TESTING FOR A TREATMENT EFFECT IN THE PRESENCE OF NON-RESPONDERS

by

Dennis D. Boos and Cavell Brownie

Institute of Statistics Mimeograph Series 1663

June 1985

Testing for a Treatment Effect in the Presence of Non-responders

by

Dennis D. Boos and Cavell Brownie
Department of Statistics
North Carolina State University
Box 8203, Raleigh, NC 27695-8203, USA

June, 1985

Summary

Good (1979) introduced a new randomization test for the two-sample problem where a proportion p of the treatment group does not respond to the treatment, and suggested that the Wilcoxon test is not effective for this situation. We show to the contrary that the Wilcoxon test is quite useful when $p = .5$ and point out some important deficiencies in his definition of a one-tailed randomization test.

Key Words: Non-responders; randomization test; Wilcoxon test; Pitman efficiency.

1. Introduction

In studies to compare treatment and control groups, it is not unusual for subjects in the treatment group to exhibit greater variability as well as a different mean response. In some situations this increased variability may be due to the presence of subjects in the treatment group who are unaffected by the treatment. For example, in toxicological studies, differences among animals in their tolerance for a given dose level of a compound can result in some animals in the treatment group performing like the controls, while others "respond" to the treatment. Other areas where the non-response to treatment of some subjects has been observed include nutritional supplementation trials (Garby, Irnell, and Werner, 1969) and behavioral toxicology (Nation et al., 1984).

The problem of detecting a treatment effect when the treatment group contains non-responders was considered by Good (1979) who proposed the following statistical model. Let $X_1,\ldots,X_m$ be the responses of the control group, assumed independent with common distribution function $F(x)$, and let $Y_1,\ldots,Y_n$ be the responses of the treatment group, independent with common distribution function $G(x)$. The testing situation is then

$$\begin{aligned} &H_0: \ G(x) = F(x) \\ &\text{vs.} \qquad\qquad\qquad\qquad\qquad\qquad\qquad\qquad (1.1) \\ &H_A: \ G(x) = pF(x-\Delta) + (1-p)F(x) \ , \end{aligned}$$

so that under H_A an average proportion p of the treatment group is shifted by Δ compared to the control group. If we further specify $\Delta > 0$, then the test is one-sided.

Good (1979) suggested a randomization test for (1.1) based on the family of test statistics

$$\nu(\theta) = \theta\left(\frac{1}{m} + \frac{1}{n}\right)^{-1}(\bar{X} - \bar{Y})^2 + (1 - \theta)\sum_{i=1}^{n}(Y_i - \bar{Y})^2, \quad 0 \le \theta \le 1,$$

where $\bar{X}$ and $\bar{Y}$ represent the means for the control and treatment groups, respectively. Note that setting $\theta = 1$ yields Fisher's randomization t-test. Good (1979) suggested using $\theta = .67$ with the idea that $\nu(.67)$ will be sensitive to both an increase in variance and a mean shift in the treatment group when $p < 1$ and $\Delta \neq 0$.

We believe that Good's paper addresses a real problem, and his test statistic appears effective when p is small. However, aspects of his paper are misleading or incorrect, and because his test is now being used by toxicologists (eg. Weiss, 1980, Cory-Slechta et al., 1980, Cox, 1981, Nation et al. 1984), we think it is important to point out the following deficiencies:

1) Good suggests that the Wilcoxon test is not effective for the testing situation (1.1). We show in Section 2 that this assertion is incorrect at least for $p \ge .6$, and is apparently based on a misunderstanding of Pitman efficiency calculations found in Hodges and Lehmann (1956).

2) Good's test statistic is not readily adapted to carrying out one-tailed tests, and his method for performing a one-tailed randomization test is not valid in general. In Section 3 we explain why this is so and describe how to get an appropriate one-sided P-value.

3) Good's statistic is intuitive but is also subject to misinterpretation in the case where the treatment group has more variability, but the same mean, as the control group. Thus, one might interpret rejection by $\nu(.67)$ as indicating a mean shift when it is due solely to a large variance in the

treatment group. Results in Table 1 of Section 4 show that standard tests such as the t-test do not lend themselves to this misinterpretation.

The Monte Carlo study which produced the results referred to above is described in Section 4, and Section 5 contains a summary with examples.

2. Pitman Efficiency and the Wilcoxon Test

In his introduction Good states that the power of the t-test under (1.1) and F=normal is reduced for $p < 1$ and that "the Wilcoxon test fares even worse." He supports this latter conclusion about the Wilcoxon test by referring to Pitman efficiency calculations of Hodges and Lehmann (1956). However, these calculations are only relevant to the case where p is small (approaching 0 as $n \to \infty$ for the Pitman calculations). If p is held fixed and Δ goes to 0 as $n \to \infty$, then the usual Pitman relative efficiency between the Wilcoxon test and the t-test obtains, i.e.,

$$e(\text{Wilcoxon},t) = 12\sigma_F^2 \left[\int f^2(x)dx\right]^2, \qquad (2.1)$$

where σ^2_F = Variance of F and f is the density of F. When F is normal, then (2.1) is $3/\pi = .955$.

Empirical results from the Monte Carlo study described in Section 4 further demsonstrate the inaccuracy of Good's statement for small samples. In Table 2 which gives empirical powers for a shift of $\Delta = 1$ for $m = n = 8$, and F = standard normal, we find empirical efficiencies of .94, .97, and .98 for $p = 1$, .8, and .6, respectively. These empirical efficiencies are the ratios of the empirical power of the Wilcoxon test to that of the t-test after the Wilcoxon powers were slightly reduced to correspond to exact $\alpha = .05$ level tests. For the case $m = n = 20$ and $\Delta = .6$, analogous empirical efficiencies calculated from

the results in Table 3 are .95, .99, and .97, respectively. Thus, for F normal, the performance of the Wilcoxon test is roughly the same as the t-test and even seems to be improving relative to the t-test as p goes from 1.0 to 0.6. For the other distributions in Tables 2 and 3, the Wilcoxon is superior to the t-test.

One intuitive reason for the Wilcoxon test to perform well when $p \in [.6,1]$ is that nonresponders in the treatment group appear as outliers and the Wilcoxon and other robust tests tend to downweight such observations. Of course, when p is small then the responders appear as outliers and the Wilcoxon will ignore them and tend to declare in favor of H_0.

3. One-tailed Randomization Tests

To discuss the calculation of randomization significance levels or P-values, additional notation is needed. Let $\bar{X}_0$, $\bar{Y}_0$, $\nu_0(\theta)$ represent statistics calculated from the original observations $X_1,\ldots,X_m$ and $Y_1,\ldots,Y_n$, while $\bar{X}_r$, $\bar{Y}_r$, $\nu_r(\theta)$ represent these statistics calculated from a partitioning of $(X_1,\ldots,Y_m, Y_1,\ldots,Y_n)$ into two samples of sizes m and n.

For a two-tailed test of (1.1) based on $\nu(\theta)$, the randomization P-value associated with an observed $\nu_0(\theta)$ is the proportion of the $\binom{m+n}{n}$ partitions for which $\nu(\theta)$ is at least as large as $\nu_0(\theta)$. We write this as

$$P = \#\{\nu_r(\theta);\ \nu_r(\theta) \geq \nu_0(\theta)\}/\binom{m+n}{n} .$$

For a one-sided test of (1.1) with $\Delta > 0$, Good recommends the following. "Compute ν for each (randomly selected) partition of the original observations. ...For a one-tailed test, reject the null hypothesis if the treatment mean exceeds...the control mean and if ν based on the original observations exceeds a proportion two alpha of all computed values." Undoubtably, the intent in

this last sentence was "...exceeds a proportion $(1 - 2\alpha)$...," but this is only a minor point. Our main criticism concerns Good's use of all $\binom{m+n}{n}$ partitions (rather than only those for which $\bar{Y}_r \succ \bar{X}_r$) as a reference set to determine if $\nu_0(\theta)$ supports H_A. We now try to explain this point in terms of P-values.

The one-sided P-value based on Good's procedure is

$$P_{Good} = \tfrac{1}{2}[\#\{\nu_r(\theta);\ \ \nu_r(\theta) \geq \nu_0(\theta) \text{ and } \bar{Y}_r \succ \bar{X}_r\} \qquad \text{if } \bar{Y}_0 \succ \bar{X}_0$$

$$+ \#\{\nu_r(\theta);\ \ \nu_r(\theta) \geq \nu_0(\theta) \text{ and } \bar{Y}_r \preccurlyeq \bar{X}_r\}]/\binom{m+n}{n}$$

$$\geq \tfrac{1}{2} \text{ otherwise.}$$

Thus, he has just taken the two-sided randomization P-value and used half of its value whenever $\bar{Y}_0 \succ \bar{X}_0$. When $\theta = 1$ and $m = n$ this method is correct because $(\bar{Y}_r - \bar{X}_r)^2$ has the same distribution over the partitions where $\bar{Y}_r \succ \bar{X}_r$ as it does for those where $\bar{Y}_r \prec \bar{X}_r$. However, when $0 \prec \theta \prec 1$, the permutation distribution of $\nu_r(\theta)$ is different over those two sets of partitions. A more appropriate one-sided P-value, which uses as a reference set those partitions for which $\bar{Y}_r \succ \bar{X}_r$, is

$$P = \#\{\nu_r(\theta);\ \ \nu_r(\theta) \geq \nu_0(\theta) \text{ and } \bar{Y}_r \succ \bar{X}_r\}/\binom{m+n}{n} \qquad \text{if } \bar{Y}_0 \succ \bar{X}_0$$

$$\geq \tfrac{1}{2} \text{ otherwise .} \qquad (3.1)$$

The problem with Good's method for one-sided inference is not easily detected when estimating test power by Monte Carlo methods from *repeated samples* of original observations from *symmetric distributions.* This is because with $m = n$ the *average* permutation distributions over the two sets of partitions $\{\bar{Y}_r \succ \bar{X}_r\}$ and $\{\bar{Y}_r \prec \bar{X}_r\}$ are equal and one would not suspect a problem for individual samples. However, when F is asymmetric then equality

of the permutation distributions may not hold even in an average sense. For F skewed to the right as in the lognormal distribution used by Good or the extreme value distribution that we used, there is a tendency for $\Sigma(Y_i - \bar{Y})^2$ and hence $\nu(\theta)$ to be large when $\bar{Y} > \bar{X}$. Among all partitions for a skewed sample there is a positive association between $(\bar{Y}_r > \bar{X}_r)$ and $\nu_r(\theta)$ large; the result is that P_{Good} is too small. In Monte Carlo sampling H_0 is rejected too often since skewed samples appear frequently. This result is evident in Good's Table 2 which contains empirical powers for a one-tailed $\alpha = .05$ level test when F is lognormal. In the null case (p = 0.0) the estimated size of $\nu(.67)$ is .083, whereas Fisher's randomization t ($\equiv \nu(1)$) has estimated size .046 which is appropriately close to .05. We obtained similar results with the extreme value distribution, where in Table 1 of Section 4 the estimated size of Good's one-tailed test is .085 for a nominal level of .05. In contrast, note that our method using (3.1) to calculate P-values resulted in an estimated size of .048 for the same situation.

In actual practice, when m and n are not small, say >6, the P-values are usually estimated by sampling N partitions from the possible partitions. The change in the P-value formula is just to replace $\binom{m+n}{n}$ by N in (3.1). When $m = n$ and $\#\{\bar{Y}_r = \bar{X}_r\} = 0$, a more efficient estimate of the P-value is given

$$\hat{P} = \tfrac{1}{2}\#\{\nu_r(\theta);\ \nu_r(\theta) \geq \nu_0(\theta) \text{ and } \bar{Y}_r > \bar{X}_r\}/\#\{\bar{Y}_r > \bar{X}_r\} \quad \text{if } \bar{Y}_0 > \bar{X}_0$$

$$\geq \tfrac{1}{2} \text{ otherwise.}$$

4. Monte Carlo Results

A Monte Carlo study was conducted for sample sizes $m = n = 8$ and $m = n = 20$ and for three different distributions F: normal, t distribution with 3 degrees of freedom (t_3), and extreme value distribution $F(x) = \exp(-\exp(-x))$.

The t_3 and extreme value distributions were included to provide information concerning robustness in the presence of heavy tails or moderate skewness.

Samples were generated from G and F in (1.1) for each type of F and for $p = 1.0, 0.8$, and 0.6 and for several values of Δ. When $\Delta > 0$ and $p = .8$ or $p = .6$ the random variable Y_i from the mixture distribution G was obtained by generating a random variable Z_i from F, and setting $Y_i = Z_i + \Delta$ or $Y_i = Z_i$ according to the outcome of an independent Bernoulli trial with success probability p.

We describe first some results for the null situation when $m = n = 8$. In addition to $\Delta = 0$ we included a Behrens-Fisher type null situation where F and G have the same mean but different variances. Table 1 gives the

--- Insert Table 1 Here ---

estimated α levels for 7 different one-tailed tests when the nominal α is set to .05. The tests are 1) the usual pooled t-test, 2) the unequal variances t-test used by SAS (see Ray, 1982, p. 219) with Satterthwaite approximation for degrees of freedom, 3) the t-test based on trimmed means with one observation trimmed from each end of each sample (see Yuen and Dixon, 1973), 4) the unequal variances analogue for the trimmed t, 5) the Wilcoxon rank sum test, 6) Good's statistic $\nu(.67)$ with his method of obtaining a one-tailed test by randomization, and 7) $\nu(.67)$ with the one-tailed test obtained by our method explained in Section 3. Each entry in Table 1 is based on 1000 Monte Carlo replications and has approximate standard error $[(.05)(.95)/1000]^{1/2} = .007$. The P-values for the $\nu(.67)$ calculations were estimated by using 500 partitions randomly selected with replacement from the $\binom{16}{8} = 12870$ total partitions. This latter randomness for each Monte Carlo replication tends to average out when forming the estimated power from all replications and adds only a

negligible amount to the .007 standard error mentioned above.

Columns 1, 4, and 6 of Table 1 suggest that all the tests hold their level fairly well when G = F except for $\nu(.67)$ with Good's method of obtaining a one-tailed test. The exact level of the Wilcoxon test is .052 when G = F, and for the pooled t the exact level is of course .05 for G = F = normal. Columns 2, 3, 5, and 7 of Table 1 are for the Behrens-Fisher situation where the means are equal but the variances are different. Here we see that $\nu(.67)$ is sensitive to different variances as expected from the nature of the test statistic. The other tests hold their levels fairly well except for the last column where the samples are from the extreme value distribution. It appears that the unequal variances t and its trimmed analogue are not sufficiently different from their pooled counterparts to be of much use. We might add that Hettmansperger (1984, p. 174-175) has shown that the asymptotic level of the Wilcoxon test in the Behrens-Fisher problem for symmetric densities ranges between .05 and .087.

--- Insert Table 2 Here ---

In Table 2 we show results for only four statistics: the pooled versions of the t and trimmed t, the Wilcoxon, and Good's $\nu(.67)$ with our method of obtaining a one-tailed test. G and F are given by H_A of (1.1) with $\Delta/\sigma_F = 1$ (σ_F^2 = variance of F) and p = 1.0, 0.8, and 0.6. The pooled trim and the Wilcoxon perform quite similarly and both outperform the t-test at the t_3 and are outperformed by the t-test at the normal. The Wilcoxon is a slight winner at the extreme value distribution. The power of Good's $\nu(.67)$ is never greater than that of the Wilcoxon although the latter should be adjusted downward by .01 for each entry to correspond to a test with α level exactly .05. As mentioned in Section 2, the ratios of powers of the Wilcoxon to the t change

little as p moves from 1.0 to 0.6.

---Insert Table 3 Here ---

Results for the case m = n = 20 are given in Table 3. We have omitted $\nu(.67)$ because Good himself states that it is most effective for small m and n. Power estimates in the null case are based on 10,000 Monte Carlo replications, and the entries for $\Delta/\sigma_F = .6$ are based on 4,000 replications. The Wilcoxon level used is actually .047, and the pooled t and pooled trim (3 from each end deleted) hold their levels well. The same patterns hold as for the case m = n = 8 except that the superiority of the Wilcoxon test relative to the t-test is heightened at the t_3 and extreme value distributions.

5. Examples and Summary

Figure 2 of Good (1979) contains an example concerning the effect of an antibiotic on vaginal titres after injection with Herpes virus. The last column of that figure is a case which appears to have one nonresponder out of five treated animals or p = .8. Good reports two-tailed P-values of .02 for the t-test and .012 for $\nu(.67)$ (note correction to Table 4 in *Biometrics*, 1980, p. 751). Applying the Wilcoxon test to these data, we obtained a rank sum for the larger group of 39 with an exact two-sided P-value of 2(.008) = .016. Thus, evidence of a treatment effect is only marginally stronger with Good's test than with the Wilcoxon or t.

The second example, from Nation, et al. (1984), concerns the effect of cadmium exposure on a passive avoidance measure in adult rats. Figure 1 of their paper gives the number of platform descents for a goup of 9 control rats, a second group of 9 rats treated daily with 1 mg/kg of body weight (Cd-1), and a third group of 9 rats treated daily with 5 mg/kg of body weight (Cd-5). The authors performed a median test with the significance level

"controlled experimentwise" and concluded that the Cd-5 group descended less frequently than the controls. We obtained one-sided P-values from the median test for control versus Cd-5 of .028 and for control versus Cd-1 of .17. The authors then said they performed Good's test and obtained P-values of .03 and .20. Their final conclusion was that only the Cd-5 group differed from the control group. Our analysis of the data (read from their Figure 1) yielded exact one-sided P-values from the Wilcoxon rank sum test of .004 for control versus Cd-5 and .025 for control versus Cd-1. A t-test yielded one-sided P-values of .002 and .02 respectively. We then ran the data through our program for Good's test using $\nu(.67)$ and got one-sided P-values of .0007 and .012, respectively, based on 4000 random partitions. Thus, the authors used the median test which tends to be inefficient, especially in small samples (see Ramsey, 1971), and then apparently computed Good's test incorrectly. The result was an incorrect inference in a situation where the t-test and Wilcoxon test lead to essentially the same conclusion as Good's test (correctly applied).

In both of these examples Good's test appears to be effective but so do the Wilcoxon test and the t-test. Moreover, the latter are much easier to use and also simpler to interpret because they do not mix variance differences with mean differences. In conclusion we feel that when the proportion of responders is likely to be .6 or above, the Wilcoxon will generally be a good choice for detecting a treatment effect.

Table 1. Empirical Power of One-tailed Tests for Several Null Situations ($\Delta = 0$), $m = n = 8$, $\alpha = .05$

	Normal			t_2		Extreme Value	
	$\sigma_G^2=\sigma_F^2$	$\sigma_G^2=2\sigma_F^2$	$\sigma_G^2=4\sigma_F^2$	$\sigma_G^2=\sigma_F^2$	$\sigma_G^2=2\sigma_F^2$	$\sigma_G^2=\sigma_F^2$	$\sigma_G^2=2\sigma_F^2$
Pooled t	.046	.050	.060	.045	.063	.042	.086
Unpooled t	.046	.046	.057	.045	.061	.039	.083
Pooled trim	.044	.056	.061	.046	.061	.052	.079
Unpooled trim	.042	.054	.056	.046	.058	.047	.073
Wilcoxon	.049	.052	.069	.053	.063	.059	.083
Good (1979)	.054	.107	.236	.048	.108	.085	.189
Good (new method)	.057	.097	.202	.056	.097	.048	.109

Note: The pooled and unpooled trim tests are based on deleting one observation from the end of each ordered sample. Each entry is based on 1000 Monte Carlo samples and has standard error $\approx .007$.

Table 2. Empirical Power of One-tailed Tests When Responders are Shifted by 1 Standard Deviation Unit ($\Delta = \sigma_F$), $m = n = 8$, $\alpha = .05$.

		Normal	t_2	Extreme value
Pooled t	p = 1.0	.62	.71	.61
	p = 0.8	.43	.56	.46
	p = 0.6	.27	.37	.31
Pooled trim (1 obs. deleted from each end)	p = 1.0	.57	.79	.60
	p = 0.8	.39	.60	.46
	p = 0.6	.26	.39	.32
Wilcoxon	p = 1.0	.59	.78	.64
	p = 0.8	.42	.59	.49
	p = 0.6	.27	.39	.33
Good (new method)	p = 1.0	.53	.67	.50
	p = 0.8	.39	.54	.38
	p = 0.6	.27	.38	.28

Note: Each entry is based on 1000 Monte Carlo samples and has standard deviation of at most $(4000)^{-\frac{1}{2}} = .016$.

Table 3. Empirical Power of One-tailed Tests, m = n = 20, α = .05.

		Normal		t_3		Extreme Value	
	Δ/σ_F	0	.6	0	.6	0	.6
Pooled t	p = 1.0	.050	.59	.047	.65	.046	.59
	p = 0.8		.44		.51		.47
	p = 0.6		.28		.34		.29
Pooled Trim	p = 1.0	.049	.54	.049	.80	.047	.60
(3 obs. deleted	p = 0.8		.41		.63		.47
from each end)	p = 0.6		.27		.42		.30
Wilcoxon	p = 1.0	.050	.55	.047	.78	.046	.64
	p = 0.8		.42		.62		.49
	p = 0.6		.27		.40		.31

Note: The null case entries (Δ/σ_F = 0) are based on 10,000 Monte Carlo replications and have standard error $\approx$.002. The others are based on 4000 replications and have standard error $\approx$.0034.

References

Cory-Slechta, D. A., Bissen, S. T., Young, A. M. and Thompson, T. (1981). Chronic postweaning lead exposure and response duration performance. *Toxicology and Applied Pharmachology* 60, 78-84.

Cox, C. (1981). Detection of treatment effects when only a portion of subjects respond. In *Nutrition and Behavior: Proceedings of the Franklin Research Centre's 1980 Working Conference on Nutrition and Behavior.* Sanford Miller Editor. Franklin Institute Press, Philadelphia.

Garby, L., Irnell, L. and Werner, I. (1969). Iron deficiency in women of fertile age in a Swedish community. III. Estimation of prevalence based on response to iron supplementation. *Acta Medica Scandinavica* 185, 113-117.

Good, P. I. (1979). Detection of a treatment effect when not all experimental subjects will respond to treatment. *Biometrics* 35, 483-489.

Hettmansperger, T. (1984). *Statistical Inference Based on Ranks.* Wiley, New York.

Hodges, J. L., Jr. and Lehmann, E. L. (1956). The efficiency of some nonparametric competitors of the t-test. *Annals of Mathematical Statistics* 27, 324-335.

Nation, J. R., Bourgeois, A. E., Clark, D. E., Baker, D. M. and Hare, M. F. (1984). The effects of oral cadmium exposure on passive avoidance performance in the adult rat. *Toxicology Letters* 20, 41-47.

Ramsey, F. L. (1971). Small sample power functions for nonparametric tests of location in the double exponential family. *Journal of the American Statistical Association* 66, 149-151.

Ray, A. A. (1982). *SAS User's Guide: Statistics, 1982 edition.* SAS Institute, Inc., Cary, North Carolina.

Weiss, B. (1980). In Rebuttal. *American Journal of Diseases of Children* 134, 1126-1127.

Yuen, K. K. and Dixon, W. J. (1973). The approximate behavior and performance of the two-sample trimmed t. *Biometrika* 60, 369-374.