Journal of Modern Applied Statistical Methods

Volume 16 | Issue 1 Article 3

2017

An Empirical Demonstration of the Need for Exact Tests

Vance W. Berger
Biometry Research Group, National Cancer Institute, vance917@gmail.com

Journal of Modern Applied Statistical Methods
May 2017, Vol. 16, No. 1, 34-50.
doi: 10.22237/jmasm/1493596920

ISSN 1538 − 9472

An Empirical Demonstration of the Need for Exact Tests

Vance W. Berger
National Cancer Institute
Rockville, MD

The robustness of parametric analyses is rarely questioned or qualified. Robustness, generally understood, means the exact and approximate p-values will lie on the same side of alpha for any reasonable data set; and 1) any data set would qualify as reasonable and 2) robustness holds universally, for all alpha levels and approximations. For this to be true, the approximation would need to be perfect all of the time. Any discrepancy between the approximation and the exact *p*-value, for any combination of alpha level and data set, would constitute a violation. Clearly, this is not true, and when confronted with this reality, the "No True Scotsman" fallacy is often invoked with the declaration it must have been a pathological data set, as if this would obviate the responsibility to select an appropriate research method. Ideally, a method would be selected because it is optimal, or at least appropriate, without needing special pleading, but judging by how often approximations are used when the exact values they are trying to approximate are readily available, current trends do not come close to this ideal. One possible explanation might be that there is not much information available on data sets for which the approximations fail miserably. Examples are presented in an effort to clarify the need for exact analyses.

Keywords: Chi-square test, normality, permutation tests, robustness, t-test

Introduction

Approximations are used rather often, in all sorts of contexts. Sometimes this is because the exact value is not available, or because it could be made available but only at a prohibitive cost. In no case is the approximation ever actually preferred to the exact value it is trying to approximate, for if this indeed is the case, then the approximation is not an approximation. Rather, it would then be calculated for its inherent interest.

This raises the issue of whether parametric analyses are conducted because they are of interest in their own right, or merely as approximations to exact

Vance W. Berger, PhD, is a member of the Biometry Research Group. Email him at: vb78c@nih.gov.

analyses. Though it is conceivable that in certain limited cases there is interest in a parametric analysis, it is clear, when one considers the pre-testing that generally occurs to ensure that the conditions are met to ensure the integrity of the approximation, that the parametric analyses are, in general, just approximations, nothing more. For example, if one were to test the data for normality by any method, even an informal one such as appeal to the fact that we have always just assumed normality, prior to conducting a *t*-test, then this undermines the notion that the *t*-test is conducted for inherent interest. There is interest only conditionally on the finding that the data are normal enough to merit such interest. Along these lines, Bradley (1968) noted that "A corresponding parametric test is valid only to the extent that it results in the same statistical decision [as the exact test]" (p. 85).

We must distinguish two cases here. In one case, the choice is to approximate or not to approximate; but if one does, then one cannot know how well the approximation performed since the exact value cannot be computed. In the other case, the exact value is readily available, so here the choice is to use it or the approximation. Berger (2000) pointed out the folly, in this case, of ever using the approximation. After all, how compelling is a test of normality in allowing for the use of an approximation when one can instead simply compare the two values to see how close they actually are (as opposed to how close they *should* tend to be on average)? But for that matter, given that one already has the exact value, why even consider replacing it with the approximation?

The lapse in logic that would allow a researcher to use an approximation when the very quantity it is trying to approximate is readily available is staggering, and yet this exact situation plays out in a huge number of randomized clinical trials, Bradley's aforementioned sage wisdom notwithstanding. The randomization itself allows for exact comparisons of the treatment groups by way of permutation tests (see, e.g., Fisher, 1935; Rigdon & Hudgens, 2015; Lu, Ding, & Dasgupta, 2015), and yet it is the inexact parametric tests that are used far more often, generally after going through the motions of justifying this choice by first conducting a test of the assumptions that allegedly support the use of the parametric test in question.

The only saving grace would be if it just didn't matter. Sure, the exact analyses are preferable, but given how robust the parametric analyses are, there is very little to gain and much to lose in terms of computing time. This argument may have been compelling decades ago, when it actually would have been difficult to conduct a permutation test, but today this is no longer the case. It is just as easy to do it right as it is to do it wrong. So this leaves us at the other

aspect of this argument, it just doesn't matter (and all the variations of this theme, including the assertion that there are more important issues for statisticians to concern themselves with, as if the choice of an appropriate analysis is somehow beneath the dignity of the very party charged with doing so). Moreover, even if it did not matter (at least numerically), that still would not provide a compelling argument in favor of a theoretically unsound analysis.

This much is clear, and should already suffice to eradicate parametric analyses from actual clinical trials, at least when comparing treatments. Sadly, it has not, and the widespread delinquency of researchers who simply cannot be bothered to concern themselves with the relative merits of various analyses is matched by a commensurate delinquency on the part of those authorities who could impose the need for rigor, yet somehow choose not to. And they do this while assuring patients and funding bodies that only the best research methods will be used. But at least we can fall back on robustness.

Everybody knows that parametric analyses are robust, but how many can actually provide a precise formulation of what that means, operationally? How good is good enough? What does "good enough" even mean in this case? What does convergence as the sample size increases without bound say about the discrepancy for this particular data set with its very finite sample size? These are uncomfortable questions for those who continue to embrace robustness as a justification for using approximations when in fact the exact values should be used instead. One theorem that would be useful in supporting this case would be along the lines of $|p_1 - p_2| < k/n$, where k is some universal constant, n is the sample size, and k/n bounds the absolute difference between the two p-values.

Even if this statement were true, it would still be hard to see how that would justify the substitution of the one for the other. After all, enlightened researchers recognize that each party may apply his or her own personal alpha level to the results of any clinical trial (Berger, 2004). This being the case, how much error is acceptable when, with a different choice, we can attain the ideal of no error at all? Moreover, is such a bound of the discrepancy even true? The remainder of this paper will illustrate that in fact it is not true for any reasonable value of k. We will consider the chi-square approximation to Fisher's exact test, the Smirnov test (both exact and approximate), and the t-test in the sections to follow.

Examples of the Chi-Square Test Failing

When dealing with a single 2×2 contingency table, the two most common tests seem to be Fisher's exact test and the chi-square test. Of course, the chi-square

test is used in other situations as well, and sometimes the exact test to which it is compared is not Fisher's exact test, and in some cases this test may not even have a name (but is easily defined in terms of a test statistic and a permutation mode of inference). Table 1 presents six data sets for which the chi-square *p*-value differs markedly from its exact counterpart. In Example C1, the comparison was the chi-square test to Fisher's exact test. Little (1989) pointed out that each expected cell count was over five, so the usual rule of thumb would have led one to use the chi-square test and find significance at the 0.05 level (note that the *p*-values in the table are one-sided, so Fisher's exact test is not significant).

Table 1. Data sets for which the chi-square test fails badly

N	References	Data Set*	*p*-values**
C1.	Little (1989)	{(170,2);(162,9)}	0.0299, 0.0162
C2.	Zelterman et al. (1995)		0.0424, 0.119
C3.	Cytel Software (1995, p. 11)		0.0013, 0.1342
C4.	Cytel Software (1995, p. 17)	{(3,1);(1,3)}	0.243, 0.0786
C5.	Berger and Lachenbruch (1998)	{(20,230);(35,225)}	0.063, 0.047
C6.	Hewett et al. (1999); Clancy (2000)	{(10,453);(2,364)}	NS***, 0.02

Note: Citations abbreviated for space; see Reference section below for full reference
* Data set provided only for a single 2 × 2 contingency table
** Exact *p*-value first, then chi-square *p*-value
*** Actual *p*-value not reported, nor is the full data set available

Table 2. Data from StatXact (Cytel Software, 1995)

0	7	0	0	0	0	0	1	1
1	1	1	1	1	1	1	0	0
0	8	0	0	0	0	0	0	0

Example C2 is from Table 1 of Zelterman, Chan, and Mielke (1995), which is hypothetical data in the form of two stratified 2 × 2 contingency tables. These were $\{(1, 0); (3, 9)\}$ and $\{(0, 0); (9, 5)\}$. Not only do the *p*-values differ dramatically (the exact *p*-value is 0.0424 and the approximate chi-square *p*-value is 0.119), but in fact it is the exact one that is lower. This example flies in the face of the conventional wisdom that states that permutation tests are always conservative so therefore exact *p*-values are always larger than their approximate counterparts. Zelterman et al. (1995) note "The lesson we learn ... is that the behavior of test statistics, such as Pearson's chi-square, may or may not agree with their asymptotic approximations. The only certain methods for accurate analysis of tables with small counts is to perform exact methods based on the

likelihood function" (p. 358) Example C3 is based on a sparse 3 × 9 contingency table presented in the StatXact manual (Cytel Software, 1995, p. 11), which is reproduced in Table 2.

Pearson's chi-square test of an interaction between rows and columns has a test statistic value of 22.29 with (3 – 1)(9 – 1) = 16 degrees of freedom, for a *p*-value of 0.1342. Using the same test statistic, specifically the chi-square test statistic, but using its exact distribution instead of the distributional assumption results in an exact *p*-value of 0.0013. As in Example C2, not only are the *p*-values (and the interpretations one would arrive at) grossly different from each other, but in fact it is the exact one that would demonstrate a true treatment effect (assuming that rows are treatments), whereas the approximate one would miss it. The StatXact manual notes "the need to compute the exact *p*-value, rather than relying on asymptotic results, whenever the data set is small, sparse, unbalanced, or heavily tied. The trouble is that it is difficult to identify, a priori, that a given data set suffers from these obstacles to asymptotic inference" (Cytel Software, 1995, p. 11).

Example C4 is also from the StatXact manual (Cytel Software, 1995, p. 17), and is Fisher's famous original tea-tasting experiment which led to the development of Fisher's exact test. As is well known, the experiment involved testing the claim of a British woman that she was able to distinguish between the two possible orders, milk first and then tea, or tea first and then milk, being poured into a cup. This woman was presented with eight cups of tea, in which four were of each order (and she was told this key fact). The order in which the cups were given to her was randomized. Of the four cups with milk poured first, she guessed right three times. Likewise, of the four cups with tea poured first, she guessed right three times. The chi-square test yields a *p*-value of 0.1573 two-sided or 0.0786 one-sided. The Fisher exact *p*-value is 0.243, which is not even close.

Example C5 regards data presented at the December 15, 1995 FDA Blood Products Advisory Committee meeting. Hospitalization due to a targeted respiratory disease was required by 20/250 (8.0%) patients on a biological treatment arm and 35/260 (13.5%) patients on the control arm. Pearson's uncorrected chi-square test yielded $p = 0.047$ two-sided, and significance was declared at the prospectively specified 0.05 alpha level (two-sided). But the nominal 0.05 alpha level is preserved only if the true probability of a Type-I error is no greater than 0.05. A fair question, then, is how likely one would be to obtain data at least as significant ($p < 0.047$), by using this chi-square test, assuming nothing more than random allocation of patients to treatment groups. The answer, $p = 0.063$, is provided by Fisher's exact test, which of course does not attain

statistical significance at the 0.05 alpha level. The StatXact manual points out that "The term 'asymptotically' means 'given a sufficient sample size', though it is not easy to describe the sample size needed for the chi-square distribution to approximate well the exact distribution of the Pearson statistic" (Cytel Software, 1995, p. 12).

Example C6 is based on Clancy's (2000) letter to the editor regarding Hewett, Lindenfeld, Riccobene, and Noyes' (1999) paper, in which the authors evaluated the effect of neuromuscular training on the incidence of knee injury in female athletes. There were ten injuries among 463 untrained athletes and two injuries among 366 trained athletes. The chi-square test was reported to yield $p = 0.02$. Clancy reported a non-significant p-value with Fisher's exact test, and also pointed out that one cell had both an actual and an expected cell count under five, so that Fisher's exact test would be the more reliable of the two, in keeping with conventional wisdom. Notably, Hewett, Levy, and Noyes (2000) responded to the letter by resorting to appeal to credentials, stating essentially that they used an "excellent" statistician, so therefore whatever he came up with must be correct by virtue of his coming up with it. A second "unbiased" statistician confirmed this.

Even in the absence of a reason for suspicion, suspicion must still arise when an argument is defended by appeal to credentials. This is, after all, tantamount to an admission that there is no better defense for the argument than credentials. One has to wonder just how "unbiased" the second statistician truly was, and also how many competent statisticians (with the fortitude to refuse to sign off on an analysis so poorly planned) were also contacted. Competent statisticians know to use Fisher's exact test when the expected cell counts, or any one of them, is less than five; even better statisticians would recognize the irrelevance of the expected cell counts and instead use Fisher's exact test any time it differs substantially from the chi-square test. And still better statisticians would recognize that they are not in a position to determine how close an approximation needs to be in order that it be preferred to the quantity it is trying to approximate, so they would simply use Fisher's exact test routinely.

Examples of the Approximate Smirnov Test Failing

When dealing with a single ordered $2 \times J$ table, the best test that is offered as a routine option (no programming required) in commercially available software packages is the exact Smirnov test, a standard feature of StatXact. See Section 10.1 of Hollander and Wolfe (1973) and Section 1.6 of Lehmann (1975). Note that while it is customary to speak of the Smirnov test as a two-sided approximate

test, we use this term to denote the exact one-sided version. Essentially, the only difference between the one-sided and the two-sided version is the absence or presence, respectively, of absolute values around the directed difference of CDFs to be maximized. Whether one-sided or two-sided, the approximate test that bears the same name often gives strikingly different p-values from the exact version for the same data set, as we will demonstrate in Table 3. Note that the exact Smirnov test *p*-values (but not the approximate ones) for these data sets appeared in Table 2 of Berger (2002), and some of them seem to contradict what we are presenting now in our Table 3. The reason for this is the newfound ability of StatXact to compute exact Smirnov *p*-values immediately for such large data sets, whereas only a few years ago only Monte Carlo approximations were feasible.

Table 3. Ordered 2 × *J* tables for which the approximate Smirnov test fails badly

N	References	Data Set	*p*-values*
S1.	Fentiman et al. (1983)	{(6,8,4,2,3);(3,2,8,0,10)}	0.0138, 0.0296
S2.	Fox et al. (1993)	{(1,5,16);(0,0,22)}	0.0106, 0.1947
S3.	Fox et al. (1993)	{(12,3,7);(3,7,12)}	0.0108, 0.0252
S4.	Elwood (1998)	{(33,5,545);(29,8,836)}	0.0258, 0.6823
S5.	TOAST (1998)	{(291,168,176);(270,161,215)}	0.0379, 0.1376
S6.	Clark et al. (1999)	{(207,19,80);(181,25,101)}	0.0209, 0.0988
S7.	Clark et al. (1999)	{(187,15,104);(169,32,106)}	0.0938, 0.3242
S8.	Shelton et al. (2001)	{(83,14,5);(72,12,14)}	0.0766, 0.4147
S9.	Staszewski et al. (2001)	{(149,29,104);(144,15,121)}	0.1051, 0.3238

Note: Citations abbreviated for space; see Reference section below for full reference
* Exact one-sided Smirnov *p*-value first, then the approximate one-sided Smirnov *p*-value

Notice that in each case the approximate *p*-value is much larger than its exact counterpart. This refutes the common misunderstanding that exact *p*-values are always overly-conservative and therefore larger than the approximate *p*-values they would (and should) replace. Example S1 comes from a study of talc for malignant pleural effusions. There were 46 patients, and 23 were randomized to each group: talc and mustine. Some patients were considered to be "not assessable" because they died within a month of pleurodesis. Among the other patients (who were assessed), success or failure was defined in terms of radiologic criteria of effusion control. In addition to this binary success endpoint, patients were also classified as being alive or dead at the time the article was written, and as having had or not had evidence of recurrent effusion. So all in all we have four binary endpoints:

1. Died prior to assessment or not;
2. Dead or alive at the end of the study;
3. Success or not;
4. Recurrence or not.

This would appear to give $2 \times 2 \times 2 \times 2 = 16$ outcomes, but in fact the first two binary endpoints are fusible, because being alive at the end of the study necessarily entails also being alive long enough to be assessed. So instead of $2 \times 2 = 4$ outcomes for the first two binary endpoints above, we recognize the structural zero (one cannot die prior to being assessed and also be alive at the end of the study), and remove it to create a trichotomous information preserving composite endpoint, or IPCE, (died prior to assessment, assessed but dead at study end, alive at study end). See Berger (2002) for more information on the construction of the IPCE. We also note that the two binary endpoints success (yes/no) and recurrence (yes/no) are fusible, because recurrence is possible only if success was achieved in the first place, so we again have a structural zero (one cannot recur without having succeeded in the first place). Removing it gives the IPCE (no success, success then recurrence, success without recurrence). We have gone from $2 \times 2 \times 2 \times 2 = 16$ possible outcomes to only $3 \times 3 = 9$. But in fact further savings is possible too, as becomes evident from inspection of Table 4.

Dying before assessment precludes the possibility of a success, so the two lower left cells, labeled "SZ" in Table 4, are structural zeros. We make the simplifying assumption that death supersedes recurrence, and so we equate the two cells labeled "3" in Table 4. The upper right cell labeled "RZ" was a random zero; that is, there could have been patients surviving without success, but as it turned out, none did. This leaves only five active outcomes, labeled 1-5 in Table 4:

1. Died prior to being assessed;
2. Died after being assessed but without success;
3. Died after success;
4. Alive at study end but recurred;
5. Alive at study end without recurrence.

These outcomes are, of course, in order of increasing clinical benefit, and the data, as presented in Table 3, were (6, 8, 4, 2, 3) in the mustine group and (3, 2, 8, 0, 10) in the talc group, and the one-sided (to show a benefit of talc in shifting to more favorable outcomes) Smirnov p-values were 0.0138 (exact) and

0.02955 (approximate). If one were to use the two-sided 0.05 alpha level and then cut it in half for a 0.025 one-sided alpha level (which seems to be lacking in any real basis, yet is still used quite often as a policy), then only the exact Smirnov test would show a statistically significant improvement in outcomes associated with talc.

Table 4. The construction of the IPCE for example S1

	Died Before Assessment	Assessed, then Died	Alive at Study End
No Success	1	2	RZ
Success, then Recurrence	SZ	3	4
Success, no Recurrence	SZ	3	5

Examples S2 and S3 both represent the same patients, with the same endpoint, with the same treatments. All that varies is the timing of the measurement. Specifically, Example S2 is Day 2 and Example S3 is Days 1-5, and both come from a study of combination therapy for nausea (Fox, Einhorn, Cox, Powell, & Abdy, 1993). What is so amazing is the complete reversal in the direction of the shift. The endpoint we consider is response, which is scored as complete, major, or none. Note that this endpoint is the IPCE of two component binary response endpoints presented by Fox et al., specifically the response rate and the complete response rate. Clearly, the two endpoints are fusible, since a complete response implies also a response.

At Day 2, the data were (1, 5, 16) in the ondansetron group and (0, 0, 22) in the combination (ondansetron plus dexamethasone plus chlorpromazine) group. In other words, there was absolutely no effect of the combination therapy for the response rate (22/22 vs. 21/22), but a fairly strong effect on the complete response rate (22/22 vs. 16/22). At the Days 1-5 assessment, the situation was reversed, with (12, 3, 7) in the ondansetron group and (3, 7, 12) in the combination group. Now there was not much of an effect of the combination therapy for the complete response rate (12/22 vs. 7/22), but a fairly strong effect on the overall response rate (19/22 vs. 10/22). Either binary endpoint would show significance at the 5% alpha level at one time point but not at the other, with one-sided Fisher's exact test *p*-values of 0.5000 for the Day 2 overall response rate, 0.0106 for the Day 2 complete response rate, 0.0049 for the Days 1-5 overall response rate, and 0.1116 for the Days 1-5 complete response rate. The exact Smirnov test yields one-sided *p*-values of 0.0106 (Day 2) and 0.0108 (Days 1-5). The approximate test yields

one-sided *p*-values of 0.1947 and 0.02518. Once again, only the exact Smirnov test shows statistical significance at the customary 0.025 one-sided level of significance.

Example S4 is reinfarction data, in which the reinfarction could be confirmed or not, or there could be no reinfarction at all. Each patient can be scored on an ordered categorical scale with three categories, (confirmed reinfarction, reinfarction not confirmed, no reinfarction). Note that once again this is the IPCE for two binary endpoints originally presented. The data for the two treatment groups (placebo, then sotalol) are presented in Table 3, and the Smirnov test was used to compare the groups. As can be seen, the asymptotic version of the test was way off, to the point of being almost unbelievable, relative to the exact Smirnov test. The exact and approximate one-sided *p*-values were 0.0258 and 0.6823. Note that a one-sided *p*-value is not, in general, half the corresponding two-sided *p*-value, and also that a one-sided *p*-value can exceed 0.5 if the trend is in the "wrong" direction. Of course, that is not the case with the data at hand, as we tested for the direction of sotalol being superior, and the data do trend in this direction. So it is unclear why the asymptotic test would behave this way. One must ask if the data themselves might suggest the need for the exact version of the test. Berger (2000) reports that "It is unclear how one would determine the advisability of the approximate test, but if one were to 'think unconditionally' then the small middle margin would not be a concern. The large sample sizes (over 500 per group), coupled with expected cell counts that all exceed five, would certainly be reassuring" (p. 1322).

Example S5 concerns danaparoid for acute ischemic stroke. The TOAST Investigators (The Publications Committee for the Trial of ORG 10172 in Acute Stroke Treatment Investigators [TOAST], 1998) presented two binary endpoints, favorable outcomes (yes or no) and very favorable outcomes (yes or no), but again these two binary endpoints are clearly fusible, since a very favorable outcome implies also a favorable outcome. The IPCE is an ordered categorical outcome variable with categories for (unfavorable, favorable, very favorable). The TOAST Investigators inexplicably and indefensibly excluded some randomized patients from the analysis they called "intent-to-treat", but of course the correct intent-to-treat analysis would include all patients randomized. For now, we note that this set of patients can be classified by favorable outcomes at Day 7 as (291, 168, 176) in the placebo arm and (270, 161, 215) in the danaparoid group. The one-sided Smirnov *p*-values are 0.0379 (exact) and 0.1376 (approximate).

Examples S6 and S7 both come from the study of Clark et al. (1999) comparing rt-PA to placebo for ischemic stroke. The primary endpoint was a

complete recovery, defined as an NIHSS score of 0 or 1, at Day 90. A second binary endpoint was clinical improvement, defined as either a complete recovery (inexplicably now defined as an NIHSS score of 0, in contrast to the earlier definition which included also an NIHSS score of 1) or a change from baseline of at least 11 points. If we ignore the inconsistency in how "complete recovery" is defined (first as an NIHSS score of 0 or 1, then as just 0), then clearly a complete recovery implies a clinical improvement, so the two endpoints are fusible, and we really have a single trichotomous endpoint, (no improvement, clinical improvement, complete recovery), where "no improvement" is short hand for either no improvement or improvement not reaching the threshold for clinical improvement. With this endpoint, the data appear to be (we cannot be sure, since only proportions, and not actual patient counts, were presented in the original report) (207, 19, 80) at Day 30 for the placebo arm and (181, 25, 101) at Day 30 for the rt-PA arm. The one-sided Smirnov test yields *p*-values of 0.02094 (exact) and 0.09884 (approximate). At Day 90 the data were (187, 15, 104) in the placebo group and (169, 32, 106) in the rt-PA group, with corresponding one-sided *p*-values of 0.0938 (exact) and 0.3242 (approximate).

Example S8 comes from a study of St. John's wort for major depression. Shelton et al. (2001) measured depression with two binary endpoints, specifically remission and response. Remission is defined as HAM-D ≤ 7 and CGI-I 1 or 2, whereas response is defined as HAM-D ≤ 12 and CGI-I 1 or 2. Clearly these two endpoints are fusible, because a remission implies a response, so the IPCE would be (no response, response without remission, remission), and the data were (83, 14, 5) in the placebo group ($n = 102$) and (72, 12, 14) in the St. John's wort group ($n = 98$). The Smirnov *p*-values were 0.0766 (exact) and 0.4147 (approximate).

Example S9 comes from a study of combination therapy in adults with HIV. Staszewski et al. (2001) presented two binary outcomes, HIV RNA levels of 50 copies per mL or less and HIV RNA levels of 400 copies per mL or less. Obviously, the former implies the latter, so we again have a trichotomous IPCE of fusible endpoints, copies (> 400, 50-400, < 50). What was called the intent-to-treat population was certainly not that, as it excluded 35 of the 562 patients randomized. The true data set, as best as it can be reconstructed from the incomplete presentation published, is (149, 29, 104) in the abacavir arm and (144, 15, 121) in the indinavir arm, each in the presence of lamivudine and zidovudine (hence combination therapy). The Smirnov *p*-values were 0.1051 (exact) and 0.3238 (approximate).

Several recurrent themes emerge from the examples in this section. First, and most obvious, notice that the exact Smirnov test always provides a lower *p*-value than the approximate Smirnov test does, and notice also that in most cases, the approximate one is not even close. It seems reasonable, then, to suggest that the approximate Smirnov test never be used in practice, even if other approximations are accepted.

Examples of the *t*-Test Failing

The *t*-test is often used for continuous outcomes when the variance is not known. It is somewhat ironic that, while we are up front about not knowing the variance, we still wish to cling to this notion that we can somehow know that the data are normally distributed, despite Geary (1947) stating clearly that no data are normally distributed. Table 5 presents four examples in which the *t*-test gave results that differed markedly from corresponding exact results.

Table 5. Data sets for which the *t*-test fails badly

N	References	*p*-values*
T1.	Williams et al. (2000); Barber and Thompson (2000)	0.01, 0.79
T2.	Chaudhry et al. (2002); Jacobs (2003)	0.054, 0.004
T3.	Chaudhry et al. (2002); Jacobs (2003)	0.21, 0.016
T4.	Chaudhry et al. (2002); Jacobs (2003)	0.054, 0.006

Note: Citations abbreviated for space; see Reference section below for full reference

Example T1 bears some similarity to Example C6, in that one set of authors argued that an approximate test should be used after it was already established that an exact method was needed. In this case, the context was open access follow-up for inflammatory bowel disease, and its effect on costs. One particular endpoint was secondary care costs. Williams et al. (2000) correctly pointed out that:

> "Because data on use of resources tend to be highly skewed, routine parametric statistics are not appropriate. We therefore assessed significance by the Mann-Whitney U-test." (p. 545)

Using this proper analysis, the between-group *p*-value for secondary care costs is presented in Table 4 of the original article as 0.01, based on a mean cost of 582 (SD = 808) in the open access arm and 611 (SD = 475) in the routine care arm (the units are not provided in the table). Barber and Thompson (2000) argued that the means are most relevant, and:

> "[T]he most appropriate simple method for comparing mean costs is the ordinary t-test. By using the means and standard deviations in each group reported by the authors, we have calculated *p*-values from *t*-tests ... one of the authors' main conclusions – that open access follow-up used fewer resources in secondary care – is not supported: The *p*-value from the t-test is 0.79." (p. 1730)

Berger (2002) noted that there are two issues here, specifically the choice of test statistic (difference of means, difference of mean ranks, difference of Van der Waerden normal scores, or something entirely different) and the mode of generating a reference distribution. Differences in means can be accompanied by differences in shape and/or spread, so the *t*-test certainly is not always the most powerful test, even to detect the difference in means. But aside from this, even if we were to decide upon the difference of means as the test statistic, this certainly should not imply that we also use an approximation instead of an exact analysis. One can easily conduct an exact *t*-test, using the difference of means as the test statistic, and the permutation reference distribution to evaluate statistical significance.

Examples T2-T4 all come from the same study. Specifically, Chaudhry, Schroter, Smith, and Morris (2002) used the approximate *t*-test for five measures of readers' perceptions of papers with and without declarations of competing interests. These measures were interest, importance, relevance, validity, and believability, and the corresponding *p*-values for the five measures were 0.004, 0.016, 0.006, 0.001, and < 0.001. Jacobs (2003) re-analyzed the data with exact methods, after pointing out the flaws in using approximate methods for the data at hand. Three of the *p*-values became non-significant, specifically interest ($p = 0.054$), importance ($p = 0.21$), and relevance ($p = 0.054$). Of course, 0.054 is close to 0.05, so one might be tempted to declare it close enough. This is bad policy, and bad statistics, and not to be confused with selecting an alpha level other than 0.05. While it is perfectly reasonable to select an alpha level other than 0.05, maybe even 0.055, this selection needs to be made prior to viewing the data (and the *p*-value). Otherwise, one is left wondering just how broad this fuzzy

inclusion region actually is. Would 0.06 have been OK? What about 0.07? Where is the line drawn? In other words, what is alpha? And if alpha is not what we said it was up front, then we have a problem with the usage of alpha, and we are drawing the bull's eye around where the dart happened to hit.

Moreover, notice that the *p*-value for importance went from 0.016 to 0.21 when the analysis went from approximate to exact. This, as well as some of the other examples in Table 1, may well surprise those who consider the choice of an exact or an approximate test to be a "fourth decimal problem" that hardly warrants the attention of today's modern statistician. The StatXact manual states that "It is wise to never report an asymptotical p-value without first checking its accuracy against the corresponding exact or Monte Carlo *p*-value. One cannot easily predict a priori when the asymptotic *p*-value will be sufficiently accurate" (Cytel Software, 1995, p. 21). This is certainly excellent advice, but we can go a step further and ask why one would then discard the gold standard, the exact permutation *p*-value, once it is in hand, to use instead an approximation to it?

Summary and Conclusions

"Robustness procedures are generally considered to be statistical methods which are insensitive to small deviations from the underlying assumptions" (Prescott, 1998, p. 3864), and often this vagueness regarding how insensitive and how small the deviations must be allows for excessive discretion in filling in the blanks. That is to say that many researchers operate as if this robustness is absolute, so that there is no sensitivity at all no matter the magnitude of the deviation or how it is quantified. In point of fact, there seems to be no reliable method for imputing an exact *p*-value based on only the combination of knowledge of the approximate *p*-value and appeal to this alleged robustness. The fact that an exact *p*-value can fall anywhere on the unit interval even once we know the value of the approximate *p*-value should serve as ample demonstration that any notion of robustness being absolute is an illusion.

There might still be a value in computing approximate *p*-values anyway, if there were some added cost or difficulty involved in computing the exact *p*-value. In some applications this in fact is the case, but certainly not in all, and it is worth the effort to determine which case we are in. If an exact *p*-value can be computed relatively easily, with no prohibitive cost, then it is difficult to imagine any valid argument for not doing so. This remains the case even if one can put forth a compelling argument in favor of presenting an approximate *p*-value. For example, it may be the case that precedent favors the approximate *p*-value, which has

always been computed in the past. We want to see how the present data compare to past data sets, and those older ones were summarized, for example, with *t*-tests, and we do not have access to the complete data that would enable us to conduct exact analyses of those older data sets. In this case, it seems reasonable to compute the *t*-test on the new data set for the sake of comparing apples to apples and oranges to oranges, but this does not preclude the possibility of also computing an exact *p*-value in addition to the approximate one. Under no circumstances should we ever pretend to know the exact *p*-value without actually computing it.

Acknowledgements

The review team offered helpful comments that resulted in a far improved final version.

References

Barber, J. A., & Thompson, S. G. (2000). Would have been better to use t-test than Mann Whitney U-test. *BMJ, 320*, 1730. doi: 10.1136/bmj.320.7251.1730

Berger, V. W. (2000). Pros and cons of permutation tests in clinical trials. *Statistics in Medicine*, *19*(10), 1319-1328. doi: 10.1002/(sici)1097-0258(20000530)19:10<1319::aid-sim490>3.0.co;2-0

Berger, V. W. (2002). Improving the information content of categorical clinical trial endpoints. *Controlled Clinical Trials*, *23*(5), 502-514. doi: 10.1016/s0197-2456(02)00233-7

Berger, V. W. (2004). On the generation and ownership of alpha in medical studies. *Controlled Clinical Trials*, *25*(6), 613-619. doi: 10.1016/j.cct.2004.07.006

Berger, V. W., & Lachenbruch, P. A. (1998). *Robust permutation tests and randomized clinical trials*. Unpublished internal document, United States Food and Drug Administration.

Bradley, J. V. (1968). *Distribution-free statistical tests*. Upper Saddle River, NJ: Prentice Hall.

Chaudhry, S., Schroter, S., Smith, R., & Morris, J. (2002). Does declaration of competing interests affect readers' perceptions? A randomized trial. *BMJ*, *325*, 1391-1392. doi: 10.1136/bmj.325.7377.1391

Clancy, W. G. (2000). Letter to the editor. *American Journal of Sports Medicine, 28*(4), 615.

Clark, W. M., Wissman, S., Albers, G. W., Jhamandas, J. H., Madden, K. P., & Hamilton, S. (1999). Recombinant tissue-type plasminogen activator (Alteplase) for ischemic stroke 3 to 5 hours after symptom onset: The ATLANTIS study: A randomized controlled trial. *JAMA: Journal of the American Medical Association, 282*(21), 2019-2026. doi: 10.1001/jama.282.21.2019

Cytel Software. (1995). *StatXact 3 for Windows: Statistical software for exact nonparametric inference: User manual*. Cambridge, MA: Cytel Software Corporation.

Elwood, J. M. (1998). *Critical appraisal of epidemiological studies and clinical trials* (2nd ed.). Oxford, UK: Oxford University Press.

Fentiman, I. S., Rubens, R. D., & Hayward, J. L. (1983). Control of pleural effusions in patients with breast cancer: A randomized trial. *Cancer, 52*(4), 737-739. doi: 10.1002/1097-0142(19830815)52:4<737::AID-CNCR2820520428>3.0.CO;2-8

Fisher, R. A. (1935). *The design of experiments*. Edinburgh, UK: Oliver and Boyd.

Fox, S. M., Einhorn, L. H., Cox, E., Powell, N., & Abdy, A. (1993). Ondansetron versus ondansetron, dexamethasone, and chlorpromazine in the prevention of nausea and vomiting associated with multiple-day cisplatin chemotherapy. *Journal of Clinical Oncology, 11*(12), 2391-2395. doi: 10.1200/jco.1993.11.12.2391

Geary, R. C. (1947). Testing for normality. *Biometrika, 34*(3/4), 209-242. doi: 10.2307/2332434

Hewett, T. E., Lindenfeld, T. N., Riccobene, J. V., & Noyes, F. R. (1999). The effect of neuromuscular training on the incidence of knee injury in female athletes. *The American Journal of Sports Medicine, 27*(6), 699-706.

Hewett, T. E., Levy, M., Lindenfeld, T. N., & Noyes, F. R. (2000). Author's response. *The American Journal of Sports Medicine 28*(4), 615-616.

Hollander, M., & Wolfe, D. A. (1973). *Nonparametric statistical methods*. New York, NY: Wiley.

Jacobs, A. (2003). Clarification needed about possible bias and statistical testing. *BMJ USA, 3*, 93.

Lehmann, E. L. (1975). *Nonparametrics: Statistical methods based on ranks*. San Francisco, CA: Holden Day.

Little, R. J. A. (1989). Testing the equality of two independent binomial proportions. *The American Statistician, 43*(4), 283-288. doi: 10.2307/2685390

Lu, J., Ding, P., & Dasgupta, T. (2015). Construction of alternative hypotheses for randomization tests with ordinal outcomes. *Statistics & Probability Letters, 107*, 348-355. doi: 10.1016/j.spl.2015.09.013

Prescott, P. (1998). Robustness. In *The Encyclopedia of Biostatistics*. (Vol. 5, pp. 3864-3869). Chichester, UK: Wiley.

The Publications Committee for the Trial of ORG 10172 in Acute Stroke Treatment Investigators. (1998). Low molecular weight heparinoid, ORG 10172 (danaparoid), and outcome after acute ischemic stroke: A randomized controlled trial. *JAMA: Journal of the American Medical Association, 279*(16), 1265-1272. doi: 10.1001/jama.279.16.1265

Rigdon, J, & Hudgens, M. G. (2015). Randomization inference for treatment effects on a binary outcome. *Statistics in Medicine, 34*(6), 924-935. doi: 10.1002/sim.6384

Shelton, R. C., Keller, M. B., Gelenberg, A., Dunner, D. L., Hirschfeld, R., Thase, M. E.,… Halbreich, U. (2001). Effectiveness of St John's wort in major depression: A randomized controlled trial. *JAMA: Journal of the American Medical Association, 285*(15), 1978-1986. doi: 10.1001/jama.285.15.1978

Staszewski, S., Keiser, P., Montaner, J., Raffi, F., Gathe, J., Brotas, V.,… Spreen, W. (2001). Abacavir-Lamivudine-Zidovudine vs Indinavir-Lamivudine-Zidovudine in antiretroviral-naive HIV-infected adults: A randomized equivalence trial. *JAMA: Journal of the American Medical Association, 285*(9), 1155-1163. doi: 10.1001/jama.285.9.1155

Williams, J. G., Cheung, W. Y., Russell, I. T., Cohen, D. R., Longo, M., & Lervy, B. (2000). Open access follow-up for inflammatory bowel disease: Pragmatic randomized trial and cost effectiveness study. *BMJ, 320*, 544-548. doi : 10.1136/bmj.320.7234.544

Zelterman, D., Chan, I. S. F., & Mielke, P. W. (1995). Exact tests of significance in higher dimensional tables. *The American Statistician, 49*(4), 357-361. doi: 10.2307/2684573