Journal Of Modern Applied Statistical Methods
Winter 2002, Vol. 1, No. 1, 74-82

Parametric Analyses In Randomized Clinical Trials

Vance W. Berger
Biometry Research Group
National Cancer Institute

Clifford E. Lunneborg
Statistics and Psychology
University of Washington

Michael D. Ernst
Department of Mathematical Sciences
Indiana University – Purdue University Indianapolis

Jonathan G. Levine
Lincoln Technologies, Inc.

One salient feature of randomized clinical trials is that patients are randomly allocated to treatment groups, but not randomly sampled from any target population. Without random sampling parametric analyses are inexact, yet they are still often used in clinical trials. Given the availability of an exact test, it would still be conceivable to argue convincingly that for technical reasons (upon which we elaborate) a parametric test might be preferable in some situations. Having acknowledged this possibility, we point out that such an argument cannot be convincing without supporting facts concerning the specifics of the problem at hand. Moreover, we have never seen these arguments made in practice. We conclude that the frequent preference for parametric analyses over exact analyses is without merit. In this article we briefly present the scientific basis for preferring exact tests, and refer the interested reader to the vast literature backing up these claims. We also refute the assertions offered in some recent publications promoting parametric analyses as being superior in some general sense to exact analyses. In asking the reader to keep an open mind to our arguments, we are suggesting the possibility that numerous researchers have published incorrect advice, which has then been taught extensively in schools. We ask the reader to consider the relative merits of the arguments, but not the frequency with which each argument is made.

Keywords: Exactness, Nonparametric, Permutation test, Reality-based analyses, Robustness, Validity

Introduction

Medical errors may be classified by the broken link in the chain connecting (a) study objectives to (b) medical data bases to (c) p-values to (d) study conclusions to (e) recommendations to (f) accepted medical practice to (g) actual medical practice. Medical errors attributable to physicians deviating from accepted practice, corresponding to the last link in the chain, (f) to (g), may attract the most malpractice suits and media attention. Yet the frequent insidious errors that occur at the second link, from (b) to (c), involving inappropriate statistical methodology, may result in even more damage (Bailar, 1976). In some cases, assumptions are required to calculate p-values, but when a platinum standard analysis is available so that "significance [may be] assessed in a way not involving unverifiable assumptions" (Tukey, 1993), it would be a (b) to (c) error not to use it.

In randomized clinical trials (RCTs), the random allocation of patients to treatment groups serves as the basis for valid between-group inference. In RCTs, then, neither random sampling from a target population nor unverifiable assumptions are required (Feinstein, 1993) to construct between-group tests that allow Type I errors (false positive findings) to occur at no greater than a specified rate (α). These platinum standard tests are design-based permutation tests that use as the reference distribution the set of actual potential allocation sequences (Berger, 2000a, Section 3.1). We will refer to design-based platinum standard permutation tests as exact in the remainder of the article, yet two caveats are needed to qualify the use of the word "exact" in this context.

First, design-based tests are exact for the strong null hypothesis, which specifies that each patient would respond identically to each treatment under study. This strong null hypothesis is not the complement of the superiority alternative hypothesis. There is an indifference region in which the weak null hypothesis (specifying common population response rates or means) is true but the strong null hypothesis is not. Design-based tests need not be exact on this region.

Second, exactness is not preserved, even for the strong null hypothesis, when the analysis is based on a

Vance W. Berger is a Mathematical Statistician at the National Cancer Institute and an Adjunct Professor at the University of Maryland Baltimore County. Clifford E. Lunneborg is Emeritus Professor of Statistics and Psychology at the University of Washington, Seattle. His work focuses on the deployment of computer-intensive and design-based statistical methods. Michael D. Ernst is Assistant Professor in the Department of Mathematical Science at Indiana University – Purdue University Indianapolis. Jonathan G. Levine is at Lincoln Technologies, Inc.

randomization scheme other than the one that was actually used. Software may not always be available for constructing a test that mimics the actual randomization used. The size of the study may preclude the possibility of enumerating all possible permutations of treatment allocations consistent with the actual randomization used, and Monte Carlo approximations may be needed. So not every permutation test that is called exact is design-based, and it is not clear that permutation tests which are not design-based are more robust than parametric tests. Even design-based permutation tests, which are necessarily more robust (in the sense of keeping the power under α) than parametric tests when the strong null hypothesis is true, may not be more robust than parametric tests when the weak null hypothesis is true and the strong null hypothesis is not. Although technically this opens the door to the possibility that in some cases the parametric test may be preferable to the best available permutation test, none of us can recall this argument being used in practice to justify a parametric analysis. Without a detailed investigation of the robustness of each test in the specific situation, we would consider the best way to decide between a parametric test and a permutation test to be the conditions for its exactness.

A parametric test requires both random sampling and proper specification of the distribution from which one is sampling randomly to be exact. In some sense, a different random allocation scheme, which is all that is needed for non-design-based permutation tests to be exact, comes closer to the actual random allocation scheme than random sampling from a known distribution does. In addition, inexactness caused by the use of Monte Carlo sampling can be bounded by selection of the number of points in the sample space.

For these reasons, we consider only cases in which the permutation test (even if not design-based) can safely be presumed to be more robust than the parametric test, and we note that this covers every case we have encountered in practice. The disturbing overuse of parametric analyses in these cases cannot be explained by the lag time required for new methods to gain acceptance in practice (Altman & Goodman, 1994) – in fact permutation tests are not new (Ludbrook & Dudley, 1998, Section 4.1). More likely, this trend is due to a combination of the reluctance of journal editors to accept correctness in place of precedent (Ludbrook & Dudley, 1998) and some recent publications that endorse parametric analyses.

For example, Agresti and Coull (1998) cited the conservatism of exact methods as a reason to use approximate methods. Because their article was not especially focused on hypothesis testing or on RCTs, it does not strike us as entirely inconsistent with our views, although we do find it inappropriate to cite their article to justify the use of parametric tests in RCTs.

Other articles have specifically proposed that parametric tests be used in RCTs. For example, Barber and Thompson (2000) criticized the use of the exact Wilcoxon-Mann-Whitney (WMW) test instead of the parametric t-test in a RCT. Likewise, Shuster (1990) and Hewett et al. (2000) were both critical of Fisher's exact test for binary data.

We return to these articles after illustrating the discrepancy between the nominal α level and the actual α level. In arguing the obvious point that the actual level ought not exceed the nominal level (whether or not it is 0.05) we refute the application of Agresti and Coull's (1998) assertions to hypothesis testing in RCTs. We specialize this argument to the case of continuous data, using as an example a recent RCT to compare open access to routine appointments for inflammatory bowel disease (Williams et al., 2000a). We pay particular attention to the points made by Barber and Thompson (2000), and refute the key one about the relative merits of the t-test and the WMW test. We specialize to the comparison between the Chi-squared test and Fisher's exact test for binary data, using as an example a recent study of the effect of neuromuscular training on knee injuries in female athletes (Hewett et al., 1999). Here we refute the points made by Shuster (1990) and Hewett et al. (2000). Then, we discuss and refute some of the reasons often cited for using a parametric test instead of an exact one. Finally, we provide recommendations.

Strict Preservation of the Type I Error Rate (α)

As the probability of a false positive, the Type I error rate (α) has been called the regulator's risk in the drug evaluation context. This may suggest that only regulators need to concern themselves with the frequency, under null conditions (i.e., an ineffective medical intervention), with which analyses claim statistical significance (i.e., superior efficacy). This is a dangerous view, because even a medical error attributable to a break in the second link of the chain, (b) to (c), is still a medical error that can cause tremendous damage (Bailar, 1976). The "only assurance of the low likelihood of [the approval of ineffective compounds that have serious adverse effects] is the Type I error" which must "occur at tolerably low rates [for] the community [to] best be assured that the conclusions of the trial most likely reflect the anticipated experience of patients" (Moye, 1999, bracketed material added for clarity).

In this section, we take a careful look at the actual α level, and distinguish it from the tolerably low rate, which is (or should be) the nominal α level. In this discussion we must bear in mind the added importance of the Type I error rate due to the frequent unquestioning acceptance of positive between-group results (Berger, 2000a, Section 1; Voutilainen, 2001). We note that allowing the nominal α level to vary with the nature of the disease and the safety profile of the agent under study may be quite reasonable. In no way do we insist that 0.05 should always be used for α. We do, however, insist that there be adherence to whatever α level is selected.

Berger (2000a, Section 3.1) demonstrated the potential for a parametric test to violate this basic tenet, and yet cover it up by calculating the α level incorrectly. This occurs because a parametric test is exact as an answer to one question, yet it is used for a different question. As an illustration, consider Table 14.2 in Section 14.2.6 of Berger and Ivanova (2001), based on the 2x2 contingency table {(12,10);(3,19)} originally presented by Fox et al. (1993). Columns are response outcomes (no or yes) and rows are treatments, ondansetron (OND) vs. combination therapy (ODC). The chi-square p-values are exact as answers to the question "If one were to sample randomly from a chi-square distribution (with one degree of freedom), then what is the probability of finding results as extreme or more extreme than those we observed?".

Had the experiment actually employed random sampling from a distribution that actually had a chi-square distribution, then this question would be equivalent to "If one were to repeat the experiment performed, under null conditions (the equivalence of OND and ODC), then what is the probability of finding results as extreme or more extreme than those we observed?". In fact there was no random sampling, the population does not have a chi-square distribution, and the exact answer to the former question is not an exact answer to the latter question. To obtain an exact answer to the latter question, which is the one of interest, we hypothetically repeat the experiment. This means re-randomizing the allocation repeatedly, using the same randomness (probability structure) that was used to determine the actual allocation. Under the strong null hypothesis of no treatment effect the responses are independent of the allocation, so the responses do not change. This allows us to compute the test statistic for each of these hypothetical repeats of the experiment. This is how platinum standard permutation tests ensure exactness.

The one-sided actual α level is the probability, under the null hypothesis (OND and ODC being equally effective), of declaring that ODC is more effective than OND. This declaration will be made if the p-value is as low as or lower than the nominal α level, so the nominal α level determines the number of ODC responses required by each test to claim significant superiority of ODC. If we pick a nominal α level of 0.0250, then the rejection region consists of those outcomes for which the p-value is no greater than 0.0250. It turns out that both tests would have the same rejection region, consisting of the outcomes for which ODC has 19 or more responses. But this event occurs with probability 0.0049, and not 0.0250. For this nominal α level both tests are conservative (and equally conservative). The outcome of ODC having 18 responses has a chi-square p-value of 0.0282 and a Fisher p-value of 0.0273, so it does not qualify for inclusion in either 0.0250 rejection region (its p-value is too large). But if we change the nominal α level from 0.0250 to 0.0275, then Fisher's exact test could fit the additional outcome into its rejection region, while the chi-square test could not. Because the chi-square rejection region still has null probability 0.0049, it still has the same actual α level, 0.0049. But Fisher's exact test now has a larger rejection region, with null probability 0.0273, which serves as its actual α level. In this case, Fisher's exact test is much less conservative than the chi-square test. This information appears in the table below.

With a nominal α level of 0.0500, each test could include the 18 outcome but not the 17 outcome in its rejection region, because p(17)=0.1014>0.05 and p(17)=0.1017>0.05, respectively, for Fisher's exact test and the chi-square test. Again, both tests are equally conservative. With a nominal α level of 0.2625 the tests again diverge, with Fisher's exact test rejecting for 17 but not 16 (p=0.2628), while the chi-square test can reject for both (p=0.2624 for 16), but not for 15 (p=0.5000). With an actual α level of 0.1014, Fisher's exact test is quite conservative; but the chi-square test, with its actual α level of 0.2628, is anti-conservative. We have seen three distinct cases, but we did not see a case in which the chi-square test was simultaneously valid and less conservative than Fisher's exact test. Because the Fisher actual α -level will

Table 1. Fisher's Exact Test Vs Chi-Square Test

Nominal α	Fisher's exact test		Chi-square test		Compared to Fisher's exact
(one-sided)	cut-off	actual α	cut-off	actual α	test the chi-square test is
0.0250	19	0.0049	19	0.0049	equally conservative
0.0275	18	0.0273	19	0.0049	more conservative
0.0500	18	0.0273	18	0.0273	equally conservative
0.2625	17	0.1014	16	0.2628	anti-conservative

be that attainable p-value closest to but not exceeding the nominal Type I error rate, no test can be simultaneously valid and less conservative than it. In fact, any exact test is minimally conservative in this sense. As such, even conservatism, which is often used as an argument against exact tests (Agresti & Coull, 1998), favors the exact test unless the parametric test gains an unfair advantage by being anti-conservative.

For a more extreme example of a comparison between an anti-conservative parametric test and a conservative exact test, consider the 2x2 table {(8,2);(4,6)}. That is, there are 2/10 successes in the control group, and 6/10 successes in the active group. With a nominal α level of 0.05 one-sided, the actual α levels are 0.0099 for Fisher's exact test and 0.0849 for the chi-square test. Because 0.0849 is closer to 0.05 than 0.0099 is, some would argue that the chi-square test at the 0.0849 level is most appropriate. In fact, it may or may not be more appropriate than Fisher's exact test at the 0.0099 level, but these are not the only options. If the response variable is observed fairly soon after randomization, then one could consider an adaptive procedure in which recruitment to the study stops only when the conservatism is small enough. This might be judged to be the case if either the observed p-value interval (Berger, 2001) is entirely on one side of α or the p-value interval that contains α is itself contained in a fairly tight pre-defined interval around α. So a larger sample size might resolve this problem satisfactorily. But even without resorting to larger sample sizes, it is also clear that if the chi-square test can be run at an actual 0.0849 level, then 0.0849 is an attainable p-value, meaning that there is an outcome for which 8.49% of the outcomes are as or more extreme. This means that Fisher's exact test can also be conducted at the 0.0849 level.

So now there are two issues. First, is it acceptable to use a test with an actual α-level larger than the planned 0.05? Second, if the answer to the first question is yes, then which test should be used at the 0.0849 level? Berger (2000b) noted the inappropriate willingness of some researchers to accept the general conservatism of exact tests, without considering the extent of conservatism of the exact test in question, as sufficient reason to answer yes to the first question. Yet the extent of conservatism of any exact test may be quantified by the p-value interval (Berger, 2001). Furthermore, conservatism is not a problem, because the lower the Type I error rate the better. The attendant loss of power may be a problem, so the power needs to be considered.

In any event, if the answer to the first question is no, then clearly Fisher's exact test must be used at the 0.0099 level. Regarding the second question, we note that with a nominal α-level of 0.0849 both the Fisher and chi-square p-values would be significant exactly 8.49% of the time. There is still an important distinction, however, in that no more than 5% of the time would the Fisher p-value be significant at the 0.05 level. This is not the case, however, for the chi-square test, for which the p-value would be significant at the 0.05 level 8.49% of the time. If events that should occur with probability one in a thousand "do not occur with this frequency, there is something seriously wrong with our understanding of probability" (Bailar, 1976). Likewise, if the 5th percentile of a distribution is actually exceeded by 8.49% of the outcomes, then it is not really the 5th percentile of the distribution. If the nominal α level is planned to be 0.05, then it might be reasonable in some cases to use an actual α level that exceeds it, perhaps 0.0849. To do so, and then after the fact report 0.05 as the α level used, is tantamount to planning a study with 200 patients, actually recruiting only 180 patients, yet still reporting the actual sample size as 200. Because parametric tests are guilty of this type of deception, conservatism cannot justify their use in practice.

The Parametric t-test vs the Wilcoxon-Mann-Whitney (WMW) Test

In this section we consider the merits of the WMW test relative to the parametric t-test for unadjusted between-group comparisons on the basis of continuous data. We first point out that there are numerous versions of the WMW test (Bergmann, Ludbrook, & Spooren, 2000), and this is likely what prompted Ludbrook (1996) to support exact tests in general yet specifically criticize the WMW test. In any event, it is the exact version of the WMW test that we consider. Williams et al. (2000a) used the WMW test (the exact version, we assume) to assess resource use and costs in a RCT comparing open access to routine appointments for inflammatory bowel disease.

Barber and Thompson (2000) commented that:

1) "resource use and cost data tend to have highly skewed distributions",
2) "t-test methods are only strictly valid for data that are normally distributed",
3) "the most appropriate simple method for comparing mean costs is the ordinary t-test",
4) "use of inappropriate methods for the analysis of cost data is all too common".

It is certainly true that many distributions are highly skewed, and even bell-shaped distributions need not be normally distributed. Furthermore, even if a variable has a normal distribution in the target population of interest, allowing for non-random sampling from this normal distribution allows for the possibility of accepting or rejecting an observation on the basis of the observation itself. Using the rejection method (Hoaglin, 1983) would then allow for the retained observations to have any distribution we want them to have. Hence, a lack of random sampling necessarily precludes the possibility of asserting normality of the sampling distribution of the data. If random allocation was used, then it is the only part of the study that was "experimental" (manipulated), and the sampling distribution is a permutation distribution. This permutation distribution may converge to normality as the sample size grows infinitely large, but we feel safe in agreeing with Geary (1947) and Hunter and May (1993) that no data (based on a finite sample size) have a normal distribution. So, we agree with Barber and Thompson's (2000) first point.

The second point has an ambiguity owing to the improper placement of the word "only" in the sentence. In light of the third point, it is conceivable that "only" was meant to start the sentence and limit the class of valid analyses to the t-test. However, given the context in which this sentence appears, it seems more likely that "only" was meant to follow "valid" so as to limit the situations in which the t-test is valid to those in which the data are normally distributed. If this latter interpretation is the correct one, then we agree with the second point. In fact, the perceived robustness to non-normality of parametric methods (as will be discussed) is somewhat of an illusion (Hunter & May, 1993).

We develop our comments on the third point by first noting that Thompson and Barber (2000) claimed that "only the t-test on untransformed data can be appropriate for costs, since it is the only one that addresses a comparison of arithmetic means". Even conceding the point that it is reasonable to compare mean costs, we can still disagree with the third point, which may be interpreted broadly to include the exact t-test, although the use of the word "ordinary" makes it is more plausible that only the parametric t-test was intended. As articulated above, parametric tests (including the t-test) fail to preserve the Type I error rate in RCTs. As such, we cannot agree with the third point if it is interpreted the plausible way. In considering the more favorable interpretation of the third point, we note that our desire to maximize power to detect mean differences might suggest that the exact t-test would be ideal. However, mean differences may well be accompanied by differences in spread and/or shape (Hart, 2001), and the nature of these differences will affect which test is most powerful. In fact, one could construct an exact test using any test statistic, including the between-group mean difference in raw costs (the exact t-test), in ranks (the WMW test), or in Van der Waerden normal scores. Often the WMW test is more powerful and/or more robust than the t-test (Lachenbruch, 1992; Higgins & Blair, 2000; Weinberg and Lagakos, 2001), so we cannot agree with the third point of Barber and Thompson (2000) even if it is interpreted to include the exact t-test.

Regarding the fourth point, we note that only an exact test can protect against a Type I error attributable to assuming normality. Using an exact test based on a test statistic that is broadly powerful to detect mean differences, and getting a low p-value, Williams et al. (2000a) convincingly demonstrated that "open access greatly reduces secondary care costs" (Williams et al., 2000b). Barber and Thompson (2000) demonstrated that for this data set, either the normality assumption was sufficiently flawed or the difference in means was sufficiently accompanied by shifts in shape and/or scale that the t-test failed to detect this difference. Apparently, Barber & Thompson (2000) failed to recognize that their primary contribution is the demonstration of the truth of the fourth point, which they accomplished by illustrating that the frequently used parametric t-test can be quite misleading (Williams et al., 2000b), and is therefore inappropriate.

The Chi-squared Test vs Fisher's Exact Test

In a recent study of the effect of neuromuscular training (Hewett et al., 1999), the chi-square test was used to analyze knee injuries in female athletes. Clancy (2000) commented that "Because the observed and expected number of knee injuries was less than five in at least one cell, an approximate method is inappropriate. An appropriate method in this instance would have been a Fisher's exact test. Incidentally, use of this exact method demonstrated no statistical significance ..., suggesting that the extreme variability present in the small sample resulted in an incorrect finding when an approximate method was used. This provides all sports medicine researchers with a potent example of why appropriate statistical analysis is extremely important." We comment below on choosing tests based on expected cell counts. For now, note that Fisher's exact test is a misnomer, because as discussed above it is not exact unless there is random allocation that has as its only restriction that the treatment totals are fixed at their observed values (Berger, 2000a, Section 3.1). As the Hewett et al. (1999) study appears to have been nonrandomized, neither Fisher's exact test nor the chi-square test is exact in this context. Yet, in response Hewett et al. (2000) appeared to accept that Fisher's exact test was in fact exact, responding only that:

> "the chi-square test is unconditional in that a significance probability produced by it refers to the long-term likelihood in repeated experiments of observing an outcome more extreme than ours, regardless of the marginal cell counts in these future experiments under the null hypothesis (more applicable and inclusive to future studies). Fisher's exact test, on the contrary, is conditional and is, technically, only applicable to future experiments like ours in which the

> marginal cell counts are fixed at the exact values that we obtained in our particular study. The significance probability of the Fisher's exact test is much more limited in scope than a chi-square probability, which is one of the reasons Fisher's exact test is rarely used by statisticians."

Hence, it is reasonable to assume that the same discussion would have ensued had the study actually been a RCT. A similar set of views was expressed over ten years ago by Shuster (1990, p. 26), who stated that Fisher's exact test "is not a true p-value, since the additional proviso is made that in the replication of the experiment, you must match the total number of successes with that observed". The RCT design is well summarized by Kempthorne (1979) as Origin III sampling, for which valid probability statements about what might have happened with different samples are not supported (Berger, 2000a, Section 2.2). Applying a parametric test cannot extend the scope to which valid inferences apply, but producing the appearance of such extension can be dangerously seductive. Conditioning on the observed marginal totals, as Fisher's exact test does, is required for exactness and validity, and hence is not a weakness (Berger, 2000a, Section 4.3). In fact, by providing internal validity (exactness) through recognizing the limitations of the study design, Fisher's exact test can actually enhance, and not compromise, the possibility for external validity Berger (2000a, Section 5). As such, we find that while the chi-square test may have asymptotically good properties in the random sampling context, its use in RCTs reflects familiarity, and not appropriateness.

So Why Do Researchers Use Parametric Analyses?

In the Introduction we allowed for the possibility that one could argue convincingly that a given parametric test might be preferred to the best available permutation test provided that it were more robust in preserving the nominal Type I error rate on the indifference region and the strong null region. However, simply stating that parametric tests are robust, without comparing the robustness of a particular parametric test to that of a competing exact test, cannot be convincing. When such a comparison is not offered, and we have never seen one in practice, robustness cannot be offered as a reason to use a parametric test.

We demonstrated above that conservatism of exact tests is not a valid reason to select a parametric test either. Another reason that is often cited, especially if a preliminary test of the assumptions underlying the validity of the parametric test is conducted, is the frequent agreement of the exact and parametric tests. The lack of obvious problems resulting from all these years of using parametric tests has also been cited. One reason for using parametric tests that is *not* often cited, but may be deduced from the lack of attention dedicated to this issue, is that some may feel that this is a fourth decimal point issue that is not ready for prime time. We find no merit in any of these reasons. In the remainder of this section we provide journal editors and regulatory authorities with responses they can use if and when they encounter such arguments.

If credibility for the parametric test derives from assurances that its p-value will likely be close to the corresponding exact one, then this is tantamount to an admission that the exact test is the gold standard (or, perhaps, the platinum standard). Approximate tests cannot be any more exact than the exact tests they are trying to approximate and, as approximations to the exact tests, are correct only to the extent that they agree with the exact test. A "heads I win, tails you lose" situation then arises, because if the parametric and exact tests lead to essentially the same inference, then this is as much an argument in favor of the exact test as it is for the parametric test, and there is no benefit to using the parametric test. If they do not agree, then the exact test needs to be used.

Feinstein (1993) wrote that "a statistician defending the general use of t and chi-square tests in modern research could point to their frequent accuracy. With the same argument, an old school clinician might point out that diabetes mellitus can usually be diagnosed by tasting the urine or applying Benedict's reagent. With the availability of better and equally easy ways to diagnose diabetes, however, these old procedures were gradually replaced by techniques that are more reliable. Similarly, when the aid of computers allows permutation or randomization tests to be performed easily, the t-test and chi-square test will probably begin to disappear as routine procedures. Even without computers, however, the permutation tests are the preferable and perhaps mandatory procedures to be used". Even in those cases in which the two tests agree perfectly, Altman (1982, p. 67) makes a compelling case that setting a precedent for poor methodology encourages other researchers to use the same poor methodology in the future. It is likely that in some of these future studies the results will be materially affected.

For these reasons, we believe that even if parametric tests tend to agree with exact tests, meaning that with "high" (left undefined) probability, the parametric p-value will be "close" (also left undefined) to the exact p-value, they still should not be used. Even with preliminary tests of the assumptions, the general similarity of the two tests does not exclude the possibility of discordant results between the tests for given data sets. This is because the preliminary test has as its null hypothesis the conditions that would allow for the use of the parametric test. The null hypothesis cannot be proven by a formal test of hypothesis, especially when the test suffers from poor power, as the preliminary tests to detect conditions that would render the parametric test unreliable (such as non-normality) often do.

For example, Little (1989) presented a 2x2 table, with cell counts {(170,2);(162,9)}. Because each expected cell count is at least 5, the chi-square test would be used, yet for one-sided testing (which is generally conducted at the 0.025 level) the chi-square test would find significance (p=0.0162) and Fisher's exact test would not (p=0.0299). Barber and Thompson (2000), Berger (2000a, Section 2.3), and Clancy (2000) presented other real data examples in which use of the parametric test matters more than in the fourth decimal. Given the danger in restricting the use of exact analyses to cases in which the need for such use is obvious, the prudent approach is to be suspicious of assumptions even when there is no apparent reason to be suspicious. The only way to validate a particular parametric p-value, and ensure that it differs from the corresponding exact p-value in only the fourth decimal place, is to compare it to the exact p-value. If the exact p-value needs to be computed to validate the approximate one, then why not simply use the exact one, the ready availability of which renders the extent to which an approximate test approximates it irrelevant (Berger, 2000a)?

As for the precedent for using parametric analyses without obvious damage, we note that p-values provided by inappropriate methodology are numbers between zero and one, and look just like p-values produced by appropriate methodology. Alarms do not go off when an inappropriate method is used. In fact, two forces conspire to conceal the damage caused by the use of inappropriate methodology by separating the manifestation of this damage from the antecedent usage of the inappropriate methodology. Specifically, when inappropriate methodology causes damage, there is both a diffusion of the damage to a set of patients who do not act or think as a unified individual and a lag time in the manifestation of this damage. Add to this that the patients may be sick anyway, and there is little hope of ever tracing the damage back to the cause. That damage actually does occur as the result of medical errors, and often goes unnoticed, has been well documented (Moore, 1995). How much easier would the life of an epidemiologist be if every risk were easily identified and linked to the damage it caused? In fact there have always been real risks that were not mitigated by our ignorance of their existence. The lack of an identifiable victim complaining about the use of parametric tests cannot be interpreted as the lack of a victim.

Conclusion

It has been said that for evil to prevail all it takes is a few good people to stand by and do nothing. The same could be said for the "scandal of poor medical research" (Altman, 1994). To avoid being part of the problem, all involved parties should insist on quality methodology. This applies especially to regulators and medical journal editors, who, given their "public duty to ensure that reports of research provide valid information" (WAME, 2001), might be seen as functioning as *de facto* regulators. Given that papers with poor methodology can cause harm (in numerous ways) and cannot be "unpublished" (Altman, 1982), consumers of medical publications (including practicing physicians and HMOs) should hold these publications to rigid standards before accepting and acting on the results (by altering reimbursement or prescribing patterns). That is, "because low p-values are not themselves persuasive but require solid methodology as a foundation, we must resist the pressure to view data positively that were produced from poor methodology" (Moye, 1999). Patients might want to ask their physicians about the evidence on which a decision is based. Given the importance of analyzing RCT data with methods that are applicable to and appropriate for RCTs, medical schools might consider offering degrees specifically in RCT design and analysis. Granting institutions might want to ensure that medical research is supported by a reality-based trialist who will build robustness into the analyses by making a minimum of unverifiable assumptions.

Because those who claim to be methodological experts often disagree among themselves, there is both conflicting information and misinformation being taught in schools and published in both the medical and the statistical literature. Two steps might put medical researchers in a better position to evaluate the analyses a statistician proposes. First, medical researchers could think hard about how best to analyze the data, possibly reading Feinstein (1993) carefully if the study is a RCT. Second, the medical researcher could require the statistician to justify the proposed analyses with logic and reason, instead of (or in addition to) references. It would help if the statistician would provide an informed consent document to spell out the assumptions and limitations of the proposed analyses. See the Appendix for an example dealing with parametric analyses. Although it is unlikely that statisticians who use parametric analyses would make themselves look bad by providing such a document, the medical researcher could bring some version of this document to the attention of the statistician to initiate the discussion about the analyses planned.

Developers of statistical software and authors of text books should offer analyses with a minimum of required assumptions and should make explicit the assumptions and limitations of all analyses. This presently is not the case (Bergmann, Ludbrook, & Spooren, 2000). Reality-based trialists should, when confronting a researcher endorsing a parametric analysis, consider the advise of Bross (1990), who wrote that "if we politely call a method 'dubious', the criticism can be brushed off as a 'difference of opinion between experts'. However, if most statisticians call a method 'fraudulent', the criticism cannot be brushed off so easily.". Hopefully this article will prove useful to reality-based trialists in their efforts to argue effectively against parametric analyses, or at least using parametric analyses without carefully checking their situation-specific robustness.

As for the researcher who wants to resist reality-based analyses, and maintain the *status quo* of routinely using parametric tests, we agree with Bross (1990) that "the user of a statistical method has the responsibility for dealing with the *scientific* question: Are the assumptions valid? In particular, when human health and safety might be jeopardized ..., a statistician has a direct responsibility to protect the public health and safety by following fail-safe principles in dealing with any assumptions". Given the logical basis for reality-based analyses, it is likely only a matter of time before the medical profession catches on that the normal theory that perplexed them in medical school actually has little or no place in RCTs. When this happens, and proper analyses become the rule instead of the exception, the emperor's new clothes will be seen for what they are, and some may well wonder why the naked emperor was allowed to rule for so long. We would not want to be in a position of having to explain why right up until the time that we were forced to cease and desist we continued to use inappropriate methods that resulted in medical errors leading to unnecessary morbidity and mortality. Although much work remains to fully elucidate the optimal methods for comparing medical interventions, and while there may never be a bias-proof system, there can be no excuse for not getting the easy ones right.

References

Agresti, A., & Coull, B. A. (1998). Approximate is better than 'exact' for interval estimation of binomial proportions. *The American Statistician*, *52*, 119-126.

Altman, D. G. (1982). Statistics in medical journals. *Statistics in Medicine*, *1*, 59-71.

Altman, D. G. (1994). The scandal of poor medical research. *British Medical Journal*, *308*, 283-284.

Altman, D. G., & Goodman, S. N. (1994). Transfer of technology from statistical journals to the biomedical literature: Past trends and future predictions. *JAMA*, *272*, 129-132.

Bailar, J. C. (1976). Bailar's laws of data analysis. Clin Pharmacol Ther, *20*, 113-120.

Barber, J. A., & Thompson, S. G. (2000). Would have been better to use t-test than Mann-Whitney U-test. *British Medical Journal*, *320*, 7251, 1730.

Berger, V. W. (2000a). Pros and cons of permutation tests. *Statistics in Medicine*, *19*, 1319-1328.

Berger, V. W. (2000b). Comment on Ludbrook and Dudley. *The American Statistician*, *54*, 85-86.

Berger, V. W. (2001). The p-value interval as an inferential tool. *Journal of the Royal Statistical Society D (The Statistician)*, *50*, 1, 79-85.

Berger, V. W., & Ivanova, A. (2001). Permutation tests for phase III clinical trials. In (S. P. Millard and A. Krause (Eds.): *Applied statistics in the pharmaceutical industry with case studies using S-PLUS*. Springer-Verlag, New York, 349-374.

Bergmann, R., Ludbrook, J., & Spooren, W. P. J. M. (2000). Different outcomes of the Wilcoxon-Mann-Whitney test from different statistics packages. *The American Statistician*, *54*, 1, 72-77.

Bross, I. D. (1990). How to eradicate fraudulent statistical methods: Statisticians must do science. *Biometrics*, *46*, 1213-1225.

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

Feinstein, A. R. (1993). Permutation tests and 'statistical significance'. *MD Computing*, *10*, 1, 28-41.

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*, 2391-2395.

Geary R. C. (1947). Testing for normality. *Biometrika*, *34*, 209-242.

Hart, A. (2001). Mann-Whitney test is not just a test of medians: Differences in spread can be important. *British Medical Journal*, *323*, 391-393.

Hewett, T. E., Levy, M., Lindenfeld, T. N., & Noyes, F. R. (2000). Letter to the editor. *The American Journal of Sports Medicine*, *28*, 4, 615-616.

Hewett, T. E., Lindenfeld, T. N., Riccobene, J. V., and 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.

Higgins, J. J., & Blair, RC. (2000). Comment on Ludbrook and Dudley. *The American Statistician*, *54*, 86.

Hoaglin, D. C. (1983). Generation of random variables. In N. L. Kotz and C. B. Read (Eds.): *The Encyclopedia of statistical sciences, Volume 3*. NY: John Wiley and Sons, 376-382.

Hunter, M. A., & May, R. B. (1993). Some myths concerning parametric and nonparametric tests. *Canadian Psychology*, *34*, 384-389.

Kempthorne, O. (1979). In dispraise of the exact test: Reactions. *Journal of Statistical Planning and Inference*, *3*, 199-213.

Lachenbruch, P. A. (1992). The performance of tests when observations have different variances. *The Journal of Computations and Simulations, 40*, 83-92.

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

Ludbrook, J. (1996). The Wilcoxon-Mann-Whitney test condemned. *British Journal of Surgery, 83*, 136-137.

Ludbrook, J. L., & Dudley, H. (1998). Why permutation tests are superior to t and F tests in biomedical research. *The American Statistician, 52*, 2, 127-132.

Moore, T. (1995). *Deadly medicine.* NY: Harper-Collins.

Moye, L. A. (1999). End-point interpretation in clinical trials: The case for discipline. *Controlled Clinical Trials, 20*, 40-49.

Shuster, J. J. (1990). *Handbook of sample size guidelines for clinical trials.* Boca Raton: CRC Press.

Thompson, S. G., & Barber, J. A. (2000). How should cost data in pragmatic randomized trials be analyzed?. *British Medical Journal, 320*, 1197-1200.

Tukey, J. W. (1993). Tightening the clinical Trial. *Controlled Clinical Trials, 14*, 266-285.

Voutilainen, P. E. (2001). Assessment of grouping variable should have been blind in trial of dementia. *British Medical Journal, 322*, 1491.

Weinberg, J. M., & Lagakos, S. W. (2001). Linear rank tests under general alternatives, with application to summary statistics computed from repeated measures data. *The Journal of Statistical Planning and Inference, 96*, 109-127.

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

Williams, J. G., Cohen, D. R., & Russell, I. T. (2000b). Authors' reply. *British Medical Journal, 320*, 7251, 1730.

World Association of Medical Editors (2001). Report of the world association of medical editors: Agenda for the future. *Croatian Medical Journal, 42*(2), 121-126.

Appendix

Sample Informed Consent Document for Statisticians to Provide to Medical Researchers (or Vice Versa)

By signing this form you agree that you have been informed of the following. In trusting me with your data, you recognize that I might perform analyses that are technically correct only if various conditions are true. The reality is that these conditions could not possibly be true. Yet in basing the analyses on the truth of these conditions we can follow the tradition of using such parametric analyses. It is unlikely that the results we obtain will differ very much from those we would have obtained had we used exact methods, which are readily available. In fact, it would not be difficult for me to compare the approximate results to the exact results, but I will not do so, because, as stated, it is unlikely that they will differ by very much. This means that there is the possibility that the parametric results will differ sufficiently from the exact results to lead to different conclusions. These conclusions may then be inappropriate, but this would not be discovered right away, because I will not compute the exact results. In the event that in the future it is revealed that damage resulted from the use of improper statistical methods, you agree to indemnify me.