The Stata Journal (2015)
15, Number 1, pp. 292–300

Nonparametric pairwise multiple comparisons in independent groups using Dunn's test

Alexis Dinno
School of Community Health
Portland State University
Portland, OR
alexis.dinno@pdx.edu

Abstract. Dunn's test is the appropriate nonparametric pairwise multiple-comparison procedure when a Kruskal–Wallis test is rejected, and it is now implemented for Stata in the `dunntest` command. `dunntest` produces multiple comparisons following a Kruskal–Wallis k-way test by using Stata's built-in `kwallis` command. It includes options to control the familywise error rate by using Dunn's proposed Bonferroni adjustment, the Šidák adjustment, the Holm stepwise adjustment, or the Holm–Šidák stepwise adjustment. There is also an option to control the false discovery rate using the Benjamini–Hochberg stepwise adjustment.

Keywords: st0381, dunntest, kwallis, Dunn's test, Kruskal–Wallis test, multiple comparisons, familywise error rate, Bonferroni, Šidák, Holm, Holm–Šidák, false discovery rate

1 Introduction

One-way omnibus tests, such as the common one-way analysis of variance (ANOVA), typically pose null hypotheses that measurements across some number of groups are all derived from a common distribution. One might think of such tests answering generic questions like, Does one need to bother looking more closely between groups for differences? Without evidence to reject the null hypothesis of such tests, one's work moves on to new topics. On the other hand, if the null hypothesis of an omnibus test is rejected, the question becomes, Which of these groups is different from which? If one used an ANOVA to test for mean difference, upon rejection of the null hypothesis, one would make multiple pairwise comparisons using t tests for mean difference in unpaired data. However, the ANOVA has restrictive assumptions concerning the distributions of the groups under scrutiny: the groups must have equal variances, and the measures in each group must be continuous, normally distributed variables.

The nonparametric Kruskal–Wallis test (Kruskal and Wallis 1952) is a nonparametric analog to the one-way ANOVA that sacrifices the precision of discriminating means for the discrimination of stochastic dominance (that is, the probability that a randomly drawn observation from one group will be greater than a randomly drawn observation from another). However, the test can do so regardless of how the measures are distributed in each group. If one assumes that the measures are continuous and that the unspecified distributions in each group differ only in their centrality, then one can under-

stand the Kruskal–Wallis test as an omnibus test for median difference. Upon rejection of the null hypothesis of this test, one would conduct multiple pairwise comparisons for stochastic dominance or median difference.

It is clear that the appropriate test for such comparisons is a nonparametric analog to the t test—the rank-sum test (Wilcoxon 1945; Mann and Whitney 1947) is an example—but the application is not as straightforward. One can use ANOVA's strict assumption about equal variances with the t tests that follow rejection of the null hypothesis of an ANOVA by using the pooled estimate of variance when calculating the standard error of the t test statistics. If one used a Kruskal–Wallis test, one would ignore this assumption, which is important when interpreting the median difference but more important when interpreting the rank-sum test as part of a family of inferences in the omnibus hypothesis. The ranks of the data on which the tests are based change if they are reranked in a pairwise fashion. Dunn's (1964) insight was to retain the rank sums from the omnibus test and to approximate a z-test statistic to the exact rank-sum statistic. Dunn's test is the appropriate procedure following a Kruskal–Wallis test.

Making multiple pairwise comparisons following an omnibus test redefines the meaning of α, which usually represents the probability of falsely rejecting the null hypothesis for one test, within the inferential framework of the hypothesis test. Dunn (1961) described how to address this issue with a Bonferroni adjustment, which can modify the rejection level for any test by dividing α by the total number of tests and requires a much smaller p-value to reject any test. This adjustment leaves α numerically intact but multiplies the p-value. This forms the basis of the familywise error rate (FWER) redefinition of α to signify the probability of falsely rejecting the null hypothesis in one test out of all tests performed. The Bonferroni adjustment introduced the FWER, but additional improvements followed: the Šidák (1967) adjustment, which is a slightly more powerful yet similar approach; Holm's sequential adjustment; and the Holm–Šidák (1979) sequential adjustment, sometimes credited to Holland and Copenhaver (1988), which treats subsequent pairwise hypothesis tests as parts of different families on the basis of whether previous tests were rejected. Finally, Benjamini and Hochberg (1995) reasoned that α should be interpreted as a desired false discovery rate (FDR) and should reflect how the expected rate of false discoveries changes after some pairwise tests are rejected in sequence.

Dunn's (1964) test has grown in popularity over the past two decades (figure 1).[1] The test is frequently used with multiple-comparison adjustments. During the past two decades, out of 1,097 cited articles, 778 included the term "Bonferroni", 11 included the term "Sidak" and excluded the term "Holm–Sidak", 111 included the term "Holm" and excluded the term "Holm–Sidak", 183 included the term "Holm–Sidak" (none included "Holland" and "Copenhaver"), and 14 included the terms "Benjamini" and "Hochberg". Dunn's increasingly used test is now implemented for Stata.

1. Data from a search on 6 March 2014 of Google Scholar for citations with the exact phrase "Dunn's test" for each of the years 1994–2013.

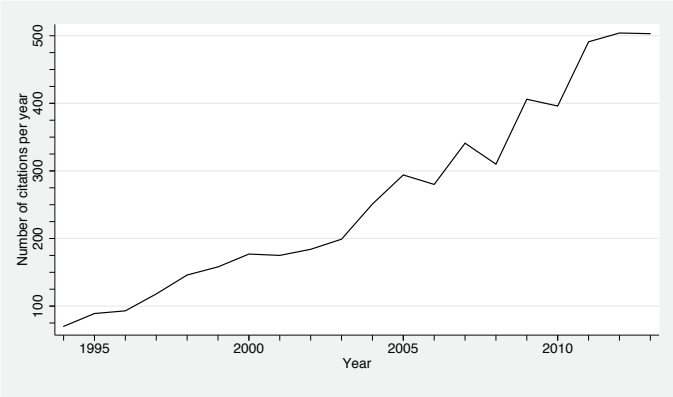

Figure 1. Citations indexed by Google Scholar including "Dunn's test" over two decades

2 The dunntest command

2.1 Syntax

`dunntest` *varname* [*if*] [*in*], `by(`*groupvar*`)` [`ma(`*method*`) nokwallis nolabel wrap level(`#`)`]

2.2 Description

`dunntest` reports the results of Dunn's (1964) test for stochastic dominance among multiple pairwise comparisons following a Kruskal–Wallis test of stochastic dominance among k groups Kruskal and Wallis (1952) using `kwallis` (see [R] **kwallis**). `dunntest` performs $m = k(k-1)/2$ multiple pairwise comparisons using z-test statistics. The null hypothesis in each pairwise comparison is that the probability of observing a random value in the first group that is larger than a random value in the second group equals one half; this null hypothesis corresponds to that of the Wilcoxon–Mann–Whitney rank-sum test (see [R] **ranksum**, and note that the `porder` option provides an explicit estimate of this probability). As in the rank-sum test, if the data are assumed to be continuous and the distributions are assumed to be identical except for a shift in centrality, Dunn's (1964) test may be understood as a test for median difference. In the syntax diagram above, *varname* refers to the variable recording the outcome, and *groupvar* refers to the variable denoting the population. `dunntest` accounts for tied ranks. `by()` is required.

2.3 Options

`by(`*groupvar*`)` specifies a variable that identifies the groups. `by()` is required.

`ma(`*method*`)` specifies the method of adjustment used for multiple comparisons and takes one of the following values: `none`, `bonferroni`, `sidak`, `holm`, `hs`, or `bh`. The default is `ma(none)`. These methods perform as follows:

`none` specifies that no adjustment for multiple comparisons be made.

`bonferroni` specifies a Bonferroni adjustment where the FWER is adjusted by multiplying the p-values in each pairwise test by m (the total number of pairwise tests). Stata will report a maximum Bonferroni-adjusted p-value of 1.

`sidak` specifies a Šidák adjustment where the FWER is adjusted by replacing the p-value of each pairwise test with $1-(1-p)^m$ according to Šidák (1967). Stata will report a maximum Šidák-adjusted p-value of 1.

`holm` specifies a Holm adjustment where the FWER is adjusted sequentially by adjusting the p-values of each pairwise test, ordered from smallest to largest, with $p(m+1-i)$, where i is the position in the ordering according to Holm (1979). Stata will report a maximum Holm–adjusted p-value of 1. Because the decision to reject the null hypothesis in sequential tests depends both on the p-values and how they are ordered, the comparisons rejected by this method at the alpha level (two-sided test) are underlined in the output.

`hs` specifies a Holm–Šidák adjustment where the FWER is adjusted sequentially by adjusting the p-values of each pairwise test, ordered from smallest to largest, with $1-(1-p)^{m+1-i}$, where i is the position in the ordering according to Holm (1979). Stata will report a maximum Holm–Šidák-adjusted p-value of 1. Because the decision to reject the null hypothesis in sequential tests depends both on the p-values and how they are ordered, the comparisons rejected by this method at the alpha level (two-sided tests) are underlined in the output.

`bh` specifies a Benjamini–Hochberg adjustment where the FDR is adjusted sequentially by adjusting the p-values of each pairwise test, ordered from largest to smallest, with $p\{m/(m+1-i)\}$, where i is the position in the ordering according to Benjamini and Hochberg (1995). Stata will report a maximum Benjamini–Hochberg-adjusted p-value of 1. Such FDR-adjusted p-values are sometimes called q-values. Because the decision to reject the null hypothesis in sequential tests depends on both the p-values and how they are ordered, the comparisons rejected by this method at the alpha level (two-sided test) are underlined in the output.

`nokwallis` suppresses the display of the Kruskal–Wallis test table.

`nolabel` causes the Dunn's test tables to display the actual data codes rather than the value labels.

`wrap` requests that Stata not break up wide tables to make them readable.

`level(#)` specifies the compliment of $\alpha \times 100$. The default is `level(95)` (or as set by `set level` [see [R] **level**]) and corresponds to $\alpha = 0.05$.

2.4 Stored results

`dunntest` stores the following in `r()`:

Scalars

`r(df)`	degrees of freedom for the Kruskal–Wallis test	`r(chi2_adj)`	χ^2 adjusted for ties for the Kruskal–Wallis test

Matrices

`r(Z)`	vector of Dunn's z-test statistics	`r(P)`	vector of (possibly adjusted) p-values for Dunn's z-test statistics

3 Remarks

▷ Example 1

Stata comes with data from the 1980 U.S. Census, and the documentation for the `kwallis` command works through an example to test whether the variable `medage` (median age of the population) varies by the variable `region` (Northeast, North Central, South, and West). The `dunntest` command defaults to presenting output from the omnibus `kwallis` command and follows it with a table of pairwise comparisons.

```
. sysuse census
(1980 Census data by state)
. dunntest medage, by(region) ma(none)
Kruskal-Wallis equality-of-populations rank test
```

region	Obs	Rank Sum
NE	9	376.50
N Cntrl	12	294.00
South	16	398.00
West	13	206.50

```
chi-squared =    17.041 with 3 d.f.
probability =     0.0007
chi-squared with ties =    17.062 with 3 d.f.
probability =     0.0007

                     Comparison of medage by region
                             (No adjustment)
Row Mean-|
Col Mean |         NE    N Cntrl      South
---------+---------------------------------
 N Cntrl |   2.698212
         |     0.0035
         |
   South |   2.793742  -0.067405
         |     0.0026     0.4731
         |
    West |   4.107611   1.477266   1.652733
         |     0.0000     0.0698     0.0492
```

The `kwallis` output appears as it does in the example in the manual. Below the output, there is a table that provides all six pairwise comparisons for the four regions. The table's title indicates the *varname* and *groupname*, and the subtitle indicates which method of adjustment is used (in this example, `dunntest` has defaulted to `No adjustment`). The row and column header labels indicate that the test results are based on the difference in mean ranks for each group, and the table entries give the pairwise z-test statistics with p-values beneath.

◁

▷ Example 2

Suppose one wants to adjust for multiple comparisons using the Holm–Šidák adjustment.

```
. sysuse census
(1980 Census data by state)
. dunntest medage, by(region) nokwallis ma(hs)

                     Comparison of medage by region
                           (Holm-Sidk)
Row Mean-|
Col Mean |          NE    N Cntrl      South
---------+------------------------------------
 N Cntrl |    2.698212
         |      0.0139
         |
   South |    2.793742  -0.067405
         |      0.0130     0.4731
         |
    West |    4.107611   1.477266   1.652733
         |      0.0001     0.1347     0.1404
```

Because `ma(hs)` was included as an option, the p-values of the tests rejected for a FWER of $\alpha = 0.05$ are underlined.

◁

▷ Example 3

In her 1964 article, Dunn included frequencies of individuals in seven broad occupational categories (for example, executives and sharecroppers) and the individuals' eligibility for home care defined by three exclusive categories (eligible for home care, ineligible for home care because of the lack of a responsible person, ineligible for home care because of the unavailability of a responsible person). She used these data in an analysis to illustrate her new test. She applied her test theory both to linear combinations between groups and to pairwise differences of mean ranks. In her worked example, she presented the results for only one pairwise test concerning whether occupational class among those with eligible home care is stochastically dominant over the occupational class of those for whom a responsible person is unavailable. We tested her data using `dunntest`, and our figures agree precisely with her results.

```
. use homecare
(Occupation and Home Care Eligibility for 383 Patients)
. dunntest occupation, by(eligibility) ma(none) nokwallis

                Comparison of occupation by eligibility
                            (No adjustment)
Row Mean-|
Col Mean |    Eligible   No respo
---------+----------------------
No respo |   -0.155969
         |      0.4380
         |
Responsi |   -2.022198  -1.441206
         |      0.0216     0.0748
```

4 Methods

4.1 Dunn's test

Dunn's z-test statistic (1) approximates exact rank-sum test statistics by using the mean rankings of the outcome in each group from the preceding Kruskal–Wallis test ($\overline{W}_i = W_i/n_i$, where W_i is the sum of ranks, and n_i is the sample size for the ith group) and basing inference on the differences in mean ranks in each group. To compare group A with group B, we calculate

$$z_i = \frac{y_i}{\sigma_i} \tag{1}$$

where i is one of the 1 to m multiple comparisons, $y_i = \overline{W}_A - \overline{W}_B$, and σ_i is the standard deviation of y_i, given by (2),

$$\sigma_i = \sqrt{\left\{ \frac{N(N+1)}{12} - \frac{\sum_{s=1}^{r} \tau_s^3 - \tau_s}{12(N-1)} \right\} \left(\frac{1}{n_A} + \frac{1}{n_B} \right)} \tag{2}$$

where N is the total number of observations across all groups, r is the number of tied ranks, and τ_s is the number of observations tied at the sth specific tied value. When there are no ties, the term with the summation in the denominator equals zero, and the calculation of (2) simplifies considerably.

4.2 Multiple-comparison adjustments

Here we describe each of the multiple-comparison adjustment procedures. p^* indicates an adjusted p-value. p refers to p-values that have the standard two-sided test interpretation $p = \mathrm{P}(|Z| \geq |z|)$. p_i refers to p-values as the order for the sequential procedures described below.

The Bonferroni adjustment multiplies each p-value by m, as shown in (3).

$$p^* = pm \tag{3}$$

The Šidák adjustment corrects the Bonferroni adjustment's error by defining the FWER and gives a slightly smaller p^*, as shown in (4).

$$p^* = 1 - (1 - p)^m \tag{4}$$

Holm's stepwise adjustment controls the FWER by ordering all m p-values from smallest to largest, providing a Bonferroni adjustment based on i and m, and fails to reject all pairwise tests, starting with the first test for which $p^* > \alpha/2$, as shown in (5).

$$p_i^* = p(m + 1 - i) \tag{5}$$

The Holm–Šidák stepwise adjustment follows Holm's method but applies the Šidák adjustment based on i and m, as shown in (6).

$$p_i^* = 1 - (1 - p)^{(m+1-i)} \tag{6}$$

The Benjamini–Hochberg stepwise adjustment controls the FDR by ordering all m p-values from largest to smallest and adjusting p by multiplying by $m/(m + 1 - i)$. It fails to reject all pairwise tests, starting with the first test for which $p^* > \alpha/2$, as shown in (7). Simes (1986) first described this adjustment procedure, or one very similar to it, but did not provide the FDR interpretation.

$$p_i^* = p\frac{m}{(m + 1 - i)} \tag{7}$$

5 References

Benjamini, Y., and Y. Hochberg. 1995. Controlling the false discovery rate: A practical and powerful approach to multiple testing. *Journal of the Royal Statistical Society, Series B* 57: 289–300.

Dunn, O. J. 1961. Multiple comparisons among means. *Journal of the American Statistical Association* 56: 52–64.

———. 1964. Multiple comparisons using rank sums. *Technometrics* 6: 241–252.

Holland, B. S., and M. D. Copenhaver. 1988. Improved Bonferroni-type multiple testing procedures. *Psychological Bulletin* 104: 145–149.

Holm, S. 1979. A simple sequentially rejective multiple test procedure. *Scandinavian Journal of Statistics* 6: 65–70.

Kruskal, W. H., and W. A. Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 47: 583–621.

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.

Šidák, Z. 1967. Rectangular confidence regions for the means of multivariate normal distributions. *Journal of the American Statistical Association* 62: 626–633.

Simes, R. J. 1986. An improved Bonferroni procedure for multiple tests of significance. *Biometrika* 73: 751–754.

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

About the author

Alexis Dinno is an assistant professor at the school of Community Health at Portland State University. She is trained as a social epidemiologist with interests in applied quantitative methods, social ecology, and health equity. She wrote the `dunntest`, `dthaz`, `paran`, and `tost` packages.