option.

```
. ranksum value, by(group) porder

Two-sample Wilcoxon rank-sum (Mann-Whitney) test

       group |      obs    rank sum    expected
-------------+---------------------------------
           0 |       10         137         105
           1 |       10          73         105
-------------+---------------------------------
    combined |       20         210         210

unadjusted variance      175.00
adjustment for ties      -37.63
                     ----------
adjusted variance        137.37

Ho: value(group==0) = value(group==1)
             z =   2.730
    Prob > |z| =   0.0063

P{value(group==0) > value(group==1)} = 0.820
```

3 The Mann–Whitney statistic: A useful measure of effect size

The last line of the output states that the probability of an observation in group 0 having a true value that is higher than an observation in group 1 is 82%. In reality, the limitations of measurement scales will often produce cases where the two values are tied. So the parameter is calculated on the basis of the percentage of cases in which a random observation from group 0 is higher than a random observation from group 1, plus half the probability that the values are tied (on the rationale that if the values are tied, the true value is greater in group 1 in half the randomly selected pairs and greater in group 2 in the other half of them).

This parameter forms the basis of the Mann–Whitney test, a parameter that is a very useful measure of effect size in many situations. A researcher will frequently be faced with a measurement scale that either is not interval in nature (such as a five-point ordered scale) or has no naturally defined underlying measurement units. Typical examples of the latter are scales to measure moods, attitudes, aptitudes, and quality of life. In such cases, presenting mean differences between groups is uninformative. The Mann–Whitney statistic, on the other hand, is highly informative. It tells us the likelihood that a member of one group will score higher than a member of the other group (with the caveat above about the interpretation of tied values). In the analysis of controlled treatment trials, this measure is equivalent to the probability that a person assigned to the treatment group will have a better outcome than a person assigned to the control group. Using this can overcome the problem of many outcome scales used in assessing treatments being measured in arbitrary units.

The statistic has a long history of being rediscovered and, consequently, goes under a variety of names. The history dates back to the original articles in which Wilcoxon (1945, 1950) described the test. He failed to give the test a name and failed to specify the hypothesis it tested. The importance of the article by Mann and Whitney (1947) is that they made the hypothesis explicit: their article is entitled "On a test of whether one of two random variables is stochastically larger than the other". However, the statistic they actually calculated in the article, U, is not the probability that one variable is larger than the other. Birnbaum (1956) pointed out that transforming U by dividing it by its maximum value resulted in a useful measure, but he failed to name the measure. In 1976, Herrnstein proposed the transformation, but the article (Herrnstein, Loveland, and Cable 1976) appeared in an animal psychology journal, and his proposal to assign the letter rho to the transformed value was extremely unoriginal.

Perhaps as a result, the literature is now replete with statistics that are nothing other than Mann–Whitney statistics under other names. Bross (1958), setting out the calculation and use of ridit statistics, noted that the mean ridit score for a group was the probability that an observation from that group would be higher than an observation from a reference population. Harrell's C statistic, which is a measure of the difference between two survival distributions, is a special case of the Mann–Whitney statistic, and indeed, in the absence of censored data, it reduces to the Mann–Whitney statistic (Koziol and Jia 2009). Likewise, the tendency to refer to the Mann–Whitney statistic as the area under the receiver operator characteristic curve is common in literature evaluating diagnostic and screening tests in medicine, and is extremely unhelpful. The name entirely obscures what the test actually tells us, which is the probability that a person with the disorder or condition will score higher on the test than a person without it. The area under the receiver operator characteristic curve has been proposed as a measure of effect size in clinical trials (Brumback, Pepe, and Alonzo 2006), which would extend the bafflement to a new population of readers.

As a measure of effect size, the Mann–Whitney statistic has been renamed not once, but repeatedly, and with willful obstinacy. McGraw and Wong (1992) proposed what they called "a common language effect size" that was none other than the Mann–Whitney statistic. The measure was generalized by Vargha and Delaney (2000), who in the process renamed it the measure of stochastic superiority. It has since been further generalized by Ruscio (2008), who renamed the statistic A. He points out that it is insensitive to base rates and more robust to several other factors (for example, extreme scores and nonlinear transformations), in addition to its excellent generalizability across contexts. Finally, Acion et al. (2006) rebranded it the "Probabilistic index"; they advocated it as an intuitive nonparametric approach to measuring the size of treatment effects.

Stata users can use the `ranksum` command to calculate the Mann–Whitney statistic. More usefully, Newson's (1998) package `somersd` provides confidence intervals. Newson named the statistic "Harrell's C".

```
. somersd group value, transf(c) tdist
Somers´ D with variable: group
Transformation: Harrell´s c
Valid observations: 20
Degrees of freedom: 19

Symmetric 95% CI for Harrell´s c
------------------------------------------------------------------------------
             |              Jackknife
       group |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       value |        .18    .0711805     2.53   0.020     .0310175    .3289825
------------------------------------------------------------------------------
```

Note that the output of `somersd` will have to be manipulated in this case to provide the statistic comparing the higher and lower groups. It reports the probability that an observation in the first group will be higher than an observation in the second group. A simple way to overcome this is to reverse the group codes by using Cox's (2003) `vreverse`, one of several user-written commands to reverse variables (like all user-written commands, it may be located and installed within Stata by using the `findit` command).

```
. vreverse group, generate(r_group)
note: group not labeled

. somersd r_group value, transf(c) tdist
Somers´ D with variable: r_group
Transformation: Harrell´s c
Valid observations: 20
Degrees of freedom: 19

Symmetric 95% CI for Harrell´s c
------------------------------------------------------------------------------
             |              Jackknife
     r_group |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       value |        .82    .0711805    11.52   0.000     .6710175    .9689825
------------------------------------------------------------------------------
```

This accords with the earlier output from `ranksum` but has the added advantage of presenting us with a useful confidence interval.

4 How and when to test medians

The Mann–Whitney test is a test for equality of medians only under the very strong assumption that both of the two distributions are symmetrical about their respective medians or, in the case of asymmetric distributions, that the distributions are of the same shape but differ in location. Thus the common belief that the test compares medians is true only under some implausible circumstances.

Nevertheless, there are times when a researcher explicitly wishes to test for a difference in medians. For example, mean length of hospital stay is generally skewed and often badly affected by small numbers of very long admissions, so the 50th percentile of the stay distribution may be of more interest than the mean. A researcher might therefore be tempted to test for equality of medians with the median test. However, once again, we need to be cautious:

```
. median value, by(group)

Median test

   Greater |
  than the |         group
    median |         0          1 |     Total
-----------+----------------------+----------
        no |         6         10 |        16
       yes |         4          0 |         4
-----------+----------------------+----------
     Total |        10         10 |        20

          Pearson chi2(1) =   5.0000   Pr = 0.025

   Continuity corrected:
          Pearson chi2(1) =   2.8125   Pr = 0.094
```

The median test does not actually test for equality of medians. Instead, it tests a *likely consequence* of drawing two samples from populations with equal medians: that a similar proportion of observations in each group will be above and below the grand median of the data. And as we can see, the test provides at least some support for the idea that the medians are different.

On the other hand, quantile regression does test for the equality of medians. It is the direct equivalent of the t test for medians, because the mean minimizes squared error, whereas the median minimizes absolute error:

```
. qreg value group
Iteration  1:  WLS sum of weighted deviations =  27.481539

Iteration  1: sum of abs. weighted deviations =         30
Iteration  2: sum of abs. weighted deviations =         26
Iteration  3: sum of abs. weighted deviations =         24

Median regression                                    Number of obs =         20
  Raw sum of deviations       24 (about 5)
  Min sum of deviations       24                     Pseudo R2     =     0.0000

------------------------------------------------------------------------------
       value |      Coef.   Std. Err.      t    P>|t|     [95% Conf. Interval]
-------------+----------------------------------------------------------------
       group |          0   1.221911     0.00   1.000     -2.56714     2.56714
       _cons |          5   .8640215     5.79   0.000     3.184758    6.815242
------------------------------------------------------------------------------
```

This accords with what we know of the data: the medians are identical and the difference in the medians is 0. Quantile regression is also more powerful in that it can be extended to other quantiles of interest and it can be adjusted for covariates.

Note that the difference in medians between two *groups* does not correspond to the median difference between *individuals*. To confirm this, we can use Newson's (2006) **cendif** command:

```
. cendif value, by(group)
Y-variable: value
Grouped by: group
Group numbers:
      group |      Freq.     Percent        Cum.
------------+-----------------------------------
          0 |         10       50.00       50.00
          1 |         10       50.00      100.00
------------+-----------------------------------
      Total |         20      100.00
Transformation: Fisher´s z
95% confidence interval(s) for percentile difference(s)
between values of value in first and second groups:
   Percent    Pctl_Dif     Minimum     Maximum
        50           2           0           4
```

The median difference between members of group 0 and members of group 1 is 2, with a 95% confidence interval from 0 to 4. This measure, usually called the Hodges–Lehman median difference (Hodges and Lehmann 1963) is, as Newson (2002) points out, a special case of Theil's (1950a; 1950b; 1950c) median slope.

The researcher therefore has to relate the statistical procedure to the purpose of the study: computing the difference in medians tests for a difference between two conditions. In the case of hospital stay, the effect of a new treatment on length of stay could be tested by comparing the median stay in two groups, one of whom received the new treatment and the other acting as control. However, when the interest is the impact on the individual, the difference in medians between groups is misleading. Here the expected effect of treatment on the individual is best measured by the median difference between patients in the treatment and control groups.

5 Conclusion

The Mann–Whitney test is based on a parameter that is of considerable interest as a measure of effect size, especially in situations in which outcomes are measured on scales that either are ordinal or have arbitrary measurement units. Unfortunately, it has been often misrepresented as a test for the equality of medians, which it is not. The Mann–Whitney statistic also has been subject to so many rediscoveries and rebrandings over the years.

Those who wish to test the equality of medians between two groups should avoid the Mann–Whitney test and should consider `qreg` as a more powerful and versatile alternative to the median test. It is important, however, to distinguish between the difference between the medians of two groups, which measures the effect of a policy or condition on the location of the distribution of the outcome of interest, and the median difference between individuals in the two groups (Hodges–Lehman median difference), which is a measure of the expected benefit to the individual associated with being a member of the superior group.

6 References

Acion, L., J. J. Peterson, S. Temple, and S. Arndt. 2006. Probabilistic index: An intuitive non-parametric approach to measuring the size of treatment effects. *Statistics in Medicine* 25: 591–602.

Birnbaum, Z. W. 1956. On a use of the Mann-Whitney statistic. In *Proceedings of the Third Berkeley Symposium on Mathematical Statistics and Probability*, ed. J. Neyman, 13–17. Berkeley, CA: University of California Press.

Bross, I. D. J. 1958. How to use ridit analysis. *Biometrics* 14: 18–38.

Brumback, L. C., M. S. Pepe, and T. A. Alonzo. 2006. Using the ROC curve for gauging treatment effect in clinical trials. *Statistics in Medicine* 25: 575–590.

Cox, N. J. 2003. vreverse: Stata module to reverse existing categorical variable. Statistical Software Components S434402, Department of Economics, Boston College. http://ideas.repec.org/c/boc/bocode/s434402.html.

Herrnstein, R. J., D. H. Loveland, and C. Cable. 1976. Natural concepts in pigeons. *Journal of Experimental Psychology: Animal Behavior Processes* 2: 285–302.

Hodges, J. L., Jr., and E. L. Lehmann. 1963. Estimates of location based on rank tests. *Annals of Mathematical Statistics* 34: 598–611.

Koziol, J. A., and Z. Jia. 2009. The concordance index C and the Mann-Whitney parameter Pr(X>Y) with randomly censored data. *Biometrical Journal* 51: 467–474.

Mann, H. B., and D. R. Whitney. 1947. On a test of whether one of two random variables is stochastically larger than the other. *Annals of Mathematical Statistics* 18: 50–60.

McGraw, K. O., and S. P. Wong. 1992. A common language effect size statistic. *Psychological Bulletin* 111: 361–365.

Newson, R. 1998. somersd: Stata module to calculate Kendall's tau-a, Somers' D and median differences. Statistical Software Components S336401, Department of Economics, Boston College. http://ideas.repec.org/c/boc/bocode/s336401.html.

———. 2002. Parameters behind "nonparametric" statistics: Kendall's tau, Somers' D and median differences. *Stata Journal* 2: 45–64.

———. 2006. Confidence intervals for rank statistics: Percentile slopes, differences, and ratios. *Stata Journal* 6: 497–520.

Ruscio, J. 2008. A probability-based measure of effect size: Robustness to base rates and other factors. *Psychological Methods* 13: 19–30.

Theil, H. 1950a. A rank invariant method of linear and polynomial regression analysis, I. *Proceedings of the Koninklijke Nederlandse Akademie Wetenschappen, Series A – Mathematical Sciences* 53: 386–392.

———. 1950b. A rank invariant method of linear and polynomial regression analysis, II. *Proceedings of the Koninklijke Nederlandse Akademie Wetenschappen, Series A – Mathematical Sciences* 53: 521–525.

———. 1950c. A rank invariant method of linear and polynomial regression analysis, III. *Proceedings of the Koninklijke Nederlandse Akademie Wetenschappen, Series A – Mathematical Sciences* 53: 1397–1412.

Vargha, A., and H. D. Delaney. 2000. A critique and improvement of the CL common language effect size statistics of McGraw and Wong. *Journal of Educational and Behavioral Statistics* 25: 101–132.

Wilcoxon, F. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin* 1: 80–83.

———. 1950. Some rapid approximate statistical procedures. *Annals of the New York Academy of Sciences* 52: 808–814.

About the author

Ronán Conroy is a biostatistician at the Royal College of Surgeons in Dublin, Ireland.