Comparing Statistical Software Packages: The Case of the Logrank Test in StatXact

Herman CALLAERT

This article can be read as a companion to an earlier article by R. A. Oster in which StatXact, LogXact, and Stata were examined. In most cases, comparing software packages on their main features, such as accuracy, user friendliness, and adequate documentation, is sufficient for giving the reader interesting and useful information. In some instances, however, one has to take a different perspective for discovering the mechanism behind discrepancies. An excellent software package, using accurate computations and ample documentation, may nevertheless not do what its menu implies. The case of the logrank test in StatXact is a good example, illustrating the need to always check procedures and formulas against labels and names.

KEY WORDS: Exact statistical methods; Exponential scores test; Nonparametric test; Savage test.

1. INTRODUCTION

An examination of statistical software packages, followed by a discussion and a list of motivated recommendations, is extremely useful for the practicing statistician. A recent example of such an examination is the article by R. A. Oster (2002), in which the merits of exact statistical methods were described, while analyzing the software packages StatXact 5, LogXact 4.1, and Stata 7. Typical topics in this kind of analysis are: installation and hardware requirements, documentation, data entry and data management, available procedures, accuracy and reliability, technical support, and so on. Also, ease of use (does the user have to type commands or is the package menu driven?) is often an important benchmark. All of the above topics are important, and Oster's article contains a wealth of information on the packages discussed.

But even if a statistical software package is statistically and computationally sound and comes with extensive information on its procedures and formulas, it nevertheless remains to be checked whether the same name indeed refers to the same statistical quantity. When an identical statistical test yields a p value of 4.68% with one package and 5.23% with another package, what is happening? Can it solely be contributed to the difference between exact methods and asymptotic procedures? Or is there numerical instability in one of the packages? Or is there a flaw in the programming? These are the type of questions asked naturally. It is rather uncommon to suspect a package of using a standard name for a statistical test and to associate it with a nonstandard formula. But it happens, as will be demonstrated in this article.

Herman Callaert is Professor of Statistics, Center for Statistics, Limburgs Universitair Centrum, B-3590 Diepenbeek, Belgium (E-mail: herman.callaert@luc.ac.be).

2. ONE STATISTIC, DIFFERENT NAMES

It is not uncommon to encounter different formulas for the same statistic. Moreover, those formulas do not have to be algebraically equivalent in order to generate "equivalent" statistics, leading to identical p values and hence to identical statistical decisions. The variety of expressions often stems from simple transformations, such as location and scale changes. A classical example is the relation between the Wilcoxon statistic W and its representation in the Mann–Whitney form U, where

$$U = W - \frac{1}{2}n(n+1) \tag{1}$$

for n treated subjects versus m controls $(n + m = N)$ in a two-sample comparison. Introducing $0 - 1$ variables U_i such that $U_i = 0$ if the ith ordered outcome (in the total group of N ordered observations) comes from a control, the classical form of the Wilcoxon statistic, defined as the sum of the ranks of the treated subjects, is

$$W = \sum_{i=1}^{N} iU_i. \tag{2}$$

It then takes a moment to recognize the following formula in Kalbfleisch and Prentice (1980, p. 147) as being equivalent with the Wilcoxon statistic

$$\sum_{i=1}^{N} \left(1 - 2\prod_{k=1}^{i} \frac{N-k+1}{N-k+2}\right) U_i. \tag{3}$$

However, (3) is equivalent with (2), and expression (3), adapted for ties and censoring, is the formula used in the StatXact manual (2001, p. 220). When statisticians choose "Wilcoxon–Mann–Whitney" from a menu in a statistical computing package, they expect to obtain the right p value for the observations in their experiment. Whether the package uses formula (1), (2), or (3) is, after all, immaterial to the practitioner (the developer of course should use those formulas which lead to optimal stability and numerical accuracy).

3. ONE NAME, DIFFERENT STATISTICS

Consider the two-sample problem where the null hypothesis H_0 states that $G(x) = F(x)$, while the alternative H_1 indicates that the treatment outcomes Y are stochastically larger than the control outcomes X (note that $Y \sim G(x)$ and $X \sim F(x)$). When dealing with failure times, the logrank statistic is a widely used nonparametric test statistic. Although a typical experiment (such as a clinical trial) usually contains censored observations, we will not consider this aspect here.

 DOI: 10.1198/0003130031900

When no ties are present, the Savage (or exponential scores) statistic (Lehmann 1998, p. 103–104)

$$\mathrm{S} = \sum_{i=1}^{N} \left(\sum_{j=1}^{i} \frac{1}{N-j+1} \right) U_i, \tag{4}$$

and the logrank statistic (Cox and Oakes 1984, p. 98)

$$\mathrm{LR} = \sum_{i=1}^{N-1} \left(U_i - \sum_{j=i}^{N} \frac{U_j}{N-i+1} \right) \tag{5}$$

are just rescaled versions of each other, and hence are equivalent for performing statistical analyses.

The statistics (4) and (5) are unsuitable in the presence of ties, and adjustments have to be made. In the following formulas e is used to denote the number of distinct outcomes (i.e., the number of ties), and n_i represents the number of observations equal to the ith smallest value. Furthermore $\sum_{k=1}^{i} n_k = \ell_i$ and $\ell_0 = 0$. The classical logrank statistic for ties (see, e.g., Kalbfleisch and Prentice 1980, p. 80; or Cox and Oakes 1984, p. 104) is defined as

$$\mathrm{LR} = \sum_{i=1}^{e-1} \left(U_i - n_i \sum_{j=i}^{e} \frac{U_j}{N-\ell_{i-1}} \right), \tag{6}$$

and this statistic is different from the Savage statistic adapted for ties, as used in StatXact (2001, p. 219):

$$\mathrm{S} = \sum_{i=1}^{e} \left[\frac{1}{n_i} \sum_{k=\ell_{i-1}+1}^{\ell_i} \left(\sum_{j=1}^{k} \frac{1}{N-j+1} \right) - 1 \right] U_i \tag{7}$$

which for easy comparison with formula (6) can also be rewritten in the equivalent form

$$\mathrm{S}^* = \sum_{i=1}^{e-1} \widetilde{v}_i \left(U_i - n_i \sum_{j=i}^{e} \frac{U_j}{N-\ell_{i-1}} \right), \tag{8}$$

with

$$\widetilde{v}_i = \frac{N-\ell_{i-1}}{n_i} \sum_{j=\ell_{i-1}+1}^{\ell_i} \frac{1}{N-j+1}, \tag{9}$$

and with $S^* = -S$.

Hence, choosing "Logrank" in the StatXact menu, and looking up in the manual the scores which are called "Logrank Scores," does not guarantee that one is working with the usual logrank test as defined in (6). In fact, when ties are present, StatXact uses the statistic (7) which is different from (6), leading to different p values and to possibly different statistical decisions.

4. AN EXAMPLE

Consider a two-sample experiment where the outcome is survival time (measured in months). Out of 15 subjects, 7 are randomly assigned to treatment (code = 1) while 8 subjects serve as the control (code = 2). The configuration of ties in the outcome variable is seen to be: $e = 6$, $n_1 = 2$, $n_2 = 3$, $n_3 = 1$, $n_4 = 2$, $n_5 = 3$, and $n_6 = 4$. The censoring indicator is set equal to 1, indicating complete cases for all observations. The data, as entered into StatXact in the "CaseData" format, are shown in Table 1.

Now choose "Logrank" in the StatXact menu, and run the test on the data in Table 1. The special feature of StatXact is its use of exact methods for statistical analysis. This leads to an observed value of its logrank statistic of 3.190 and a two-sided p value of 4.68%, indicating that there is a statistically significant difference between treatment and control (at the 5% level). When the same analysis is performed with the classical procedures in SAS (which are based on asymptotic methods), one obtains a value of -2.841 for the observed logrank statistic with an associated two-sided p value of 5.23%, indicating that a significant difference can not yet be claimed at the 5% level.

Since the logrank test statistic is used both in StatXact and in SAS, a straightforward explanation of the observed discrepancy in p values (4.68% versus 5.23%) should stress the difference in the computation of the null-distribution: StatXact is able to compute the exact null-distribution whereas SAS (and most other classical packages) use an asymptotic approximation. However, as explained in paragraph 3, the same name does not cover the same statistic, and hence the above comparison is not a fair one.

5. A FAIR COMPARISON

The usual logrank test, as defined in (6), is unfortunately not

Table 1. AU: Please Give Table Caption

Case#	*TREAT*	*OUTCOME*	*CENSOR*
00001	1	1	1
00002	1	1	1
00003	1	5	1
00004	1	6	1
00005	1	6	1
00006	1	6	1
00007	1	6	1
00008	2	2	1
00009	2	2	1
00010	2	2	1
00011	2	3	1
00012	2	4	1
00013	2	4	1
00014	2	5	1
00015	2	5	1

Table 2. Author: Please Give Table Title

CASE#	*TREAT*	*LRSCORES*
00001	1	−0.866667
00002	1	−0.866667
00003	1	0.114896
00004	1	1.114896
00005	1	1.114896
00006	1	1.114896
00007	1	1.114896
00008	2	−0.635897
00009	2	−0.635897
00010	2	−0.635897
00011	2	−0.535897
00012	2	−0.313675
00013	2	−0.313675
00014	2	0.114896
00015	2	0.114896

immediately available in StatXact. However, StatXact can generate exact null distributions for permutation tests with general scores. How this works is explained in section 9.13 of the StatXact manual (2001); it is important to pay attention to the specific standardization of the statistic as used by StatXact. One can then compute appropriate scores and feed them into the general permutation test. This yields an exact null distribution and hence p values based on exact procedures. It is shown in the appendix how scores have to be computed in order that the general permutation test is exactly equal to the usual logrank test. This leads to the data shown in Table 2 to be entered into StatXact:

Choosing "Permutation" from the StatXact menu and applying this test with the scores of Table 2 yields 2.841 as the observed value of the test statistic (which up to its sign coincides with the observed value of the logrank statistic in SAS) with a two-sided p value of 5.05%.

Now we are in a position of making a fair comparison since both procedures use structurally the same statistic. The only difference is that the StatXact p value is based on the exact null distribution of the usual logrank statistic, whereas the SAS p value stems from the asymptotic approximation. This is the explanation for the difference in p values that are produced by the two packages (5.05% versus 5.23%).

6. CONCLUSION

In his "Discussion and Recommendations" section, Oster (2002) wrote that "... all statisticians, and all data analysts and researchers who perform categorical and/or nonparametric statistical analysis, need to have StatXact. StatXact has capabilities that are not found in any other statistical software package, and will correctly analyze datasets that are small, sparse, unbalanced, and not normally distributed." I fully agree with this statement, and it therefore is even more important to point out the possible confusion that can result from choosing "Logrank" in the StatXact menu. The logrank test statistic is widely used in the analysis of lifetime data, and the full strength of StatXact can be used to correctly analyze those types of data through exact methods. How this can be done has been discussed in this article.

APPENDIX: RELATIONS BETWEEN DIFFERENT REPRESENTATIONS OF LINEAR RANK TEST STATISTICS

Many common nonparametric test statistics can be captured into the general framework of linear rank test statistics. These in turn can be represented in a variety of ways, of which the following three representations are classical (see Section 2 for the definition of U_i):

$$T = \sum_{i=1}^{N} a_i U_i, \tag{A.1}$$

which yields the sum of the scores of the treated subjects,

$$T^c = \sum_{i=1}^{N} w_i U_i, \tag{A.2}$$

which is the "centered" version, with $w_i = a_i - \bar{a}$ (and hence with $E(T^c) = 0$); and

$$T^{0-E} = \sum_{i=1}^{N-1} v_i \left(U_i - \sum_{j=1}^{N} \frac{U_j}{N-i+1} \right) \tag{A.3}$$

which is, up to its sign, equal to T^c, but now expressed in an "observed minus expected" framework. Note that

$$v_i = -w_i - \frac{1}{N-i} \sum_{k=1}^{i} w_k.$$

Formulas (A.1), (A.2), and (A.3) are valid when there are no ties, but they lead to different (but structurally equivalent) expressions for many well-known statistics, such as the Wilcoxon–Mann–Whitney and the Savage/exponential scores/logrank statistics.

The presence of ties leads to the use of midranks. Following the notation of Section 3 for the configuration of ties, formulas (A.1)–(A.3) generalize as follows (note that $0 \le U_i \le n_i$ and $\sum_{i=1}^{e} U_i = n$):

$$\widetilde{T} = \sum_{i=1}^{e} \widetilde{a}_i U_i, \tag{A.4a}$$

with

$$\widetilde{a}_i = \frac{1}{n_i} \sum_{k=\ell_{i-1}+1}^{\ell_i} a_k. \tag{A.4b}$$

The centered version now equals to

$$\widetilde{T}^c = \sum_{i=1}^{e} \widetilde{w}_i U_i, \tag{A.5a}$$

with

$$\widetilde{w}_i = \widetilde{a}_i - \bar{a} = \frac{1}{n_i} \sum_{k=\ell_{i-1}+1}^{\ell_i} w_k. \tag{A.5b}$$

Finally, the observed minus expected format can be written as

$$\widetilde{T}^{0-E} = \sum_{i=1}^{e-1} \widetilde{v}_i \left(U_i - n_i \sum_{j=i}^{e} \frac{U_j}{N-\ell_{i-1}} \right), \tag{A.6a}$$

with

$$\widetilde{v}_i = -\widetilde{w}_i - \frac{1}{N-\ell_i} \sum_{k=1}^{i} n_k \widetilde{w}_k. \tag{A.6b}$$

The definition (6) of the usual logrank test implies that the scores $\widetilde{v}_i$ in (A.6a) are all identical equal to one. From (A.6b) it then follows that the scores $\widetilde{w}_i$ in (A.5a) are equal to

$$\widetilde{w}_i = \sum_{k=1}^{i} \frac{n_k}{N-\ell_{k-1}} - 1. \tag{A.7}$$

Since StatXact is programmed to use linear rank statistics in their "centered" form (A.5a), the scores (A.7) have to be imputed into

the general permutation test in order to perform an exact analysis for the logrank test (6). For a configuration of ties as in the example of Section 4, formula (A.7) leads to $\widetilde{w}_1 = -0.866667$, $\widetilde{w}_2 = -0.635897$, $\widetilde{w}_3 = -0.535897$, $\widetilde{w}_4 = -0.313675$, $\widetilde{w}_5 = 0.114896$ and $\widetilde{w}_6 = 1.114896$.

[Received September 2002. Revised May 2003.]

REFERENCES

Cox, D. R., and Oakes, D. (1984), *Analysis of Survival Data*, New York: Chapmann and Hall.

CYTEL Software Corporation (2001), *StatXact 5. User Manual*, Cambridge, MA: Author.

Kalbfleisch, J. D., and Prentice, R. L. (1980), *The Statistical Analysis of Failure Time Data*, New York: Wiley.

Lehmann, E. L. (1998), *Nonparametrics: Statistical Methods Based on Ranks* (revised 1st ed.), Upper Saddle River, NJ: Prentice Hall.

Oster, R.A. (2002), "An Examination of Statistical Software Packages for Categorical Data Analysis Using Exact Methods". *The American Statistician*, 56, 235–246.