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.

The *Stata Journal* is indexed and abstracted by *CompuMath Citation Index*, *Current Contents/Social and Behavioral Sciences*, *RePEc: Research Papers in Economics*, *Science Citation Index Expanded* (also known as *SciSearch*, *Scopus*, and *Social Sciences Citation Index*.

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

http://www.stata-journal.com

Subscriptions are available from StataCorp, 4905 Lakeway Drive, College Station, Texas 77845, telephone 979-696-4600 or 800-STATA-PC, fax 979-696-4601, or online at

http://www.stata.com/bookstore/sj.html

Subscription rates listed below include both a printed and an electronic copy unless otherwise mentioned.

U.S. and Canada		Elsewhere	
Printed & electronic		**Printed & electronic**	
1-year subscription	$ 98	1-year subscription	$138
2-year subscription	$165	2-year subscription	$245
3-year subscription	$225	3-year subscription	$345
1-year student subscription	$ 75	1-year student subscription	$ 99
1-year university library subscription	$125	1-year university library subscription	$165
2-year university library subscription	$215	2-year university library subscription	$295
3-year university library subscription	$315	3-year university library subscription	$435
1-year institutional subscription	$245	1-year institutional subscription	$285
2-year institutional subscription	$445	2-year institutional subscription	$525
3-year institutional subscription	$645	3-year institutional subscription	$765
Electronic only		**Electronic only**	
1-year subscription	$ 75	1-year subscription	$ 75
2-year subscription	$125	2-year subscription	$125
3-year subscription	$165	3-year subscription	$165
1-year student subscription	$ 45	1-year student subscription	$ 45

Back issues of the *Stata Journal* may be ordered online at

http://www.stata.com/bookstore/sjj.html

Individual articles three or more years old may be accessed online without charge. More recent articles may be ordered online.

http://www.stata-journal.com/archives.html

The *Stata Journal* is published quarterly by the Stata Press, College Station, Texas, USA.

Address changes should be sent to the *Stata Journal*, StataCorp, 4905 Lakeway Drive, College Station, TX 77845, USA, or emailed to sj@stata.com.

The *Stata Journal* (ISSN 1536-867X) is a publication of Stata Press. Stata, STATA, Stata Press, Mata, MATA, and NetCourse are registered trademarks of StataCorp LP.

The Stata Journal (2013)
13, Number 2, pp. 379–381

Bonferroni and Holm approximations for Šidák and Holland–Copenhaver q-values

Roger B. Newson
National Heart and Lung Institute
Imperial College London
London, UK
r.newson@imperial.ac.uk

Abstract. I describe the use of the Bonferroni and Holm formulas as approximations for Šidák and Holland–Copenhaver formulas when issues of precision are encountered, especially with q-values corresponding to very small p-values.

Keywords: st0300, parmest, qqvalue, smileplot, multproc, multiple-test procedure, familywise error rate, Bonferroni, Šidák, Holm, Holland, Copenhaver

1 Introduction

Frequentist q-values for a range of multiple-test procedures are implemented in Stata by using the package `qqvalue` (Newson 2010), downloadable from the Statistical Software Components (SSC) archive. The Šidák q-value for a p-value p is given by $q_{\rm sid} = 1 - (1-p)^m$, where m is the number of multiple comparisons (Šidák 1967). It is a less conservative alternative to the Bonferroni q-value, given by $q_{\rm bon} = \min(1, mp)$. However, the Šidák formula may be incorrectly evaluated by a computer to 0 when the input p-value is too small to give a result lower than 1 when subtracted from 1, which is the case for p-values of 10^{-17} or less, even in double precision. q-values of 0 are logically possible as a consequence of p-values of 0, but in this case, they may be overliberal. This liberalism may possibly be a problem in the future, given the current technology-driven trend of exponentially increasing multiple comparisons and the human-driven problem of ingenious data dredging. I present a remedy for this problem and discuss its use in computing q-values and discovery sets.

2 Methods for q-values

The remedy used by the SSC packages `qqvalue` and `parmest` (Newson 2003) is to substitute the Bonferroni formula for the Šidák formula for such small p-values. This works because the Bonferroni and Šidák q-values converge in ratio as p tends to 0. To prove this, I show that for $0 \le p < 1$,

$$dq_{\rm bon}/dp = m \quad \text{and} \quad dq_{\rm sid}/dp = m(1-p)^{m-1}$$

and that the Šidák/Bonferroni ratio of these derivatives is $(1-p)^{m-1}$, which is 1 if $p = 0$. By L'Hôpital's rule, it follows that the ratio $q_{\rm sid}/q_{\rm bon}$ also tends to 1 as p tends to 0.

 st0300

A similar argument shows that the same problem exists with the q-values output by the Holland–Copenhaver procedure (Holland and Copenhaver 1987). If the m input p-values, sorted in ascending order, are denoted p_i for i from 1 to m, then the Holland–Copenhaver procedure is defined by the formula

$$s_i = 1 - (1 - p_i)^{m-i+1}$$

where s_i is the ith s-value. (In the terminology of Newson [2010], s-values are truncated at 1 to give r-values, which are in turn input into a step-down procedure to give the eventual q-values.) The remedy used by `qqvalue` here is to substitute the s-value formula for the procedure of Holm (1979), which is

$$s_i = (m - i + 1)p_i$$

whenever $1 - p_i$ is evaluated as 1. This also works because the two s-value formulas converge in ratio as p_i tends to 0. Note that the Holm procedure is derived from the Bonferroni procedure by using the same step-down method as is used to derive the Holland–Copenhaver procedure from the Šidák procedure.

3 Methods for discovery sets

The SSC package `smileplot` (Newson and the ALSPAC Study Team 2003) also implements a range of multiple-test procedures by using two commands, `multproc` and `smileplot`. However, instead of outputting q-values, `smileplot` outputs a corrected critical p-value threshold and a corresponding discovery set, defined as the subset of input p-values at or below the corrected critical p-value. The Šidák-corrected critical p-value corresponding to an uncorrected critical p-value p_{unc} is given by $c_{\mathrm{sid}} = 1 - (1 - p_{\mathrm{unc}})^{1/m}$ and may be overconservative if wrongly evaluated to 0. In this case, the quantity that might be wrongly computed as 1 is $(1 - p_{\mathrm{unc}})^{1/m}$. When this happens, `smileplot` substitutes the Bonferroni-corrected critical p-value $c_{\mathrm{bon}} = p_{\mathrm{unc}}/m$. However, this is a slightly less elegant remedy in this case because the quantity $(1 - p_{\mathrm{unc}})^{1/m}$ is usually evaluated to 1 because m is large and not because p_{unc} is small.

To study the behavior of the Bonferroni approximation for large m, we define $\lambda = 1/m$ and note that

$$dc_{\mathrm{bon}}/d\lambda = p_{\mathrm{unc}} \quad \text{and} \quad dc_{\mathrm{sid}}/d\lambda = -\ln(1 - p_{\mathrm{unc}})(1 - p_{\mathrm{unc}})^{\lambda}$$

implying (by L'Hôpital's rule) that in the limit, as λ tends to 0, the Šidák/Bonferroni ratio of the two derivatives (and therefore of the two corrected thresholds) tends to $-\ln(1 - p_{\mathrm{unc}})/p_{\mathrm{unc}}$. This quantity is not as low as 1 but is 1.150728, 1.053605, 1.025866, and 1.005034 if p_{unc} is 0.25, 0.10, 0.05, and 0.01, respectively. Therefore, the Bonferroni approximation in this case is still slightly conservative for a very large number of multiple comparisons over a range of commonly used uncorrected critical p-values, but is less conservative than the value of 0, which would otherwise be computed.

This argument is easily generalized to the Holland–Copenhaver procedure. In this case, `smileplot` initially calculates a vector of m candidate critical p-value thresholds by using the formula

$$c_i = 1 - (1 - p_{\mathrm{unc}})^{1/(m-i+1)}$$

for i from 1 to m and selects the corrected critical p-value corresponding to a given uncorrected critical p-value from these candidates by using a step-down procedure. If the quantity $(1 - p_{\mathrm{unc}})^{1/(m-i+1)}$ is evaluated as 1, then `smileplot` substitutes the corresponding Holm critical p-value threshold

$$c_i = p_{\mathrm{unc}}/(m - i + 1)$$

which again is conservative as $m - i + 1$ becomes large (corresponding to the smallest p-values from a large number of multiple comparisons), but is less conservative than the value of 0, which would otherwise be computed.

Newson (2010) argues that q-values are an improvement on discovery sets because, given the q-values, different members of the audience can apply different input critical p-values and derive their own discovery sets. The technical issue of precision presented here may be one more minor reason for preferring q-values to discovery sets.

4 Acknowledgment

I would like to thank Tiago V. Pereira of the University of São Paulo in Brazil for drawing my attention to this issue of precision with the Šidák and Holland–Copenhaver procedures.

5 References

Holland, B. S., and M. D. Copenhaver. 1987. An improved sequentially rejective Bonferroni test procedure. *Biometrics* 43: 417–423.

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

Newson, R. 2003. Confidence intervals and p-values for delivery to the end user. *Stata Journal* 3: 245–269.

Newson, R., and the ALSPAC Study Team. 2003. Multiple-test procedures and smile plots. *Stata Journal* 3: 109–132.

Newson, R. B. 2010. Frequentist q-values for multiple-test procedures. *Stata Journal* 10: 568–584.

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

About the author

Roger B. Newson is a lecturer in medical statistics at Imperial College London, UK, working principally in asthma research. He wrote the `parmest`, `qqvalue`, and `smileplot` Stata packages.