# 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 following rejection of a Kruskal-Wallis test, and is now implemented for Stata in the `dunn` package. `dunn` produces the multiple comparisons following a Kruskal-Wallis *k*-way test using Stata's built-in `kwallis` command, and includes options to control the family-wise error rate using Dunn's proposed Bonferroni adjustment, the Šidák adjustment, Holm stepwise adjustment, or the Holm-Šidák stepwise adjustment. There is an additional option to instead control the false discovery rate using the Benjamini-Hochberg stepwise adjustment.

**Keywords:** dunn, kwallis, Dunn's test, Kruskal-Wallis test, multiple comparisons, family-wise error rate, Bonferroni, Šidák, Holm, Holm-Šidák, false discovery rate

## 1 Introduction

Oneway omnibus tests, such as the common oneway ANOVA, typically pose null hypotheses that measurements across some number of groups are all derived from a common distribution. We might think of such tests as answering questions of a generic form "do we need to bother looking more closely between groups for differences?" Without evidence to reject the null hypothesis of such tests, the researcher's work moves on to new topics. One the other hand, if the null hypothesis of an omnibus test is rejected, the work at hand grows as the question "which of these groups is different from which?" Were one to use ANOVA to test for mean difference, upon rejection of the null hypothesis, one would proceed to make multiple pairwise comparisons using *t* tests for mean difference in unpaired data. However the ANOVA comes with some quite restrictive assumptions concerning the distributions of the groups under scrutiny: they 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 oneway ANOVA which sacrifices the precision of discriminating means for a discrimination of stochastic dominance (i.e. the probability than a randomly drawn observation from one group will be greater than a randomly drawn observation in another), buts gains the ability to do so regardless of the distributions of the measures in each group. If the modest additional assumptions that the measures are continuous, and that the (unspecified) distributions in each group differ only in terms of their centrality, then the Kruskal-Wallis test may be understood as an omnibus test for median difference. Upon rejection of the null hypothesis of this test, one would proceed to conduction multiple pairwise comparisons for stochastic dominance (or median difference).

That the appropriate test for such comparisons is a nonparametric analog to the $t$ test seems straightforward—the rank-sum test (Wilcoxon 1945; Mann and Whitney 1947) is just such a beast—but the application is not quite straightforward. With the $t$ tests that follow rejection of the null hypothesis of an ANOVA make use of ANOVA's strict assumption about equal variances by using the pooled estimate of variance in calculating the standard error of the $t$ test statistics. Were one to forge ahead following a Kruskal-Wallis, one would ignore this assumption which is important for the interpretation of median difference, but more fundamentally important for interpreting the rank sum tests as part of a family of inferences bound up in the omnibus hypothesis: the very ranks of the data upon which the tests are based change if reranked in a pairwise fashion. Dunn's insight (1964) was to retain the rank sums from the omnibus test, and to construct a $z$ test statistic approximation 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 $\alpha$ (which usually means 'the probability of falsely rejecting the null hypothesis for a single test') within the inferential framework of hypothesis test. Dunn (1961) described a strategy for addressing this issue with a 'Bonferroni adjustment' which, looked at one way, modifies the rejection level for any individual test by dividing $\alpha$ by the total number of tests, requiring a much smaller $p$-value to reject any test. Looked at another way, this adjustment leaves alpha numerically intact, but multiplies the $p$-value. This forms the basis of the 'family-wise error rate' (FWER) redefinition of $\alpha$ to mean 'the probability of falsely rejecting the null hypothesis once out of all tests performed'. While the Bonferroni adjustment introduced the logic of the FWER, improvements followed, including Šidák's adjustment (1967) is a slightly more powerful, but similar approach, Holm's sequential adjustment, and the Holm-Šidák sequential adjustment (1979; sometimes credited to Holland and Copenhaver 1988), which treat subsequent pairwise hypothesis tests as parts of different families based on whether or not previous tests were rejected. Finally, Benjamini and Hochberg (1995) reasoned that $\alpha$ should be interpreted as a desired 'false discovery rate' (FDR) and should reflect how the expected rate of false discoveries changes after rejecting some pairwise tests in sequence.

Dunn's test has been growing in popularity over the past two decades (Figure 1).[1] These are frequently used with some form of multiple comparisons adjustment. During these same two decades, of the 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 March 6th, 2014 search of Google Scholar for citations with the exact phrase "Dunn's test" for each of the years 1994–2013, inclusive.

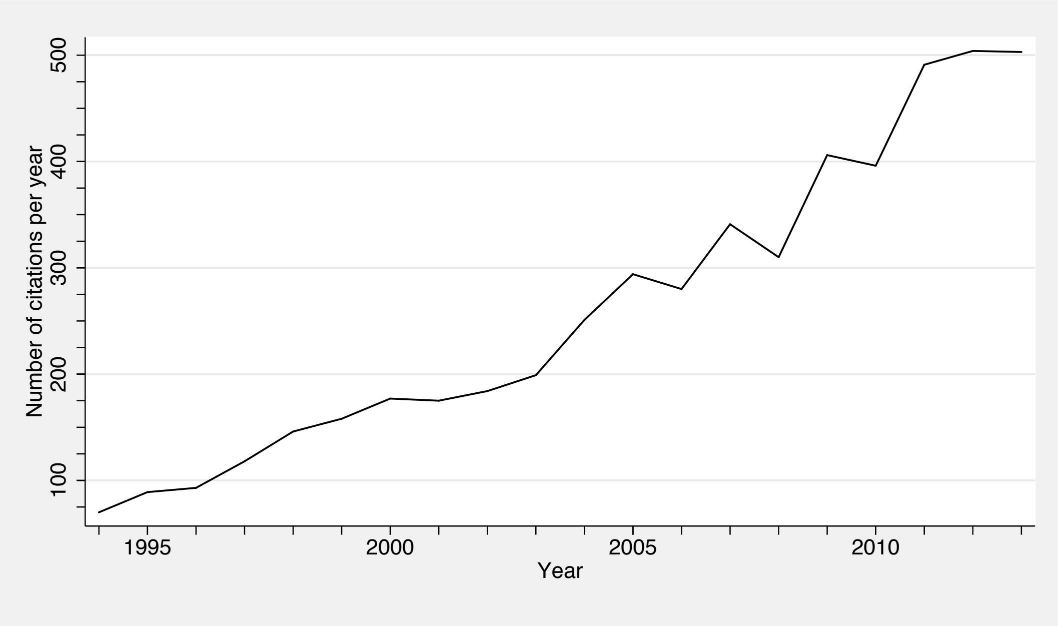

Figure 1: Citations indexed by Google Scholar including "Dunn's test" over two decades

# 2 The dunn command

## 2.1 Syntax

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

where *method* is one of

`none|bonferroni|sidak|holm|hs|bh`

## 2.2 Description

`dunn` reports the results of Dunn's test (1964) for stochastic dominance among multiple pairwise comparisons following a Kruskal-Wallis test of stochastic dominance among $k$ groups Kruskal and Wallis (1952) using [R] **kwallis**. `dunn` 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**). Like the rank-sum test, if the data can be assumed to be continuous, and the distributions are assumed identical except for a shift in centrality, Dunn's 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. `dunn` accounts for tied ranks. `by()` is required.

## 2.3 Options

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

`ma(`*method*`)` is required. It specifies the method of adjustment used for multiple comparisons, and must take one of the following values: `none`, `bonferroni`, `sidak`, `hs`, or `bh`. `none` is the default method assumed if the `ma` option is omitted. These methods perform as follows:

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

`bonferroni` specifies a Bonferroni adjustment where the family-wise error rate (FWER) is adjusted by multiplying the $p$-values in each pairwise test by $m$ (the total number of pairwise tests) as per Dunn (1961). 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 $1-(1-p)^m$ as per 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 as ordered from smallest to largest with $p(m+1-i)$, where $i$ is the position in the ordering as per Holm (1979). Stata will report a maximum Holm-adjusted $p$-value of 1. Because in sequential tests the decision to reject the null hypothesis depends both on the $p$-values and their ordering, those comparisons rejected with 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 as ordered from smallest to largest with $1-(1-p)^{m+1-i}$, where $i$ is the position in the ordering as per Holm (1979). Stata will report a maximum Holm-Šidák-adjusted $p$-value of 1. Because in sequential tests the decision to reject the null hypothesis depends both on the $p$-values and their ordering, those comparisons rejected with this method at the $\alpha$ level (two-sided tests) are underlined in the output.

`bh` specifies a Benjamini-Hochberg adjustment where the false discovery rate (FDR) is adjusted sequentially by adjusting the $p$-values of each pairwise test as ordered from largest to smallest with $p[m/(m+1-i)]$, where $i$ is the position in the ordering as per Benjamini and Hochberg (1995). Stata will report a maximum Benjamini-Hochberg-adjusted $p$-value of 1. Such FDR-adjusted $p$-values are sometimes refered to as $q$-values in the literature. Because in sequential tests the decision to reject the null hypothesis depends both on the $p$-values and their ordering, those comparisons rejected with 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 actual data codes to be displayed rather than the value labels in the Dunn's test tables.

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

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

# 3 Saved results

Scalars

| | | | |
|---|---|---|---|
| `r(df)` | degrees of freedom for the Kruskal-Wallis test | `r(chi2_adj)` | $\chi^2$ adjusted for ties for the Kruskal-Wallis test |

Matrices

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

# 4 Remarks

### ▷ Example 1

Stata comes with data from the 1980 U.S. Census, and the documentation for the `kwallis` command ([R] **kwallis**) 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 `dunn` 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)
. dunn 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 just as from the example in the manual. Below that there is a table giving 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 being used (in this example `dunn` has defaulted to "No adjustment"). The row and column header labels indicate that the test is 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 we want to adjust for multiple comparisons using the Holm-Šidák adjustment:

```
. use homecare
(Occupation and Home Care Eligibility for 383 Patients)
. dunn medage, by(region) nokwallis ma(hs)
                              Comparison of medage by region
                                      (Holm-Sidák)
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 we included `ma(`*hs*`)` as an option, the $p$-values of those tests that we would reject for a FWER of $\alpha = 0.05$ are underlined.

◁

▷ **Example 3**

In her 1964 paper, Dunn included frequencies of individuals in seven exclusive broad occupational categories (e.g. ranging from executives through share croppers), and their eligibility for home care defined by three exclusive categories (eligible for home care, ineligible for home care due to the lack of a responsible person, ineligible due to the unavailability of a responsible person). She used these data in an analysis illustrating her new test. She was interested in applying her test theory both to linear combinations between groups in general, and in the specific case of testing pairwise differences of mean ranks. In her worked example she presented the results for only a single pairwise test: 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. If we test her data using `dunn`, we find that our figure agree precisely with her result:

```
. use homecare
(Occupation and Home Care Eligibility for 383 Patients)
. dunn 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
```

◁

# 5 Methods

## 5.1 Dunn's test

Dunn's $z$ test statistic (1) approximates exact rank sum test statistics by using the mean rankings of the outcome ($\overline{W} = W/n$) in each group from the preceding Kruskal-Wallis test, and basing inference on the differences in mean ranks in each group; for a comparison between one group $A$ and another group $B$

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

where $i$ is one of the 1 to $m$ multiple comparisons, $y_i = \overline{W}_A - \overline{W}_B$, and $\sigma_i$ 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 $\tau_s$ is the number of observations tied at the $s^{\text{th}}$ specific tied value. When there are no ties, the term with the summation in the denominator equals zero, and the calculation of (2) simplifies.

## 5.2 Multiple comparisons adjustments

Each of the multiple comparisons adjustment procedures are described here; $p^*$ indicates an adjusted $p$-value, $p$-values, $p$, have the standard two-sided test interpretation: $p = \text{P}(|Z| \geq |z|)$, and $p_i$ will refer to $p$-values as order for the sequential procedures described below.

The Bonferroni adjustment simple multiplies each $p$-values by $m$ as in (3).

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

The Šidák adjustment corrects the Bonferroni adjustment's error in defining the FWER, and gives a slightly smaller $p^*$ as 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 in (5).

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

The Holm-Šidák stepwise adjustment is as per Holm's method, but applies the Šidák adjustment based on $i$ and $m$, as 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 $m/(m+1-i)$, and fails to reject all pairwise tests starting with the first test for which $p^* > \alpha/2$, as in (7).

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

## 6 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 (Methodological)* 57(1): 289–300.

Dunn, O. J. 1961. Multiple Comparisons Among Means. *Journal of the American Statistical Association* 56(293): 52–64.

———. 1964. Multiple Comparisons Using Rank Sums. *Technometrics* 6(3): 241–252.

Holland, B. S., and M. D. Copenhaver. 1988. Improved Bonferroni-Type Multiple Testing Procedures. *Psychological Bulletin* 104(1): 145–149.

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

Kruskal, W. H., and A. Wallis. 1952. Use of ranks in one-criterion variance analysis. *Journal of the American Statistical Association* 47(260): 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(318): 626–633.

Wilcoxon, F. 1945. Individual comparisons by ranking methods. *Biometrics Bulletin* 1(6): 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, and wrote the `dunn`, `dthaz`, `paran`, and `tost` Stata packages.