Published in final edited form as:

Stat Med. 2009 July 20; 28(16): 2170–2184. doi:10.1002/sim.3581.

Student T-tests for Potentially Abnormal Data†

Jonathan J. Shuster^{1,*†}

¹Department of Epidemiology and Health Policy Research, College of Medicine, University of Florida, Gainesville, FL 32610, U.S.A.

SUMMARY

When the one sample or two-sample t-test is either taught in the class room, or applied in practice to small samples, there is considerable divergence of opinion as to whether or not the inferences drawn are valid. Many point to the “Robustness” of the t-test to violations of assumptions, while others use rank or other robust methods because they believe the t-test is not robust against violations of such assumptions. It is quite likely, despite the apparent divergence of these two opinions, that both arguments have considerable merit. If we agree that this question cannot possibly be resolved in general, the issue becomes one of determining, before any actual data have been collected, whether the t-test will or will not be robust in a specific application. This paper describes Statistical Analysis System (SAS) software, covering a large collection of potential input probability distributions, to investigate both the null and power properties of various one and two sample t-tests and their normal approximations, as well as the Wilcoxon two-sample and sign-rank one sample tests, allowing potential practitioners to determine, at the study design stage, whether the t-test will be robust in their specific application. Sample size projections, based on these actual distributions, are also included. This paper is not intended as a tool to assess robustness after the data have been collected.

Keywords

Robustness; Satterthwaite approximation; sign-rank test; T-test; Wilcoxon test

1. INTRODUCTION

An investigation into the robustness of the one and two-sample t-tests in specific applications is timely, as this paper is being written during the centennial of Student's seminal paper [1]. An exceptional historical perspective of the development of the t-test, both prior to and up to 25 years after the publication, including its refinement into its current forms, appears in an article by Senn [2]. One of the references in Senn [2], attributed to Pfanzagl and Sheynin [3], in fact shows that Student may have been predated by as much as 32 years by Lüroth [4] in discovering the t-distribution. Senn [2] also contains a number of closely related references concerning or from that time period. For other historical perspectives, see Hanley, Julien, and Moodie [5] and Zabell [6] (with discussion). A healthy skeptic would argue that there is no such thing as normally distributed data, and hence the t-test will violate its assumptions for all small sample applications. If indeed the t-test is theoretically invalid, then why is it still so widely used today? Those who apply it routinely believe in its robustness. Some rather remarkable investigations, for example D'Agostino,

†This paper is a study design article, useful to consultants, clients, teachers, and students.

*Correspondence to: Jonathan J. Shuster, Department of Epidemiology and Health Policy Research, College of Medicine, University of Florida, P.O. Box 100177, Gainesville, FL 32610-0177, U.S.A..

†jshuster@biostat.ufl.edu

Chase, and Belanger [7], Sullivan and D'Agostino [8] and for equal sample sizes, Zimmerman [9] show excellent robustness in situations previously thought otherwise. On the other hand, Strasak and colleagues [10], who clearly do not believe in robustness properties of the t-test, criticize contributors to the New England Journal of Medicine for failing to verify assumptions when they utilized t-tests. In addition, a propensity for outliers can make the t-test, under the null hypothesis, reject far below the nominal rate, as well as yield counterintuitive results. As Dr. Ronald Shiffler (now Dean of the College of Business Administration at Georgia Southern University) pointed out in a personal communication, while he was a PhD student at the University of Florida, if you move any single observation far enough in either direction, the result of a significant t-test will become non-significant. That outlying observation is seen as different from all others, while the t-test sees the remaining observations as essentially the same. This principle is very nicely illustrated by Senn [11], where his Figure 1 plots the value of the t-statistic as a function of one single observation (the tenth), leaving the other nine fixed. Rank methods do not have this problem. Hence, for a more complete assessment, the corresponding properties of the two most popular rank-based methods, the Wilcoxon two-sample and sign-rank one sample tests are included as output in the macros.

In the next section, we describe SAS macros that can be used to help one design hypothesis driven research studies, including an assessment of the robustness of the t-test in a specific application. Section 3 provides 43 numeric illustrations covering five practical experimental situations, while the final section is devoted to a discussion.

2. SAS MACROS

When one is conducting a hypothesis driven experiment, some information about the underlying response profile is typically available from pilot or historical studies. In fact, good information is needed to conduct a sample size calculation. The array of distributions chosen for the macros is designed to be very simple and very flexible. They can designate the cumulative distributions at up to ten points (discrete or continuous). This should be sufficient accuracy to adequately approximate virtually any pilot data derived distribution. The macros conduct replications of the experiment to estimate the probability of rejection (power) at a user prescribed two-sided nominal p-value, under the null or alternative hypothesis, via a Monte Carlo experiment. Instructions on how to use the SAS macros are given later in subsection 2.1. The basic principle is that the user supplies the following input:

- Sample size information.
- Hypothesized center (mean or median) for one-sample macro only
- Type I error (typically 0.05)
- Seed numbers for generating the random variables
- Number of replications of the Monte-Carlo experiment
- Whether the distributions are discrete or continuous
- Cumulative distribution information at up to 10 points

For One Sample situations, use this macro:

```
%macro srpower(n,center,alpha, reps, seed,dc,
y1,f1,y2,f2,y3,f3,y4,f4,y5,f5,y6,f6,y7,f7,y8,f8,y9,f9,y10,f10);
```

For Two Sample situations, use this macro:

```
%macro wilpower(n1, n2,alpha, reps, seed1,seed2,dc,
y1,f1,g1,y2,f2,g2,y3,f3,g3,y4,f4,g4,y5,f5,g5,
y6,f6,g6,y7,f7,g7,y8,f8,g8,y9,f9,g9,y10,f10,g10);
```

The input parameters are described in Table I below.

Note that when using continuous distributions (Parameter dc=1), the distributions are defined as follows. First, the cumulative distribution at each y value is defined by f and g. The mass between consecutive y values is distributed uniformly over the interval. If the first y (y1) is positive, and f1(g1) is not zero, the respective mass f1(g1) is uniformly distributed between 0 and y1. If y1 is not zero, but f1(g1) is zero, then no mass below y1 is allocated for the respective distribution f(g). With dc=1, the example below yields a continuous uniform distribution over the interval from (−1 to 1) for the f's. For the g's a continuous distribution, with 10 % of the mass is uniformly distributed from −2 to −1, 20% of the mass is uniformly distributed from −1 to 0, 30 % of the mass is uniformly distributed from 0 to 1, and the remaining 40% of the mass is uniformly distributed from 1 to 2. The sample sizes are 27 and 54, the type I error is $\alpha=0.05$, 100000 replications are to be done, with seeds 6371 and 7357.

For ease of data entry, it is convenient to write the macro call as below, with one line for each statement about y,f, and g. The one sample macro spower would look very similar, except there is only one sample size, a hypothesized center is needed, there is only one seed, and only one distribution is provided (f's).

```
%wilpower(27,54,.05,100000,6371,7357,1,
-2,0,0,
-1,0,.1,
0,.5,.3,
1,1,.6,
2,1,1,
.....,
.....,
.....,
.....,
.....);
```

The output of each macro includes empirical power estimates and standard errors for each method, population means and standard deviations, and reiterates all of the input parameters. Total run time is also provided. The one sample macro covers the classical t-test, its normal approximation, and the large sample normal approximation to the sign-rank test as cutoffs. The two-sample macro covers the t-test, the normal approximation to the t-test, the z-test (unpooled variance) by the normal approximation and the Satterthwaite [12] approximation, as well as the Wilcoxon test. All tests are two-sided. The output also provides sample size projections for 80% and 90% power for the two-sided alpha level test, when the input is under non-null conditions. We use the “Brownian motion” projection per Shuster and Chang [13], a concept borrowed from sequential methods. For the one sample t-test, the sum of observations is asymptotically Brownian motion for large sample sizes, with sample size taking the role of time. In the two sample case with approximately fixed sampling weights allocated to the two groups $\{W_1=\theta, W_2=(1-\theta)\}$, the process $[\Sigma Y_{1j}/\theta - \Sigma Y_{2j}/(1-\theta)]$ is approximately Brownian motion with sample size again taking the role of time. For the two sample Wilcoxon test, again with sampling weights θ and $(1-\theta)$, Shuster, Chang, and Tian [14] demonstrate the Brownian motion connection with sample size serving as the time variable. Virtually the same derivation as Shuster, Chang, and Tian [14] can be used for the Wilcoxon sign-rank test, yielding asymptotic Brownian motion with the role of time being once again played by sample size. For all four statistical methods, the sample size to produce power π , at two-sided type I error α takes the form:

$$n=K\left[\left\{\Phi^{-1}(\pi)-\Phi^{-1}(\alpha/2)\right\}/\Delta\right]^2, \quad (2.1)$$

where K is a constant, Δ is the effect size, and $\Phi^{-1}(\gamma)$ represents the inverse function of the cumulative distribution of the standard normal distribution.

Projection of the Sample size needs (m_j) by Brownian motion, based on the output from the macro is

$$m_j=n_j\left[\left\{\Phi^{-1}(\gamma)-\Phi^{-1}(\alpha/2)\right\}/\left\{\Phi^{-1}(\pi)-\Phi^{-1}(\alpha/2)\right\}\right]^2 \quad (2.2)$$

where n_j = actual sample sizes used in the macro, and γ is the empirical power for the actual sample size used in the macro. This equation is obtained from (2.1), substituting the actual estimated power γ and desired power π into (2.1) and noting that K and Δ are the same.

When the empirical power is insufficient for a Brownian motion approximation to be considered reliable (arbitrarily defined as below 20%), the output also provides the classical confidence interval inversion sample size projection for the parametric methods as described below.

For the one sample parametric test when the Brownian motion projection is deemed unreliable, we employ (2.3) below, adopted from formula (5) of Case and Ambrosius, Chapter 19 [15].

$$m_1=\left[\sigma\left\{\Phi^{-1}(\pi)-\Phi^{-1}(\alpha/2)\right\}/(\mu-\text{center})\right]^2 \quad (2.3)$$

where μ and σ are the population mean and standard deviation as calculated from the actual distribution of the y_j and center is the hypothesized mean.

For the two sample parametric test, when the Brownian motion projection is deemed unreliable, we employ (2.4) below, adopted from formula (7) of Shuster, Chapter 12 [16]

$$m=\left\{\sigma_1^2/\theta+\sigma_2^2/(1-\theta)\right\}\left[\left\{\Phi^{-1}(\pi)-\Phi^{-1}(\alpha/2)\right\}/(\mu_1-\mu_2)\right]^2 \quad (2.4)$$

where the μ 's and σ 's are the population means and standard deviations as calculated from the actual distributions of the y_j with θ , the fraction of observations assigned to group 1. The individual projected sample sizes are rounded from

$$m_1=\theta m \text{ and } m_2=m-m_1$$

It may take an extra iteration to refine the estimates, if after looking at the empirical power, it is determined to be too high or too low side against projections.

2.1. How to use the SAS macros

It is presumed that the user has access to PC SAS on a stand alone machine or through a network server to use the macros.

Step 1: (first time use only). Go to the following website http://ehpr.ufl.edu/sample_size_programs

- (a) Using Internet Explorer or another browser that is directly linked to SAS, double click on the appropriate link (onesample.sas or twosample.sas) under SAS

Monte Carlo Design (T-Test). This will load the program into the SAS Program window.

- (b) Otherwise, (or if this does not work), right click on the appropriate link (onesample.sas or twosample.sas) and select “Save Target As”, and save this to an accessible location (hard drive or folder). Next, using Windows Explorer, launch the program by double clicking on the appropriate icon (onesample.sas or twosample.sas)

Step 2 (first time only). When the program is loaded in the SAS program window from Step 1(a) or step 1(b), press the function key F3, to launch the program to exactly replicate two of the numerical examples used in the next section (1.3 with $n=20$ for one sample, and 2.12, $n_1=27$ and $n_2=54$ for the two-sample.) The output will exactly replicate the results as in Tables III and V.

Step 3. For a user supplied application, scroll down to lines 381–392 (One-sample macro) or lines 450–461 (Two-sample macro) and replace the title and data from the supplied example with the application's title and data. Internally, the macros have complete descriptions and instructions (in Green) in lines 26–165 (One-sample) and lines 30–156 (Two-sample). Save the program under a new name in a permanently accessible location.

Step 4. Run the program by pressing the function key F3, and save the output as a .RTF file (rich text format is recommended.). There is a simple facility included to automatically save to rtf described in lines 1–9 in both programs. This output is prettier than using “save as” from the file command menu of the output window. For first time use, it is recommended that this facility be ignored. The macro will still run.

Note that the macros can be stacked to run several applications in series. We did all of the numerical examples in just two runs (one for one sample applications and one for the two sample applications.)

Note: If the work is saved per Step 3, the program can be accessed via Windows Explorer, by double clicking on its icon, and the program rerun by pressing the F3 function key once loaded into the program window.

Note: The program needs to be run separately for the null and alternate hypotheses. For the one sample t-test, the null center parameter is the mean of the distribution of f 's. In addition, for the sign-rank test, the distribution of f 's must be symmetric about this mean. For the two-sample null parametric case, the distribution of the f 's and g 's must have the same mean. In addition, for a true null distribution of a Wilcoxon test, the corresponding f 's and g 's must be identical. We see in Example 5 in the next section, where the f 's and g 's have the same means but the distributions are different, the Wilcoxon test performs poorly.

3. NUMERICAL EXAMPLES

Five experimental situations are illustrated. One sample examples have Macro input parameters listed in Table II and Macro output parameters listed in Table III. Tables IV and V provide the corresponding input and output for the two-sample examples, respectively. Note that we employ cumulative distributions, which in this author's experience, lend themselves well to description by biomedical researchers.

Example 1: One Sample Binary Outcome.

This is done in part for quality assurance of the operation of the macro.

A phase II clinical trial of response has a historical success rate of 60%, based on sufficient information to consider the value as known. A new therapy, with the goal of a 20% improvement to 80%, will be tested via a single treatment study. Since two-sided testing is virtually the standard, we shall employ a two-sided test. The input parameters describing the one sample macro are described under illustrations 1.1 (Null) and 1.2 (Power) in Table II.

It is a good exercise to ask the client how many subjects would be an ideal number for the study. Our client answers $N=20$. Therefore, we run the first simulations under the null and under the alternative for $N=20$. As seen in Table III, illustration 1.1, the null situation rejects more than the prescribed 5% for both the t-test and normal approximation. [It is silly to report the Wilcoxon sign-rank test for this application as it tests nothing meaningful, but the P-values for the sign-rank test are used for quality assurance, and agree with the exact binomial values within reasonable sampling error.] Power, at 63%, is somewhat below the intended 80%. The output suggests that a sample size of 30 would get us to 80% power. Hence, a second set of applications appear in illustrations 1.1 and 1.2. The power projection is still lower than ideal, and hence a new projection of 33 is utilized. The t-test still rejects slightly more than the prescribed 5%. It is surprising that the t and normal approximations do not differ with respect to power but do with respect to type I error. One outcome in question, namely 14 events in a sample of 33 is quite likely (1.8% for a success rate of 60%), but virtually zero for a success rate of 80%. This outcome is not significant for the t-test, but significant for the normal approximation.

Example 2. Two-sample Binomial Outcomes.

This example is done mainly for quality assurance purposes. A small randomized phase II trial is designed to reduce the failure rate, $P(Y=2)$, from the current 80% to 40%.

The investigators, based on a large sample approximation, calculates that they needed 22 subjects in each group. Illustrations 2.1 and 2.2 of Tables IV and V provide the empirical power under the null and alternate hypotheses. These are remarkably close to the asymptotic expectations. The exact values from binomial calculations appear in *Italics* in Table V, and the macro output is well within sampling error.

As noted by D'Agostino, Chase, and Belanger [7], the t-approximation does remarkably well in small sample applications to two-sample testing in binomial data. They found that the uncorrected chi-square is also robust, but somewhat less so than the t-test. This is not all that surprising in that it can be shown that for binary data, T , the t-statistic has the following relationship to χ^2 , the chi-square statistic:

$$T^2 = (n - 2)\chi^2 / \{n - \chi^2\} \quad (3.1)$$

Should we have been docking student exam scores for using the t-test, rather than the chi-square test, when they are asymptotically equivalent?

Example 3: A crossover trial.

While this example uses simplified distribution scenarios, it illustrates the kind of probing that can lead to selection of an analytic method before any real data are collected. This problem uses illustrations 1.3, 1.4, 1.5, and 1.6 in Tables II and III, as well as illustrations 2.3, 2.4, 2.5, and 2.6 in Tables IV and V.

Consider a randomized crossover study where each subject will receive in random order each of two treatments. We define the effect size parameter as the arithmetic mean of the

period 1 and period 2 target population effect sizes for treatment 2 minus treatment 1. The objective is to test the null hypothesis that this is zero. This hypothesis is identical to that posed in equation (2.4) of Tudor and Koch [16]. Tudor and Koch [16] and Shuster [17] discuss analytic considerations, based on saturated linear models for the nonparametric and parametric cases, respectively, and hence we only discuss a few elements here. The one sample approach simply takes the differences (treatment 2 minus treatment 1) and tests the hypothesis by a one sample test such as the t-test or sign rank test. The two sample approach takes the period 2 minus period 1 differences and compares the **orderings** via a two-sample t-test or Wilcoxon test. In this example, we look at the impact of a deterministic period effect (adding constant to all period two values). Such an addition would make the target population paired differences between treatments 2 and treatment 1 differ by ordering. Both methods are applicable for a completely randomized design, which assigns patient ordering by a coin flip. In the presence of such a deterministic period effect, the two sample design is more efficient. In addition, when the actual assigned numbers of subjects differ by ordering, the one sample, but not the two-sample parametric method yields conditionally biased estimates of the effect size. Other valid but different approaches to crossover designs can be found in Senn [18].

Illustrations 1.3 and 2.3 represent null situations under the absence of any deterministic period effect, namely discrete uniform distributions over the points $-2, -1, 0, 1$, and 2 . Illustrations 1.4 and 2.5 cover the one sample and two sample cases respectively under the alternative hypothesis, where there is no period effect. Illustrations 1.5 and 2.4 represent the null distributions where a deterministic period effect of $+1$, is added to period 2. For example, for the one sample case, -3 occurs in the situations where the “no period effect illustration” (1.3) was -2 (treatment 2-treatment 1), if and only if the treatment ordering was treatment 2 then treatment 1. This would have probability of $0.5(0.2)=0.1$. For the period effect on the “period 2 minus period 1 two sample example”, the values in illustration 2.6 are derived from those in 2.5 by simply adding one unit to each y_j value.

Under the null hypothesis, both the t-tests and nonparametric tests are close to the desired nominal rate of 0.05, but the normal approximations for the t-test reject somewhat more often than the nominal level. Note that when there is a period effect, the power of the one sample test is seriously altered, as compared to when there is no such effect. The mean difference does not change with vs. without the period effect, but the standard deviation increases. On the other hand, we obtain exactly the same power with or without the period effect in the two sample case. The required sample sizes are similar with no period effect for the one sample vs. two sample situations (39 vs. 38), but are substantially different with the period effect (61 vs. 38).

Example 4: An example dealing with anticipated outliers.

A study, designed to assess the impact of a therapeutic quality of life intervention on length of hospital stay after elective knee replacement, involved considerable interaction with the investigators. The planning parameters are somewhat simplified for this illustration. Suppose the insurance coding allows a stay of one, two, or three days, unless a complication occurs, where a ten day stay is permitted. Although the investigator felt that the intervention would have no impact on this complication rate, in accordance with standard “intent-to-treat principles”, he was discouraged from throwing away these observations. The planning null parameters are in illustration 2.7. Basically, there is about a 10% chance of complication, and otherwise, the distribution is anticipated to be uniform over 1, 2, and 3 days. The alternative distribution is in illustration 2.8. Note that the Wilcoxon test has accurate type I error properties, whereas the various two-sample tests have mixed behaviors, especially when $N=10$ per group, where the normal approximations do well, and the t-approximations

reject somewhat below the nominal rate of 5%. As for power, one needs about eight times as many patients (839 vs. 102), if one elected to use the t-test rather than the Wilcoxon test. One might wonder what effect a transformation might have upon the power. We studied a natural log transformation, as when the confidence interval is exponentiated back to its original units, it has a scale interpretation, something meaningful to clinical investigators and insurance companies. Other transformations may lead to difficult to interpret effect size estimates. The null (alternative) appear in illustrations 2.9 (2.10). Of course, the Wilcoxon test is unaffected. But the t-test requirements drop the total sample size requirement to 180 (90 per group), much better than without the transformation, but far more than the total of 102 needed by the Wilcoxon test. On the basis of these analyses, launched before a single patient had been recruited,, a highly efficient study can be designed.

Example 5: An example with unequal variances.

In small placebo controlled clinical trials, we often anticipate that the experimental group has a more volatile response than the control group. Even though we lose the highly desirable permutational basis for our inference, we may still have keen interest in the comparison of mean responses. This example, with simplified data for illustrative purposes, deals with exactly this situation for changes in hormone values. This randomized controlled two-sample trial will try to answer the question as to the mean impact of an intervention upon hormone levels. It is expected that the relative variation in controls will be much smaller than that in experimental subjects. The planning null and alternative distributions are given in illustrations 2.11 and 2.12 (continuous distributions with probability mass uniformly distributed over the subintervals immediately below the cumulative value). Since under the alternative hypothesis, the standard deviation of the experimental group (g's) is about twice that of the control group (f's), we look at designs near the optimal allocation (in proportion to the standard deviations, or two experimental subjects per control) as well as the more traditional equal allocation. Note that when the sample sizes are equal, all the competing parametric tests perform similarly both in type I error and power. When the sample sizes are unequal, the methods that pool the variance estimator tend to overestimate the true variance, and therefore tend to have type I error and power substantially lower than the methods relying on unpooled estimates of variance. The Satterthwaite corrected method had null rejection rates very close to 0.05 throughout illustration 2.11. The Wilcoxon test, while a valid test for comparing populations, is not appropriate for comparing these means or medians, since we expect in advance that one group has different variability than the other. This is reflected by the test rejecting more than the nominal 5% level when the sample sizes are equal, but less than the nominal level when the allocation is two experimental subjects per control.

4. DISCUSSION

It needs to be noted that as presently marketed, except for one case as noted below, the output of these macros cannot be obtained directly from commercial software packages, such as SAS Proc Power (<http://www.sas.com/index.html>) or nQuery (<http://www.statsol.ie/index.php?pageID=2>). Neither of these packages allow for the user to input the relevant actual probability distributions, to produce results that do not depend on either asymptotics or on normal assumptions. For nQuery, a user can input discrete probability distributions to obtain asymptotic power and sample size for the Wilcoxon test. However, the package StatXact8 (<http://www.cytel.com/>) can perform exact and simulated power and sample size calculations for the Wilcoxon test. One feature of the macros under non-null situations is that a user can put in any sample size, run the macro, and it will estimate the sample size actually needed. The real process, as seen in the examples, is iterative in nature. A reader may wonder why we use Brownian motion projections when feasible, over the more usual

asymptotic sample size calculations. The answer is that irrespective of the planning parameters, formulas such as (2.3) and (2.4), or their counterparts for the nonparametric methods, are static, and will give you the same answer irrespective of the input sample sizes. Because the Brownian motion projection uses the actual empirical power, it will provide a far more reliable estimate of what sample size to try next in the iteration.

These macros have a number of practical uses. For instructors, the central limit theorem can be illustrated by varying the sample size in the one sample macro, and varying the value of alpha. The normal approximation of the two sample t-test under null conditions can also be demonstrated by varying the sample sizes and alpha values. The impact of lack of equal standard deviations can be related to the magnitude of the difference in standard deviations and the magnitude of the sample size inequality. Both of these issues are especially important in observational studies, but example 5 shows it to have potential importance in some randomized trials as well. When students ask, how large need the sample sizes be before the normal approximation is appropriate, there is a tool to look at this on a case by case basis. Some instructors have used 30 as a magic number, but in reality, there is no general answer. The mean of 30 independent identically distributed Gamma random variables with density given in (4.1) below with shape parameter $\alpha=1/30$ is exponential, hardly Gaussian in appearance.

$$f(x) = \{\Gamma(\alpha)\}^{-1} \beta^\alpha x^{\alpha-1} \exp(-x/\beta) \quad (4.1)$$

With the exception of non-robustness to a propensity for outliers as cited by Senn [11], the issue of robustness of the t-tests (including the versions that correct for unequal variance) is hard to pin down logically. For example, the exponential distribution is highly skewed, but the t-tests not only perform well for modest sample sizes in terms of type I error, but they are locally most powerful tests in both the one and two sample settings. This should convince us to be open minded about whether or not the t-test will be appropriate for an individual application. Since we usually have some pilot data from other studies when we design a hypothesis driven study for our colleagues, it seems to be a good investment, as part of our study design, to investigate whether or not the t-test would be appropriate for this upcoming study. The side by side (matched) comparison with the Wilcoxon test allows one to choose the appropriate analysis, and estimate power, before a single observation is collected.

In consulting work, as illustrated in examples 3, 4, and 5, the macros give a very good picture of what we can expect once the study data are actually collected. The examples provide illustrations of the kinds of decision making about both study design and analysis that can be very convincing for statisticians, clinical scientists, and grant reviewers. Formal tests for assumptions, after data are collected, rather than at the design phase, is fraught with high risk, and likely, will destroy the perceived operating characteristics, Type I error and power. See Shuster [19] for a compelling example, which can be replicated using a two sample Gamma macro, not illustrated here, but available from same website described in subsection 2.1. The macro calls at the end of the Macro do actually replicate the example in [19], with a different seed process from the original. All together, the three available macros enable a user to completely avoid the use of diagnostics after data collection, by anticipating the actual distributions. The actual method of analysis would be locked into place prior to the initiation of the study. In addition, for those with SAS programming experience, the replacement of the “RANGAM” function in the Gamma Macro by (RAN) or (RANUNI plus INV) functions, can expand the range of applications to other parametric distributions. The Gamma macro comments in green on lines 23–35 describe how to do this.

Rigorous quality control was applied to the completed macros. To ensure that the macro was correctly generating the observations according to the desired distributions, the sample means and population means were compared on a wide variety of input distributions, for each macro. Second, to assure the empirical results were being done correctly, we applied the macro to binary outcome data, well beyond those illustrated in Examples 1 and 2, for both the one sample and two sample situations, where exact results were calculated directly by the SAS binomial function. The results confirmed beyond any doubt that the macros generate the correct results.

An excellent use of these macros occurs when a clinical colleague is proposing a multi-treatment single center clinical trial. Using this macro, and setting the α level at 0.0167 (0.05/3) for three treatments or at 0.0083 (0.05/6) for four treatments, we can run power analyses under realistic scenarios to see if the resources are there to do this as proposed. It is actually extraordinarily rare that more than two treatments can be accommodated with adequate power at a single center. These exercises can be very convincing, when compared to a mathematical derivation, as the simulations are much easier to understand by clinicians. After all, they were the ones to create the null and alternate distributions.

One useful application of the macro is in tweaking the value of α downward (upward), to account for rejection rates above (below) the prescribed nominal levels. This type of calibration was actually done for Fisher's Exact test in Crans and Shuster [20]. If we want to ensure adequate power when the true type I error is below 5%, we can set slightly different α levels in the program. For example, in illustrations 2.11 and 2.12, where the unpooled Z -test with the normal cutoff rejected slightly more than the desired 5% level for sample sizes of 28 and 56, if we upped the sample size from 84 to 87 (29 in group 1 and 58 in group 2), and set the value of α to 0.046 instead of 0.050, we obtain an empirical type I error for the normal cutoff of the Z -test with unpooled variance to be 0.04943 and the empirical power for this test to be 0.80218 both based on 100,000 replications.

It is possible to use the macros for mixed discrete-continuous distributions. For example if $y_1=0$, $f_1=0.4$, $y_2=1$, and $f_2=1$, and the parameter $dc=1$ (continuous), we would actually have a distribution with 40% of its mass at zero and 60% uniformly distributed between 0 and 1. Perhaps more interesting, if we set $y_1=1$, $f_1=0$, $y_2=1$, $f_2=0.4$, $y_3=2$, $f_3=1$, we obtain a distribution with 40% of its probability at 1 and 60% uniformly distributed between 1 and 2.

There is no reason why these types of macros cannot be developed for more complex designs. This paper covers some of the most basic designs, but the ideas seem readily adaptable to regression models, especially analysis of covariance, and to multi-treatment trials. Sequential methods might also be incorporated into future macros. Another future idea might be to develop a macro that estimates conditional power, in a similar manner to Cantor [21], who did this in a survival analysis setting. The data collected up to the present time are used, along with simulations of the data yet to be collected, to estimate the probability of obtaining a significant result by the end of the study, under both the null and alternate hypotheses. This could allow the trial to be terminated early either for efficacy or futility, with minimal impact on the operating characteristics.

Note that while this paper is devoted to hypothesis testing, good statistical practice dictates that confidence intervals for effect sizes be provided. Another future avenue of research might be to develop algorithms for sample size based on the width of confidence intervals, derived from actual anticipated distributions, rather than on normal assumptions or large sample approximations.

In conclusion, these macros can prove to be very useful additions to the toolbox of instructors and consultants. Usually, when designing biomedical experiments, a clinical

researcher, based on published literature and his/her own pilot data, has good ideas as to how many subjects s/he wishes to include and the difference in distributions s/he expects to find. Before a protocol is written, we can take this information and replicate the hypothetical experiment 100,000 times to gain insight as to the chances the investigator will get a significant result (i.e. power). This concept gets across far more readily than results from abstract formulas.

Acknowledgments

The author wishes to thank Mr. Douglas Theriaque, Director of the Data Services Laboratory, General Clinical Research Center, University of Florida, for his help with respect to making the macros more user friendly, including the automatic creation of rich text output files. Thanks also go to the Associate Editor, who spent considerable effort in helping the author improve the presentation. The author is grateful to the two reviewers, especially to Reviewer #1 who provided important historical information.

Contract/grant sponsor: National Institute of Research Resources, National Institutes of Health; contract/grant number:M01RR00082

REFERENCES

1. Student. The probabilistic error of a mean. *Biometrika*. 1908; 6(1):1–25.
2. Senn SJ. A century of t-tests. *Significance*. 2008; 5(1):37–39.
3. Pfanzagl J, Sheynin O. Studies in the history of probability and statistics .44. A forerunner of the t-distribution. *Biometrika*. 1996; 83(4):891–898.
4. Lüroth J. Vergleichung von zwei Werten des wahrscheinlichen Fehlers. *Astron Nachr*. 1876; 87:209–220.
5. Hanley JA, Julien M, Moodie EAE. [t DISTRIBUTION CENTENNIAL] Student's z, t, and s: What if Gosset had R? *The American Statistician*. 2008; 62(1):64–69.
6. Zabell SL. On Student's 1908 article “The probabilistic error of a mean”. *Journal of the American Statistical Association*. 2008; 103(1):1–6.
7. D'Agostino RB, Chase W, Belanger A. The appropriateness of some common procedures for testing the equality of two independent binomial populations, (C/R: 1989V43 p187–188). *The American Statistician*. 1988; 42(3):198–202.
8. Sullivan LM, D'Agostino RB. Robustness of the t test applied to data distorted from normality by floor effects. *J. Dent. Res*. 1992; 71(12):1938–1943. [PubMed: 1452898]
9. Zimmerman DW. Two separate effects of variance heterogeneity on the validity and power of significance tests of location. *Statistical Methodology*. 2006; 3:351–374.
10. Strasak AM, Zaman Q, Marinell G, Pfeiffer KP, Ulmer H. The use of statistics in medical research: A comparison of *The New England Journal of Medicine* and *Nature Medicine*. *The American Statistician*. 2007; 61(1):47–55.
11. Senn SJ. The t-test tool. *Significance*. 2008; 5(1):40–41.
12. Satterthwaite FW. An approximate distribution of estimates of variance components. *Biometrics Bulletin*. 1946; 2:110–114. [PubMed: 20287815]
13. Shuster JJ, Chang MN. Second guessing clinical trial designs. *Sequential Analysis*. 2008; 27(1):2–20.
14. Shuster JJ, Chang MN, Tian L. Design of group sequential clinical trials with ordinal categorical data based on the Mann-Whitney-Wilcoxon test. *Sequential Analysis*. 2004; 23(3):413–426.
15. Case, LD.; Ambrosius, WT. Power and Sample Size. Chapter 19 in *Methods in Molecular Biology: Elementary Biostatistics*. Ambrosius, WT., editor. Humana Press; Totawa NJ: 2007. p. 377-408.
16. Tudor G, Koch GG. Review of nonparametric methods for the analysis of crossover studies. *Statistical Methods in Medical Research*. 1994; 3:345–381. [PubMed: 7889227]
17. Shuster, JJ. Design of Experiments. Chapter 12 in *Methods in Molecular Biology: Elementary Biostatistics*. Ambrosius, WT., editor. Humana Press; Totawa NJ: 2007. p. 235-260.
18. Senn, SJ. *Cross-over Trials in Clinical Research*. Wiley; Chichester: 2002.

19. Shuster JJ. Diagnostics for assumptions in moderate to large simple clinical trials: Do they really help? *Statistics in Medicine*. 2005; 24:2431–2438. [PubMed: 15977289]
20. Crans GG, Shuster JJ. How conservative is Fisher's exact test: A quantitative evaluation of the two-sample binomial trial. *Statistics in Medicine*. 2008; 27:3598–3611. [PubMed: 18338319]
21. Cantor AB. Power estimation for rank tests using censored data: Conditional and unconditional. *Controlled Clinical Trials*. 1991; 12:462–473. [PubMed: 1657525]

Table I

Parameters of the macro and their interpretation

Parameter	Interpretation
n	Sample Size (srpower)
n1,n2	Sample sizes for group1, group 2 (wilpower)
center	Hypothesized Mean or Median (srpower)
alpha	Type I error, typically 0.05 (two-sided)
reps	Number of replications of experiment (usually at least 10,000)
seed	srpower
seed1, seed2	wilpower large odd numbers to seed the randomization ($<2^{31}$) This gives complete reproducibility for good documentation.
dc	dc =0 for discrete dc=1 for continuous
yj	y1 y2 y3 .. y10 are the outcome values (if fewer, say 5, y6,...y10 are missing, use a dot=.)
fj	Cumulative null distribution $P(Y = y_j) = f_j$ If dc=1 then we distribute mass (f1) at y1 uniformly over (0,y1), Mass (f2-f1) at y2 uniformly over (y1,y2), etc. See below if f1=0 and dc=1. (if fewer, say 5, f6,...f10 are missing, use a dot=.)
gj	Wilpower Cumulative alternative distribution $P(Y = y_j) = g_j$ If dc=1 then we distribute mass (g1) at y1 uniformly over (0,y1), Mass (g2-g1) at y2 uniformly over (y1,y2), etc. See below if g1=0 and dc=1 (if fewer, say 5, g6,...g10 are missing, use a dot=.)

%wilpower (two sample macro) and %srpower (one sample macro)

If f1=0 and dc=1 the there is no mass distributed between 0 and y1.

If g1=0 and dc=1 the there is no mass to distributed between 0 and y1.

Table II

One Sample Input Parameters.

ID	Ex	yj (fj)	Center Hypothesized	N
1.1	1N	0(.4), 1(1)	0.6	20, 30, and 33
1.2	1A	0(.4), 1(1)	0.8	20, 30, and 33
1.3	3N	-2(.2), -1(.4), 0(.6), 1(.8), 2(1)	0	20 and 39
1.4	3A	-2(.3), -1(.6), 0(.8), 1(.9), 2(1)	0	20 and 39
1.5	3N	-3(.1), -2(.2), -1(.4), 0(.6), 1(.8), 2(.9), 3(1)	0	20 and 61
1.6	3A	-3(.15), -2(.3), -1(.55), 0(.75), 1(.9), 2(.95), 3(1)	0	20 and 61

Seed for all 1711, replications for all=100000. All use DC=0 (discrete) $\alpha=0.05$ for all.

See Table I for a definition of terms.

Ex refers to Example Number in Section 3, with N=Null and A=Alternative.

Table III

One Sample outcomes.

ID	N	Power Sign-Rank	Power T-test	Power Normal	N80 Non-Par	N80 Parametric	
1.1	20		0.07107	0.07107			
			<i>0.07198</i>	<i>0.07198</i>			
	30		0.06528	0.06528			
			<i>0.06476</i>	<i>0.06476</i>			
	33		0.05817	0.07626			
			<i>0.05782</i>	<i>0.07546</i>			
1.2	20		0.62803	0.62803		30	
			<i>0.62966</i>	<i>0.62966</i>			
	30		0.76008	0.76008		33	
			<i>0.76079</i>	<i>0.76079</i>			
	33		0.79924	0.79924		33	
			<i>0.79996</i>	<i>0.79997</i>			
1.3	20	0.04204	0.05184	0.06463			
	39	0.04615	0.04966	0.05721			
1.4	20	0.47330	0.51295	0.55312	44	39	
	39	0.78285	0.79135	0.80688	41	40	
1.5	20	0.04216	0.05177	0.06462			
	61	0.04808	0.05102	0.05540			
1.6	20	0.32728	0.35890	0.39941	69	61	
	61	0.80291	0.80202	0.81243	60	61	

N=actual sample size Power is empirical estimated power at $\alpha=0.05$. N80 represents the sample sizes required for 80% power. In 1.1 and 1.2, exact binomial values are in *Italics*. Input parameters are described in Table II.

Table IV

Two Sample Input Parameters.

ID	Ex	yj (fj,gj)	(N1,N2)
2.1	2N	1(.2,.2), 2(1,1)	(22,22)
2.2	2A	1(.2,.6), 2(1,1)	(22,22)
2.3	3N	-2(.2,.2), -1(.4,.4),0(.6,.6),1(.8,.8),2(1,1)	(10,10) and (18,18)
2.4	3N	-1(.2,.2),0(.4,.4),1(.6,.6),2(.8,.8),3(1,1)	(10,10)
2.5	3A	-2(.3,.1),-1(.6,.2), 0(.8, .4),1(.9,.7),2(1,1)	(10,10) and (18,18)
2.6	3A	-1(.3,.1),0(.6,.2), 1(.8, .4),2(.9,.7),3(1,1)	(10,10)
2.7	4N	1(.3,.3), 2(.6,.6),3(.9,.9),10(1,1)	(10,10) and (51,51)
2.8	4A	1(.3,.6), 2(.6,.8),3(.9,.9),10(1,1)	(10,10),(51,51) and (420,419)
2.9	4N	0(.3,.3),.693(.6,.6),1.099(.9,.9),2.303(1,1)	(51,51) and (90,90)
2.10	4A	0(.3,.6),.693(.6,.8),1.099(.9,.9),2.303(1,1)	(51,51) and (90,90)
2.11	5N	-2(0,0),-1(0,.25), 0(.5,.5), 1(1,.75), 2(1,1)	(18,18),(12,24),(43,43), (27,54), (46,46), and (28,56)
2.12	5A	-2(0,0),-1(0,.1), 0(.5,.3),1(1,.6), 2(1,1)	(18,18),(12,24),(43,43), (27,54), (46,46), and (28,56)

Seeds for all 6371 and 7357, replications for all=100000 (except 2.8 N1/N2=420/419 used 50000.) All use DC=0 (discrete) except for 2.11 which uses DC=1 (continuous) $\alpha=0.05$ for all. See Table I for a definition of terms.

Ex refers to Example Number in Section 3, with N=Null and A=Alternative.

Table V

Two Sample outcomes.

ID	N ₁ /N ₂	Power Wilcoxon	Power T-test	Power Normal Pooled	Power Satter	Power Normal Unpooled	N80 Non-Par	N80 Parametric
2.1	22/22	0.05258	0.05268 <i>0.05150</i>	0.05452 <i>0.05341</i>	0.05268	0.05462		
2.2	22/22	0.79157	0.79157 <i>0.79115</i>	0.80574 <i>0.80498</i>	0.79157	0.80574	22/22	22/22
2.3	10/10 18/18	0.04836 0.04976	0.05116 0.05109	0.06629 0.05958	0.04995 0.05075	0.06629 0.05958		
2.4	10/10	0.04836	0.05116	0.06629	0.04995	0.06629		
2.5	10/10 18/18	0.48408 0.74894	0.50499 0.75876	0.54995 0.78005	0.50063 0.75791	0.54995 0.78005	21/21 20/20	18/18 19/19
2.6	10/10	0.48408	0.50499	0.54995	0.50063	0.54995	21/21	18/18
2.7	10/10 51/51	0.04957 0.05061	0.03103 0.04987	0.04743 0.05271	0.02258 0.04918	0.04743 0.05271		
2.8	10/10 51/51 420/419	0.23523 0.80424 1.00000	0.10276 0.17731 0.79682	0.13507 0.18299 0.79768	0.08279 0.17692 0.79670	0.13507 0.18299 0.79764	51/51 50/50	420/419 420/419 422/421
2.9	51/51 90/90	0.05061 0.04944	0.05025 0.04983	0.05327 0.05142	0.05011 0.04979	0.05327 0.05142		
2.10	51/51 90/90	0.80424 0.96280	0.55194 0.78661	0.56081 0.79053	0.55168 0.78648	0.56081 0.79053	50/50 50/50	90/90 92/92
2.11	18/18 12/24 43/43 27/54 46/46 28/56	0.06535 0.03012 0.06385 0.03140 0.06405 0.03291	0.05473 0.01914 0.05142 0.01775 0.05094 0.01825	0.06311 0.02332 0.05444 0.01926 0.05405 0.01990	0.05216 0.04981 0.05040 0.05035 0.04992 0.05064	0.06311 0.05820 0.05444 0.05406 0.05405 0.05394		
2.12	18/18 12/24 43/43 27/54 46/46 28/56	0.45055 0.38716 0.80422 0.76289 0.83212 0.77979	0.41283 0.30322 0.76389 0.64370 0.79183 0.66194	0.43850 0.32291 0.77246 0.65643 0.79918 0.67465	0.40529 0.43337 0.76112 0.77456 0.78964 0.79046	0.43850 0.46177 0.77246 0.78394 0.79918 0.79884	42/42 34/67 42/42 30/59 42/42 29/59	43/43 27/54 46/46 28/56 46/46 28/56

N₁/N₂=actual sample sizes per group. Power is empirical estimated power at $\alpha=0.05$. N80 represents the sample sizes projected for 80% power. In 2.1 and 2.2 exact binomial calculations for the tests are in *Italics*. Input parameters are described in Table IV.