18th World IMACS / MODSIM Congress, Cairns, Australia 13-17 July 2009
http://mssanz.org.au/modsim09

Power comparisons of parametric and rank tests: grouped outcomes with zero-spike

Hudson, H. M.[1]

[1] *NHMRC Clinical Trials Centre, University of Sydney*
Email: malcolm@ctc.usyd.edu.au

Abstract: Randomized controlled trials are increasingly examining preference data on the minimum benefit required to make treatment worthwhile. These studies assess patients' time and probability trade-offs as outcomes. A comparison is required between two treatment groups, as randomised in the trial. Such data often has specific features making use of standard tests of group differences inappropriate. These include:

- a 'zero spike', a substantial number of patients judging a trivial survival benefit sufficient;
- some patients judging the maximum possible benefit insufficient to make treatment worthwhile;
- a tipping point, i.e. a minimum survival benefit, specific to the patient, making treatment worthwhile, and which determines the patient's elicited response;
- grouping of minimum benefit judged sufficient, with outcomes limited to the specific categories (such as "3 months") offered patients in structured interviews.

Several alternative two-sample tests have been used with such data, but all of them potentially return invalid P-values because of these features, incompatible with test assumptions. We report on simulation modelling of the adequacy of bootstrap correction in estimating P-values and power of the alternative procedures.

In simulation studies we demonstrate that underlying continuous latent variable, ordinal discrete survival, and mixture distribution models can provide the required comparisons. Corresponding test approaches, parametric and non-parametric, are described. These tests may differ in bias and power. Under models with latent variable determining preference, our findings are of little bias in nominal P-values of parametric and rank tests considered. Substantial power differences are demonstrated. The superior choice of test is found to depend on the form of model alternative considered. Insight into appropriate choice of test is gained from consideration of location-shift and polarised alternatives.

In particular, the commonly used normal scores and Wilcoxon-Mann-Whitney tests share good performance under translation shift alternatives. However, these tests exhibit poor power in a sample where responses are drawn from two distinct distributions. In such heterogeneous samples, permutation t-test and logrank tests exhibit higher power.

Keywords: *Bias, binned counts, grouping, ordinal data, patient preference data, permutation test, power, time trade-off, rank test, Wilcoxon-Mann-Whitney test.*

1 INTRODUCTION

1.1 Survival trade-off outcomes

In cancer studies, preferences between treatments may depend on trading off discomfort and inconvenience of treatment for enhanced survival

For example. in a recent study, Blinman et al. (2008), two forms of outcome measure were considered:

- time trade-offs (TTO) – offering extra survival time
- probability trade-offs (PTO) – offering higher probability of survival

To elicit an individual's TTO and PTO, the minimum survival benefit necessary to make the discomfort and inconvenience of chemotherapy worthwhile were established. Trial participants who had experienced a full course of chemo were afterwards asked a series of questions designed to determine survival benefits that would offset the discomfort and inconvenience of their chemotherapy. The 5 year TTO questions were of the general form:

> Imagine you knew that:
> - without chemotherapy life expectancy is 5 years, and;
> - with chemotherapy, life expectancy is 5 years and *6 months*.
> In other words, having chemotherapy would increase life expectancy by *6 months*.
> Based on your own experiences of chemotherapy, which would you prefer?

The 6 month extra survival benefit quoted is subsequently replaced by lower and higher values, ranging from 1 day to 20 years (and permitting the patient to nominate that no survival gain would suffice), in a carefully designed sequence of questions. The result is a single survival benefit threshold, the patient's TTO. A similar series of questions was used to elicit the TTO with a better prognosis, life expectancy baseline of 15 years, and further series of questions for PTOs with baseline probabilities of 65% and 85% of survival exceeding 5 years and 15 years.

In this paper, we address statistical methods for group comparisons of assessed TTOs and PTOs. Since analysis methods for PTOs are identical to those for TTOs, we restrict attention henceforth to TTOs. We address tests that have become accepted in the clinical literature, and the power of these methods. We propose alternative tests that can provide better power to detect differences.

To simplify discussion, and because individual outcomes were strongly correlated for 5 year and 15 year baselines, we shall use a TTO summary calculated as the sum of 5 year and 15 year TTO (i.e. twice the mean TTO) for each patient in Figure 1. The distribution of this aggregate TTO is shown in this Figure.A responses of the maximum survival benefit offered being insufficient was coded as a TTO of 25 years; identical responses of insufficient benefit on both 5 and 15 year baselines would lead to an aggregate TTO of 50 years (i.e. an average 25 year TTO) in the Figure.

The Figure makes clear the features leading to statistical difficulties: zero-spike, high level of aggregation in outcomes, and the skewed distribution with relatively few patients requiring higher survival benefits in order to make chemotherapy worthwhile. The finding is typical of TTOs determined in earlier studies. In Duric et al. (2005) 50-70% of women judged a 1% improvement in 5 year survival rates or a 3 month improvement in life expectancy sufficient to make 4-6 months of adjuvant chemotherapy worthwhile.

1.2 Issues in statistical analysis of survival trade-offs

Consider T, the survival gain required for treatment to be judged worthwhile. There are various analysis perspectives for the distribution of T that may be adopted:

- underlying continuous outcome;
- ordinal discrete (survival categories, e.g, 'low-realistic');
- mixture distribution, where models are developed for two distinct groups: women not willing to trade any

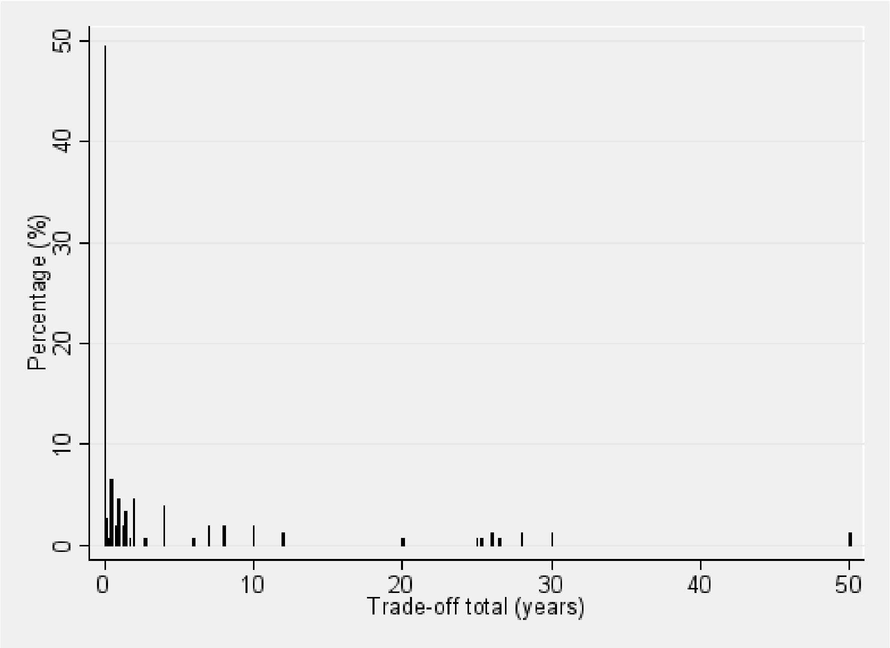

Figure 1: TTO categories and frequencies of outcomes

survival advantage for enhanced quality of life, (so T=0, discrete) and continuous (T>0) outcomes.

For continuous TTO inference, two-sample tests, nonparametric permutation and rank tests are available to evaluate treatment differences. Specific approaches include t- test, following log-transformation (an ad hoc approach), permutation t-test, and Wilcoxon-Mann-Whitney and Normal scores (Simes and Coates, 2001) rank tests.

The transformation of aggregate TTO is important to parametric and permutation tests, while rank tests are invariant to (monotone) transformation of trade off T. Rank methods require an appropriate method of breaking ties, which occur frequently because of the categorical nature of assessed response.

Since time trade-offs are elicited in categories (1 month, 3 months, etc.), outcomes form discrete distributions. Outcome categories can be *pre-assigned* scores and tests conducted on these *scores*. A t-test might score TTO, T, as $\log(1+T)$, while rank tests use as scores the order statistics under a pre-specified distribution (as with Normal scores).

For example, consider Figure 2. With a high proportion of women opting for the smallest survival benefit available, 1 day, the log transformation strongly distinguishes such women from those requiring slightly larger benefits (e.g. 1 month). This effect was undesirable, given such differences represented a clinically negligible benefit. The effect is reduced by offset in transformation; we adopted $\log(1+T/0.25)$, where 0.25 years represents a clinically meaningful survival benefit of treatment.

With zero spike, discrete and skewed distributions, P-values based on standard assumptions or asymptotics must be in doubt (Kolassa, 1995; Lesaffre et al., 1993). Non-parametric tests, for example, require the assumption of data with a continuous distribution, for which ties do not occur. Valid P-values may still be obtained, with some computational effort, by bootstrap resampling of the distribution of the test statistic. For data above, the precise P-value of a t-test on log-transformed outcomes was estimated in 10,000 bootstrap resamples. The t-test's bootstrap distribution tail area, P=0.07 in this instance, may be compared with the nominal P-value, P=0.07 also, which assumes normal distributions of (continuous) transformed outcomes. The common P-value is close to statistical significance.

Perhaps surprisingly, rank tests of the difference between treatments in this data, provide much larger P-values, exceeding 0.5, providing no indication of differences in TTO between these two treatment groups. This too is an often observed phenomenon, one occurring in previous studies comparing TTOs. Additionally, in a recent study, Leung (2007) established reasonable consistency of Normal scores, linear regression (untransformed t-test) and ordinal regression tests of treatment differences. Our aim in this paper is to evaluate the power

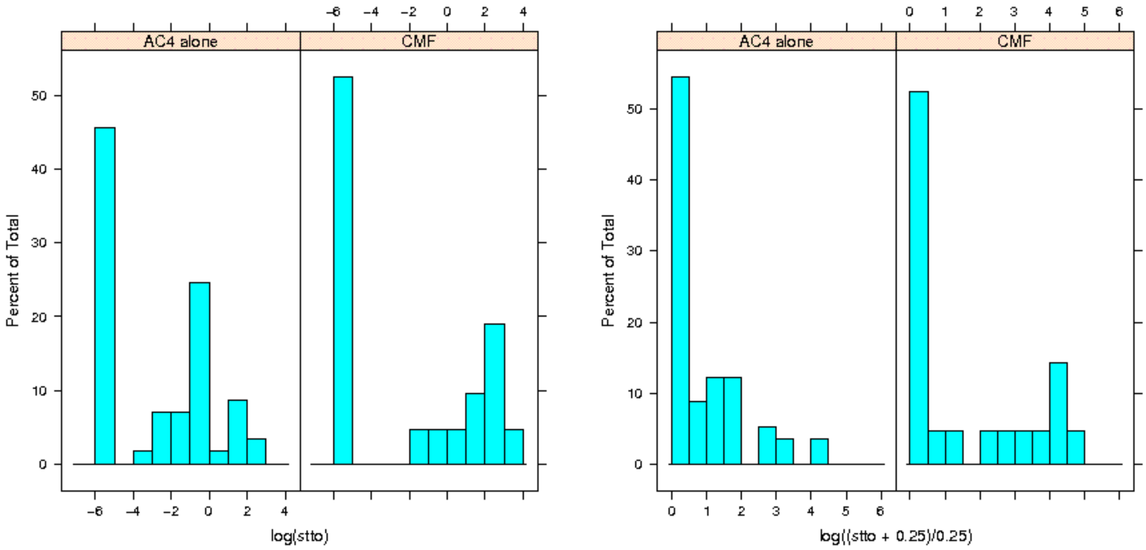

Figure 2: Comparisons by scores of two treatment groups: AC4 (n=57) and CMF (n=21). Displayed by scoring: (a) logT; (b) log[(T+0.25)/0.25].

of parametric, permutation and rank tests in two-sample simulations under specific alternatives where various treatment differences occur. We thereby formulate an explanation of when and why differences in power occur, one that may assist in choice of appropriate tests for TTO (and PTO) data.

2 SIMULATION STUDY

2.1 Approach and methods

A selection of two-sample, permutation and rank tests were considered:

1. permutation t-test on log(1+TTO/0.25) scores;
2. Wilcoxon-Mann-Whitney (rank sum) test, abbreviated as Wilcoxon RS in Tables;
3. Normal scores (rank) test;
4. Exponential scores (Savage rank) test.

All rank test statistics allowed for ties by employing a standard approach using average scores, e.g. using mid-ranks of tied observations with the Wilcoxon-Mann-Whitney statistic. The approach is commonly used as the default method for calculating such test statistics in statistical software. It equates to Efron's method with the logrank test.

A permutation test compares the value of the test statistic with those for random reallocations of the original data to the two groups. The permutation distribution of a test statistic serves to assess the significance of the observed test result. The sampling distribution is that of a random sample, without replacement, from a finite population of scores. Like a rank test, the test and corresponding P-value are distribution free. Unlike a rank test, the test statistic changes on transformation of data (recorded scores matter). The permutation distribution provides a conditional test, applicable even when the parametric assumptions of the t-test may not hold.

Validity of P-values under the null (independent samples from a discrete TTO data distribution) was assessed using a number of approaches, approximate and exact:

- as reported by standard statistical software, applying tests under inapplicable assumptions;

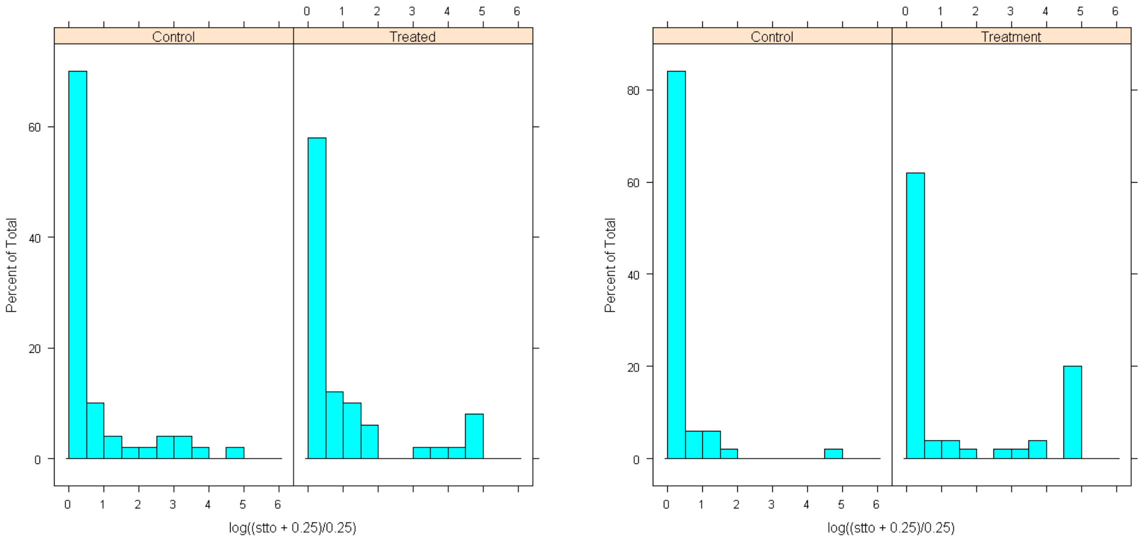

Figure 3: TTO distributions in location-shift (on logs, left) and polarised alternatives (right).

- P-values based on asymptotic normality of sampling distribution of scores (Lehmann, 1975);
- P-values obtained using Fisher's permutation distribution of each sample statistic, a test conditional on the observed data sample.

The latter were considered to be the gold standard. The exact conditional size (under the null hypothesis) and exact conditional power (under an alternative), for tests with rejection cutoff determined for nominal significance levels given below, using the normal sampling distributions, were estimated by crude Monte-Carlo method (generation of 10,000 permuted data sets). The method is similar to those discussed in Hilton and Mehta (1993) and Rabbee et al. (2003).

TTOs in two groups of equal sample size (n=m=50) were generated by grouping log-Normally distributed random numbers. These two groups correspond to women termed untreated (or controls) and treated. N=10,000 simulation samples were generated, and the empirical distribution of test results recorded.

Under the null, both groups were generated from the same underlying distribution, and the type 1 error level (α) of each test estimated in simulations at nominal test levels $\alpha = 0.1\%, 1\%, 5\%$ and 10%. Power comparisons were then conducted for two types of alternative, where TTO increased for treated women.

The first form of comparison was a location-shift alternative to the log-normal distribution of underlying TTO. This is appropriate when treatment effect is expected to change latent TTO by a common scale factor in all treated women. The asymptotic power properties (Hajek et al., 1998) of rank tests are derived under location-shift alternatives of continuous distributions. Observed TTO categories were defined by grouping the latent log-Normal random numbers in fixed intervals to form the discrete distribution of TTO. Different category cut-points and shifts in distribution were chosen to match the characteristics of study data.

The second form of alternative (termed *polarised*) was simulated by specifying the distribution of TTO as a mixture of those women content with very small survival benefits (as indicated by the randomly generated TTO) and other women requiring higher benefits, for whom the benefit required under treatment doubled from that benefit required by a control.

Figure 3 displays histograms of simulated TTO samples in one of N=10,000 simulation replicates. Note polarised alternatives retain a similar proportion of women satisfied with negligible benefits, and another group with much higher TTOs, leading to U-shaped distributions of TTO in the treated group.

Table 1: Rejection rates under the null versus nominal significance

Equal sample sizes	Effect: NULL			
N=100	Rejection rate			
Test	%	%	%	%
α-level	0.1	1	5	10
	%	%	%	%
Wilcoxon RS	0.06	0.92	5.0	9.9
Normal scores	0.07	0.90	5.0	9.9
(unconditional)	0.10	0.98	5.0	9.8
Logrank (exponential scores)	0.08	1.00	4.7	9.5
t-test (permutation)	0.02	0.69	4.7	9.8
(unconditional)	0.02	0.70	4.6	9.7

Table 2: Power under two alternatives: Location Shift alternative (left); Polar alternative (right)

Equal sample sizes	Effect: SHIFT			0.5*SD
N=100	Rejection rate*			
Test	%	%	%	%
α-level	0.1	1	5	10
	%	%	%	%
Wilcoxon RS	14	36	62	73
Normal scores	14	37	63	74
(unconditional)	15	38	63	74
Logrank (exponential scores)	13	32	57	68
t-test (permutation)	6	25	50	63
(unconditional)	7	25	50	63
*N=10000 replicated data sets				

Equal sample sizes	Effect: POLARISE			2.0*SD
N=100	Rejection rate*			
Test	%	%	%	%
α-level	0.1	1	5	10
	%	%	%	%
Wilcoxon RS	0.6	5	15	24
Normal scores	1.2	8	22	32
(unconditional)	2	8	22	32
Logrank (exponential scores)	5	21	43	57
t-test (permutation)	6	25	50	63
(unconditional)	10	36	63	75
*N=10000 replicated data sets				

2.2 Results

Type 1 error rates

The results of Table 1 compare α-levels of various tests, both conditional on the observed sample (Fisher's permutation approach) and unconditional. Conditional and unconditional P-values were found to be similar, as expected in large sample asymptotics. We note little evidence that the skew distributions and aggregation of continuous latent TTOs biased the *nominal* significance reported by standard statistical software. If anything, tests appear a little more conservative than their nominal level. This finding permits powers of tests to be calculated and compared using nominal assessment of significance.

Power

Power for a location shift alternative and polarisation alternative are provided in Table 2. The powers achieved for the sample sizes and alternatives simulated are moderate, reaching a little over 50% for tests conducted at level $\alpha = 0.05$, and up to 75% at level $\alpha = 0.10$. However, interest lies in the substantial differences in power of different tests. While slightly less powerful against the shift alternative, logrank and t-tests substantially outperform more commonly used Wilcoxon rank sum and Normal scores tests.

3 CONCLUSIONS

We found nominal type 1 error rates (based on finite sample asymptotics) to be reliable for survival TTO data. Commonly used methods, the normal scores and Wilcoxon-Mann-Whitney tests, share good performance under translation shift alternatives. However, they exhibit poor power in heterogeneous groups, relative to permutation t-test and logrank tests. Heterogeneity may result from a mixture model in which sample responses are drawn from two distinct distributions. With such data, permutation t-test and logrank tests exhibit higher power. This finding, and the comparison with existing studies, supports further application of these tests and mixture model analysis, to better evaluate treatment effects on TTO distributions with zero-spike. Ad hoc analysis, by unconditional t-test after transformation by log(1+TTO/0.25), has also been found to be unbiased and powerful.

We have restricted reporting, given limitations of space, to two instances of the location-shift and polar alternatives and to the case of equal sample sizes. In these instances providing data with zero-spike from a mixture distribution, we demonstrated that commonly applied tests have poor power with heterogeneous TTOs. Future work will concentrate on extending the scope of the simulation study, evaluating power properties of tests in other polar and mixture model alternatives. In an accompanying paper, we plan to provide further comparisons of various approaches to computing rejection rates, evaluating the accuracy of efficient asymptotic methods in this context.

References

Blinman, P., V. Duric, A. Nowak, P. Beale, S. Clarke, K. Briscoe, A. Boyce, G. Marx, J. Simard-Lebrun, M. Hudson, and M. Stockler. Patients' preferences for adjuvant chemotherapy (ACT) in early colon cancer (ECC): What makes it worthwhile. In *American Society of Clinical Oncology Annual Meeting*, 2008. (abs, poster accepted in the general gastrointestinal session).

Duric, V. M. , M. R. Stockler, S. Heritier, F. Boyle, J. Beith, A. Sullivan, N. Wilcken, A. S. Coates, and R. J. Simes. Patients' preferences for adjuvant chemotherapy in early breast cancer: what makes AC and CMF worthwhile now? *Annals of Oncology*, 16:1786–1794, 2005. doi:10.1093/annonc/mdi370.

Hajek, J., Z.Sidak, and P.K.Sen. *Theory of rank tests*. Academic Press, 2nd edition, 1998. ISBN 0126423504.

Hilton, J. F., and C. R. Mehta. Power and sample size calculations for exact conditional tests with ordered categorical data. *Biometrics*, 49(2):609–616, 1993.

Kolassa, J. A comparison of size and power calculations for the Wilcoxon statistics for ordered categorical data. *Statistics in Medicine*, 14:1577–1581, 1995.

Lehmann, E. *Nonparametrics: Statistical methods based on Ranks*. Holden-Day, 1975.

Lesaffre, E., I. Scheys, J. Frohlich, and E. Bluhmki. Calculation of power and sample size with bounded outcome scores. *Statistics in Medicine*, 12:1063–1078, 1993.

Leung, M. Master's thesis, Biostatistics Collaboration of Australia (BCA), 2007. (Workplace Project Unit).

Rabbee, N., B. Coull, and C. Mehta. Power and sample size for ordered categorical data. *Statistical Methods in Medical Research*, 12:73–84, 2003.

Simes, R., and A. Coates. Patient preferences for adjuvant chemotherapy in early breast cancer: How much benefit is needed? *Journal of National Cancer Institute Monographs*, 30:146–152, 2001.

ACKNOWLEDGEMENTS

The author thanks Ayse Bilgin, Petra Graham and Martin Stockler for careful reading and suggestions that improved this paper.