UW Biostatistics Working Paper Series

10-1-1997

Multiple Outcomes in Health Services Research: Hypothesis Tests and Power

Donald C. Martin
none

Paula Diehr
University of Washington, pdiehr@u.washington.edu

Thomas D. Koepsell
University of Washington, koepsell@u.washington.edu

Stephan D. Fihn
University of Washington, sfihn@u.washington.edu

MULTIPLE OUTCOMES IN HEALTH SERVICES RESEARCH: HYPOTHESIS TESTS AND POWER

DC Martin*[1], P Diehr[2], T Koepsell,[2] and SD Fihn[1,2]. * martin@biostat.washington.edu, [1] NW VA Health Services Research Program, [2] University of Washington

ABSTRACT

Health services research often is directed towards making small improvements in a number of outcomes that reflect many aspects of the patient's life rather than a large improvement in a single well defined outcome. A researcher might choose five scales to measure different aspects of treatment outcomes and not expect any large treatment differences on any single outcome measure. O'Brien [1984] has proposed a nonparametric statistical procedure which is particularly well suited to this type of problem and that can result in considerable increases in statistical power. As an example consider the sample sizes needed for the 5 variables below and note the much smaller sample size for the pooled rank test.

Variable	σ	σ difference	Δ	Effect Size	n1 = n2
Pooled rank test	-	-	-	.38	108
SF36 -physical	23.9	18.5	5	.27	216
SF36 -General Health	24.4	18.9	5	.26	233
SF36 -Mental Health	20.1	15.6	5	.32	154
Agina-Exertion	25.1	19.4	5	.25	252
Agina-CAD Perception	24.1	18.7	5	.26	233

This paper will briefly review O'Brien's pooled rank method and develop power calculations. A detailed power calculation example will be presented and discussed. Adding outcome variables where the treatment effect is small compared to the variability could reduce the power of the pooled rank test. The effect on power of adding poor outcome variables will be discussed.

Introduction

Frequently an experiment in health Services research is intended to impact a number of different outcomes rather than a single outcome. As an example, a treatment might be expected to improve mobility, make the patient less reliant on other care givers, reduce depression and improve a quality of life score. Often the investigate is advised to pick a single "primary" variable as the outcome and design the study as if there was only one outcome variable. This advice is usually based on the idea that multiple outcomes in a multivariate test may reduce the power or result in an unacceptably complex analysis.

O'Brien [1984] devised a simple and robust way of pooling multiple outcome variables when there are directional hypothesis on each variable and showed that the pooled test can gain power over a multivariate two group test. This paper will review O'Briens test and present a power approximations for this test. O'Brien shows that this test can be more powerful than the multivariate t test.

O'Brien Test

We will consider the two group case where one group is a control group and the other is a treatment group. For the sake of discussion we will assume that each of the variables can be described as having a direction that can be described as an improvement. As an example, a reduced depression score would be an "improvement". We will omit outcome variables that don't have a hypothesized direction of change for a successful treatment impact. O'Brien test changes the signs on outcome variables so that a positive change corresponds to an improvement and ranking all of the variables. Thus the sign would be changed on a depression score where smaller scores indicate less depression. The ranks are then computed for each variable across subjects. The ranks are then summed across outcome variables within individuals to create a summary score. For large sample sizes, say more than 10 in each group, a t test can be used on the summary score. If the sample sizes are small then a permutation t test will result in the exact test size in spite of the small correlations induced between observations by the ranking. The permutation t test is described by Good [1994] or Noreen [1989] among others.

A small example of the O'Brien Test

I fabricated a small, two groups of size 10, data set to illustrate the O"Brien pooled rank method. The second column in Table 1 is a group designator, the next three columns are the three variables. The three columns labeled rv1, rv2 and rv3 are the ranks of the three variables across cases. Note that average ranks were assigned for ties as is the usual convention in ranking. The column labeled pooled is simply the sum of the three ranks for

each case and the average rank column is the average of the three ranks.
Table 1 - Example data, three variables, ranks, pooled rank and average rank

	group	v1	v2	v3	rv1	rv2	rv3	pooled	Ave. Rank
1.	1	30	59	59	4	11.5	11	26.5	8.83
2.	1	40	78	2	7	14.5	1	22.5	7.5
3.	1	17	59	87	2	11.5	18	31.5	10.5
4.	1	39	58	60	6	10	12	28	9.33
5.	1	54	42	47	10	7	5	22	7.33
6.	1	60	29	77	11	4	15.5	30.5	10.17
7.	1	45	49	49	8	8	6	22	7.33
8.	1	51	41	17	9	6	2	17	5.67
9.	1	98	26	51	18	3	7	28	9.33
10.	1	92	50	52	16	9	8	33	11
11.	2	38	97	72	5	19	13.5	37.5	12.5
12.	2	15	89	77	1	17	15.5	33.5	11.17
13.	2	62	75	84	12	13	17	42	14
14.	2	93	95	91	17	18	20	55	18.33
15.	2	23	78	89	3	14.5	19	36.5	12.17
16.	2	76	99	72	13.5	20	13.5	47	15.67
17.	2	99	7	57	19.5	2	9.5	31	10.33
18.	2	99	39	57	19.5	5	9.5	34	11.33
19.	2	76	79	45	13.5	16	4	33.5	11.17
20.	2	78	6	43	15	1	3	19	6.33

The two group O'Brien pooled rank test is essentially by the two group t test on either the pooled ranks or the average ranks. The significance level will be the same for either test. Table 2 shows a two tailed t test on each of the three variables and the t test on the pooled ranks.

Table 2 - t tests on individual variables and the pooled rank

Variable	mean 1	S.D. 1	mean 2	S.D. 2	t	p value
Var1	52.6	25.5	65.9	30.7	1.05	.3062
Var2	49.1	15.5	66.4	35.9	1.40	.1792
Var3	50.1	25.1	68.7	17.4	1.92	.0704
Pooled	26.1	5.09	36.9	9.68	3.12	.0059

Note that none of the individual variables are significant while the pooled rank test is significant at better than the .01 level. This example illustrates both how the O'Brien

pooled rank test can gain power over the individual tests and avoid multiple decision issues such as Bonferoni corrections to the alpha levels. As previously indicated, the t test is an approximation of the exact permutation t test which should be used for small sample sizes. A permutation t test program was not readily available but the Wilcoxon rank sum test is a good alternative to the permutation t on the pooled ranks because it includes an implicit correction for the correlations induced by ranking. The rank sum test on the pooled ranks results in a rank sum of 66 with an exact p of about .002 (the limit of a 3 decimal place table) and a normal approximation of .0032. Thus it looks as if the t test was a good approximation.

Using a statistical package to compute the pooled rank test

The pooled rank test is a three step operation with most statistical packages:

1. Compute the ranks across subjects for each variable.
2. Sum the ranks for each variable for each subject to get the pooled rank.
3. Run the usual parametric test on the pooled ranks. A t test for two groups, a one way ANOVA for more than two groups, etc.

I have omitted the case where one or both samples are small since this we are usually concerned with moderate to large samples. Here are the SPSS commands used for the

Table 3 - SPSS commands for the pooled rank example

```
rank variables = v1 v2 v3.
compute pr = (rv1+rv2+rv3).
oneway pr by group
        /statistics=all.
```

example
and here are the Stata commands:

Table 4 - Stata commands for the pooled rank test example

```
egen  rv1 = rank(v1)
egen  rv2 = rank(v2)
egen  rv3 = rank(v3)
gen pr = rv1+rv2+rv3
oneway pr group
```

The three lines that begin with egen (extended variable generation) compute the ranks for

the three variables. The line that begins with gen (generate a variable) cheats the pooled rank variable as the sum of the three ranks. The line that begins with oneway performs a one way ANOVA on the pooled ranks using the variable group to define the two groups.

Power Calculations

The derivation and rationale for the power calculations are given in Appendix A. We will assume a two group design with pre and post treatment measures and 5 variables, p = 5, as given in table 3. The investigator has assumed a change of 5 points on each scale as a treatment effect. The standard deviations, σ, were taken from another study.

The first step is to approximate the standard deviation of the pre/post difference. We will assume that both the pre and the post measurements have the same standard deviation. Then the standard deviation of the difference depends on the pre and post measure correlation. A pre/post r of .5 should be the worst case since you are better off using only the post measure if the correlation is less than .5. Correlations as high as .9 are rare and the smaller correlations are conservative for power calculations. I usually try values of .6, .7 and .8 which tends to bracket most of the realistic values[1]. I will use a pre post r of .7 for this example. The estimated standard deviation of the pre/post difference is given by:

$$\sigma_{p/p\,i} = \sigma_i \sqrt{2\,(1 - r_i)} \qquad for \;\; i = 1, \ldots, p = 5 \qquad (1)$$

Of course if you have $\sigma_{p/p}$ estimates from other studies these should be used instead of an estimate based on a guess at the correlation. The next step is to compute the effect size for each variable:

$$E_i = \Delta_i / \sigma_{p/p\,i} \qquad for \;\; i = 1, \ldots, p = 5 \qquad (2)$$

If the study is not a pre/post design just use the post as $\sigma_{p/p}$ in the formulas. Then compute the average effect size:

$$E_m = (E_1 + E_2 + \ldots + E_p) / p \qquad (3)$$

---

[1] I suppose that I should offer some apology for guessing at values but to be honest, power calculations always involve a some guesswork. If you knew everything necessary to compute the power, the study would be unnecessary.

which is $E_m = .276$ for the example.

We will need to estimate the average correlation between the ranked outcome variables. I usually assume that the Spearman rank correlations are about the same size as the original Pearson product moment correlations. If we have a variety of different outcomes, the correlations are not usually very large and my usual guessing range is .2 to .5. I will use $r_m = .4$ for the example. Watch out for the case where you have several different measures of the same underlying quantity. These can have much higher correlations and will increase the mean correlation. You may want to use a guess of .7 to .9 for the correlation between alternate measures of the same quantity. The effect size for the pooled rank test is:

$$d = E_m \sqrt{p / (1 + (p - 1) r_m)} \qquad (4)$$

where d is the effect size for the pooled rank test. In this example d = .383. Table 3 presents the calculations for the individual variables and the sample sizes needed for a power of .80 for a two tailed t test at the .05 level for each variable.

Table 3- Sample data for power calculation example

Variable	Δ	σ	$r_{p/p}$	$\sigma_{p/p}$	E	$n_1 = n_2$
SF36 -physical	5	23.9	.70	18.51	.270	216
SF36 -General Health	5	24.4	.70	18.90	.265	225
SF36 -Mental Health	5	20.1	.70	15.57	.321	153
Agina-Exertion	5	25.1	.70	19.44	.257	238
Agina-CAD Perception	5	24.1	.70	18.67	.268	220

If we use this d as the treatment effect in a sample size program with σ = 1 for the two group t test we get an n of 108 for each group which is an improvement over the individual t test sample sizes which range from 153 to 238.

Power and Sample Size Calculations

Given α, n_1, n_2 and d you can compute power or given α, power, and d you can compute the sample size usually assuming equal sample sizes in each group. The following section assumes a two tailed t test and uses α/2. For a one tailed test replace α/2 with α in the formulas.

The usual textbook formulas for the approximate power of a two group t test are:

$$d = (\mu_1 - \mu_2) / \sigma \tag{5}$$

$$z_{(1-\beta)} = d / \sqrt{n_1^{-1} + n_2^{-1}} - z_{(1-\alpha/2)} \tag{6}$$

where the power is $N(z_{(1-\beta)})$ where N is the cumulative normal distribution. Note that d is the effect size and is computed by equation 4 rather than 5 for the pooled rank test. This equation slightly overestimates the power of the t test for moderate samples. An improved approximation which is quite good for more than 5 degrees of freedom and powers greater than .3 is:

$$z_{(1-\beta)} = d / \sqrt{n_1^{-1} + n_2^{-1}}(1 - z^2_{(\alpha-1)}/(4f)) - z_{(1-\alpha/2)} \tag{7}$$

where $f = n_1 + n_2 - 2$, the degrees of freedom for the t test. The usual textbook sample size approximation for equal sample sizes $n = n_1 = n_2$ is

$$n' = 2 d^2 (z_{(1-\alpha/2)} + z_{(1-\beta)})^2 \tag{8}$$

which is reasonable good for large samples but underestimates the true sample size. An improved approximation is:

$$n_1 = n_2 = n = n' + z^2_{(1-\alpha/2)} /4 \tag{9}$$

which has a maximum error of .65 for $\alpha = \{.0025, .005, .025, .05\}$. Alternately $n = n'+1$ for $\alpha = .05$ and $n = n'+2$ for $\alpha = .01$ is slightly better approximation. When n has fractional values round up to the next integer. These improved approximations have not been published but a related approximation is given in van Belle and Martin [1993].

Discussion

The pooled rank test is usually more powerful than the individual tests. It will gain power over the individual t tests in most cases. However, if a some variables have large treatment effects and these outcomes are pooled with variables that have little or no treatment effect the pooled the treatment effects of with the large outcomes will be reduced in the average effect and the pooled test can lose power. The power calculations described in this paper can be used to examine the effect of adding and dropping variables on the power of the pooled rank test. As an example adding a variable with a zero effect size to the example increases the sample to 149 from 108. Dropping the least effective variable, Angina - exertion, only increases the sample size to 110 and dropping both angina variables increases the sample size to 117. Adding another variable with the same Δ and σ as SF-36 Mental Health, the best of the variables, reduces the sample size to 99.

As is true of all pre/post designs, the power increases with large pre/post correlations. If the pre/post correlations are less than .5 then the post measure alone is less variable than the pre/post difference and the pre measure can be omitted. Thus .5 is usually a lower bound for pre/post correlations. The sample size or power is moderately sensitive to the pre post correlation. In the example making all of the pre/post correlations .6, .7, or .8 results in sample sizes of 144, 108 and 72 respectively.

Low between variable correlations gain power for the pooled rank test. If the correlations were 1 between the variables any one variable would perfectly predict all of the others so there is no improvement by including other variables. When the correlation is zero each variable brings the maximum new information to the analysis. I assume that negative correlations are unlikely for variables that are expected to have the same directional shift for a treatment effect and will not consider this possibility. If in the example we assume average between variable correlations of .1, .2, .3 and .4 we get sample sizes of 59, 75, 92 and 108 respectively. These are larger changes in sample size than those from dropping or adding variables of similar effect size. Sometimes there will be several different measures of the same outcome variable in a study. These usually have moderately high correlations, often around .7 to .9 but will usually gain power over a single variable. Just take these into account when estimating the average between variable correlation.

One or Two Tails

You might question why I have used two tailed tests in the examples when we really have one tailed hypotheses for each variable. If you did question this, I think that you are both astute and correct. My personal opinion is that the uncritical use of two tailed tests is not a good idea. I went ahead and used a two tailed outcome to avoid getting sidetracked into this issue. The argument for always using a two tailed test is that you should protect yourself against an unexpected treatment effect that is in the opposite direction from the original hypothesis. This is a good argument but it results in cost increases in a study since larger sample sizes are needed to get the same power. What I usually recommend is an

asymmetric two tailed test where an $\alpha = .01$ is used in the direction of an unexpected treatment effect and $\alpha = .04$ in the expected direction. This keeps the overall $\alpha = .05$ and results in better power for the expected direction. This essentially requires stronger evidence to accept a treatment effect in the opposite direction as the expected direction. The usual choice of .025 in each tail is based on an equal ignorance of the outcome which is not true in this situation.

The sample size/power calculations can be treated as two separate tests, one at the .04 level and the other at the ,01 level. The only region where this approximation in is inaccurate is where the effect sizes are low and there is very little power. Owens [1965] discusses the accuracy of one tailed approximations to two tailed tests for the noncentral t distribution and gives some error bounds. This is not a problem for typical power calculations since powers of less than .5 show that you are in trouble and small errors are not important.

Summary

The pooled rank test is useful where there are a variety of outcome measures with directional hypothesis for the treatment effect and the treatment is expected to have similar effects on all of the variables. If there is one really good, big effect size, outcome variable and a number of marginal variables, then a pooled rank outcome may lose power compared to a t test on the best variable.

The pooled rank test has major strengths:

> It allows the investigator to use a number of outcome variables for situations where the treatment effect is expected to impact a number of different aspects of the subject. This is often the case when a broad "quality of life" outcome is of interest.
>
> The pooled test is often more powerful that either a multivariate test or individual tests and has a better chance of detecting small but consistent treatment effects over multiple outcomes.
>
> The pooled rank test is nonparametric and robust against outliers.
>
> The test is simple to implement in standard statistical packages.
>
> The test avoids the issue of multiple hypotheses tests and possible corrections adjustments such as the Bonferoni correction to the α levels which can result in serious loss of power.

In short, this is a great tool and I wish that I had thought of it. However, if you are going

to use this in a research proposal, you will need to consider the power of the test for the research design. This paper covers power calculations for post only measurements and pre-post analyses based on change scores. The power calculations can be extended to trend statistics based on multiple time point measurements. The pooled rank test can be extended to covariance adjustments of the post measure using the pre score as a covariate. I have not considered power calculations for the covariate case.

Aknowldgements

Supported in part by Grants SDR 96-002 and CSH 91-007 from Health Services Research and Development of the Department of Veterans Affaires and by Grant Number CA 34847 from the National Cancer Institute.

References

Good, Philip [1994]. Permutation Tests. Springer-Verlag New York.

Noreen, Eric W. [1989]. Computer Intensive Methods for Testing Hypotheses. John Wiley and Co. New York.

O'Brien, Peter C. [1984]. Procedures for comparing samples with multiple end points. Biometrics 40. pp. 1079-1087.

Owen, D. B. [1965]. The power of Student's t test. J. of the American Statistical Assoc. Vol 60, no. 309 pp. 320-333. corrigenda Vol. 60 p. 1251.

Fisher, L and van Belle, G [1993]. Biostatistics: A Methodology for the Health Sciences. p. 861. John Wiley and Co. New York.

van Belle G, and Martin DC [1993]: Sample size as a function of coefficient of variation and ratio of means. *American Statistician* 47:165-167.

A Power Program for The O'Brien Pooled Rank Test

Donald C. Martin Ph.D.

Northwest Veterans Affairs Health Services Research and Development Field Program

ABSTRACT

This program implements the power and sample size calculations given in "Multiple Outcomes in Health Services Research: Hypothesis Tests and Power" by Martin, Diehr, Koepsell and Fihn. It computes power or sample sizes for the O'Brien pooled rank test and the two group t test for pre-post difference designs or post only designs.

INSTALLATION

This program is written in Microsoft Visual BASIC 3 and is restricted to Microsoft operating systems by the V.-3 license agreement. It was developed under Windows for Workgroups 3.11 and has been run under Windows 95. It should work with other versions of Windows.

To install simply copy the two files from the diskette to whatever directory that you want them in and run the OPOWER.EXE file. The GRID.VBX must be available as well. At this point there are only two files.

INTRODUCTION

First go through the paper describing the method. I am too lazy to describe all of the values needed for the calculation again in this document,

The next step is to either load the program and run it or to look at figure 1 which is the basic window. Try to identify all of the data entry boxes.

This program has two modes of operation. It can compute power for given sample sizes or sample sizes need to get a selected power. The mode of operation is determined by a pair of buttons: compute power/Compute sample size. In the compute sample size mode box labeled

desired power is enabled and the Sample Size box is disabled (dim). The Sample Size box is enabled and the Desired Power box is disabled (dim) when in the compute power mode.

To compute sample size, you will need:

1. One or two tailed test? (button) {Two tails is the default.}

2. Compute Sample Size button down. { default set up. }

3. Alpha level (box) { .05 is the default value.}

4. The number of variables (box)

5. Desired power (box) { .80 is the default value }

6. The average between variable correlation. (Average Correlation box) { .40 is the default value}

7. Assumed treatment change for each variable. (Difference column)

8. Standard deviation for each variable. (Std. Dev. column)

9. Pre-post correlation (P-P Corr. column)

10. For individual t test sample sizes the power column must be defined. { .80 is the default value}

It is not really needed but it is a good idea to enter the variable names. (Variable column) The program will use default variable names, var01, var02,... etc if you don't enter the names.

To compute power:

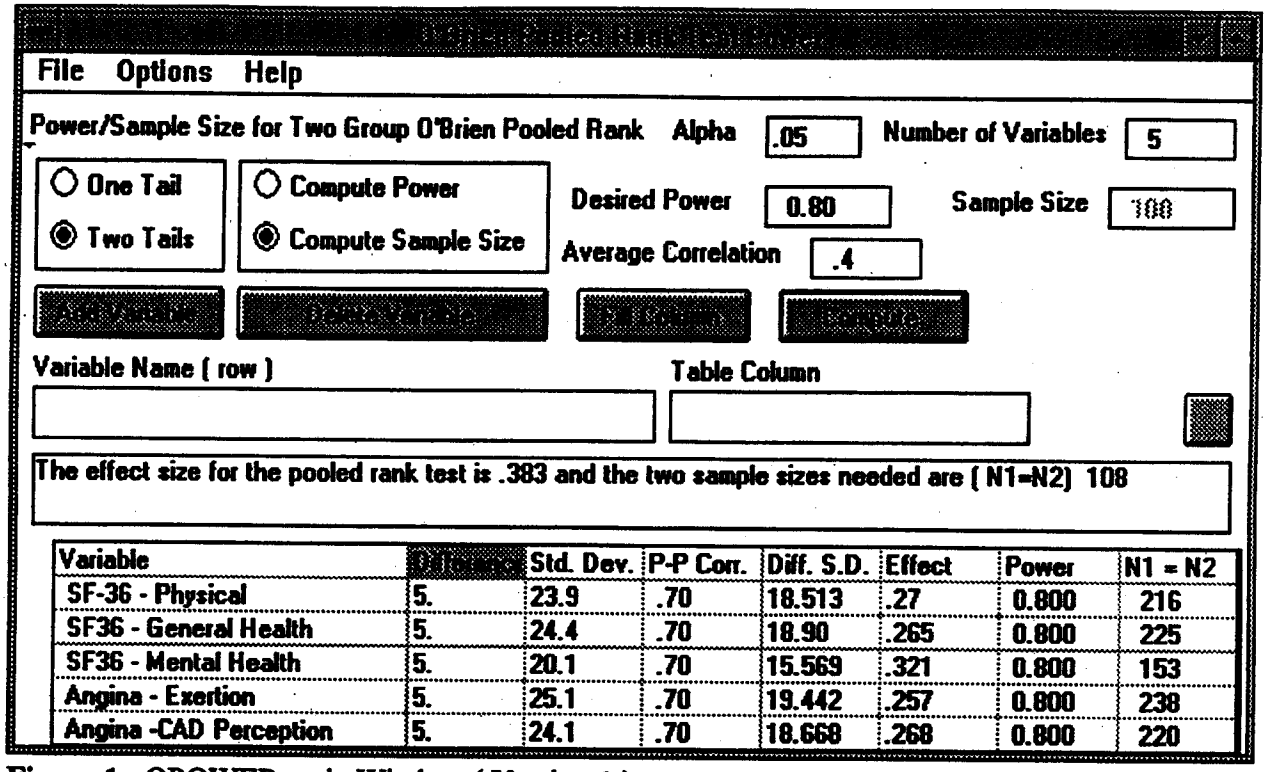
1. One or two tailed test? (button)

2. Compute Power button down.

3. Alpha level (box)

4. The number of variables (box)

5. Sample Size (box) { 100 for each of 2 groups is the default value }



Variable	Difference	Std. Dev.	P-P Corr.	Diff. S.D.	Effect	Power	N1 = N2
SF-36 - Physical	5.	23.9	.70	18.513	.27	0.800	216
SF36 - General Health	5.	24.4	.70	18.90	.265	0.800	225
SF36 - Mental Health	5.	20.1	.70	15.569	.321	0.800	153
Angina - Exertion	5.	25.1	.70	19.442	.257	0.800	238
Angina -CAD Perception	5.	24.1	.70	18.668	.268	0.800	220

Figure 1 OPOWER main Window (Version 1)

6. The average between variable correlation. (Average Correlation box)

7. Assumed treatment change for each variable. (Difference column)

8. Standard deviation for each variable. (Std. Dev. column)

9. Pre-post correlation (P-P Corr. column)

10. For t test powers on each variable, you need to fill the N1 = N2 column with sample sizes or set the entire column to the selected sample size.

It is not really needed but it is a good idea to enter the variable names. (Variable column).

The first three numeric columns of the spread sheet, Difference, Std. Dev. and P-P correlation must be filled out for each variable. You can start with an empty table and use the Add Variable button to add a line to the spread sheet or you can go to the number of variables box and set the total number of variables. The spread sheet will be updated as soon as the number of variables

box loses focus, that is done by moving your mouse to some other box or button and clicking.

To enter numeric data in the spread sheet, select a cell and enter or edit the value in the Table/Column box. To enter or change a variable name, select the name and enter or edit the name in the variable box. Be sure to look at the spread sheet values to see that your entry was correct. It is easy to make errors by failing to delete a previous numeric value. Due to time limitations, I have not programed a valid real number check and invalid numbers such as ..12 are usually converted to zeros. .

Suppose that we want to put .80 in all of the rows for power. To do this click on any power entry and enter .80 then press the fill column button which will put the selected value into all rows. This is also useful for loading the N1=N2 column for power calculations.

To delete a variable, select any cell in the row and then press the Delete Variable button. Answer OK to the are you sure that you really want to do this horrible thing message box.

Nothing happens until you press the compute button. Then the spread sheet is updated and the power or sample size is displayed in the message area. Warning! If you compute a power, the Desired Power box is updated to the pooled rank power. If you compute a sample size, the Sample Size Box is updated to the pooled rank sample size. This was put in to speed up "what if" calculations. Let me know if you find that this works against you.

The menu options are limited at this time. The File option are print and exit. Perhaps I will get around to allowing the user to save and reload the set up for a study. These would use the unavailable options Load, Save As and Save. The Option menu allows you to load the example problem given in the paper. This is useful as a time saver in trying some variations that will be described below. No help files have been implemented but there is an About option which includes a warning.

If the design is a post measure only, set the pre-post correlation, P/P column, to .5 for each variable and the calculations will be correct.

TUTORIALS

Tutorial 1 A complete set up for sample size.

Suppose that we are planning a study with 5 outcome variables with an assumed treatment effect of 5 units and standard deviations given below:

Variable	Δ	σ
SF36 -physical	5	23.9
SF36 -General Health	5	24.4
SF36 -Mental Health	5	20.1
Agina-Exertion	5	25.1
Agina-CAD Perception	5	24.1

We can start the program and use the Add Variable button to enter each variable. Note that the default variable name appears in the variable name box. If you want to replace Var01 with SF36-Physical, you will need to delete Var01. The next step is to click on the cell in the difference column and type 5. Then click on the cell in the Std. Dev. column and type 23.9. Repeat the process for the other four variables.

Assuming that we have a pre-post difference design we need to enter our guess at the pre=post correlation which is going to be .7. Select any cell in the P-P Corr. column by clicking it. Enter .7 and press the fill column button.

We now need to enter the average between variable correlation. Use .4. This number should go into the Average Correlation box. The default value is already set at .4 so you don't need to change this value. The alpha and desired power are already set at .05 and .80 which are pretty common choices. We will use the default settings of Two Tails and Compute Sample Size. We will also use the default power of .80 for the sample sizes for each of the individual variable t tests.

Press the compute button. The pooled rank effect size, .383, and the required sample size, 108 in each group, appear in the box just over the spread sheet and the sample sizes for each of the t tests appear in the N1=N2 column.

We can now print the entire screen by going to the file menu and selecting print.

Tutorial 2 Suppose that we decide on sample sizes and want power.

Suppose that we decide that a sample of 125 in each group is feasible and want to recompute the power for the pooled rank test and the individual t tests.

Press the compute power button. Now enter 125 in the sample size box for the pooled rank power. Choose any cell in the N1=N2 column and enter 125 and press fill column. Press compute. The pooled rank power is .854 and the t test powers are .566 etc.

Tutorial 3 We still like 125 in each group but we want to see the effect of dropping the angina questions from our study. Select one of the angina variables and press the Delete Variable button. Answer OK to the are you sure question. Repeat this for the second angina question. Now press compute. The pooled rank power will drop to .826.

Tutorial 4 Try a really big sample

Lets go back to our original study and raise the sample size to 250 in each group. We might want the study to have enough power to detect changes in the individual variables. Use the Add Variable button to put the two Angina variables back. Now you see why saving and reloading a set up would be useful. Of course we can cheat and reload the sample problem from the options menu which will save time. Now set all the sample sizes to 250 as before and be sure the compute power button is down. Press compute. With this gilt edge sample size we find that the pooled rank power is now .990 and the individual t test powers range from .819 to .948. A very nice study if we could afford it.

Tutorial 5 Suppose that we can only afford 100 subjects in each group.

One possibility would be to change from a two tailed test to a one tailed test. Set all of the sample sizes to 100 and press the one tailed test button, then compute. We find that we have a power of .854 for the pooled rank test but may not be able to find treatment effects for the individual variables where the powers run from .566 to .732.

Tutorial 5 Lets be brave and consider an asymmetric two tailed test.

We will use an alpha of .04 in the expected direction and of .01 in the unexpected direction with the above sample sizes of 100 in each group. This may avoid some of the criticism from the "we never use one tailed tests". Set up for power with sample sizes of 100 as before, choose a one tailed test and set alpha to ,04. We see that the power of the pooled rank test drops from .854 to .828. A minor loss compared to the two tailed power of .769 for an .05 level test.

Tutorial 6 A post treatment measure only study sample size calculation.

Suppose we go back to the study in Tutorial 1. The only change is to set the P-P Corr. column to .5. You can change the powers back to .8 from exercise 5, change the alpha back to .05 and change back to compute sample size. Then we find that the sample size needed for the pooled rank test is 179 and the individual t test sample sizes range from 255 to 397 for each group.

Tutorial 7 Lets go back to the pre-post design and see what a higher pre-post correlation does.

Change all of the pre-post correlations to .8 by selecting a cell and using the fill column button. The pooled rank sample size drops to 72 and the t tests range from 102 to 159. Using reliable measures can really help the sample size.

Tutorial 8 Lets add a really great variable to the original design and compute sample sizes.

Reset the P-P corr. column to .7 or use the sample data option. Lets add a new variable with a treatment difference of 10, a standard deviation of 25 and a pre-post correlation of .7. Compute. We see that the t test sample size for our great additional variable is 60 while the pooled rank test is 79 in each group. In this case we would be better off just to test the one really good variable instead of pooling it with a bunch of so-so variables that drag its power down.

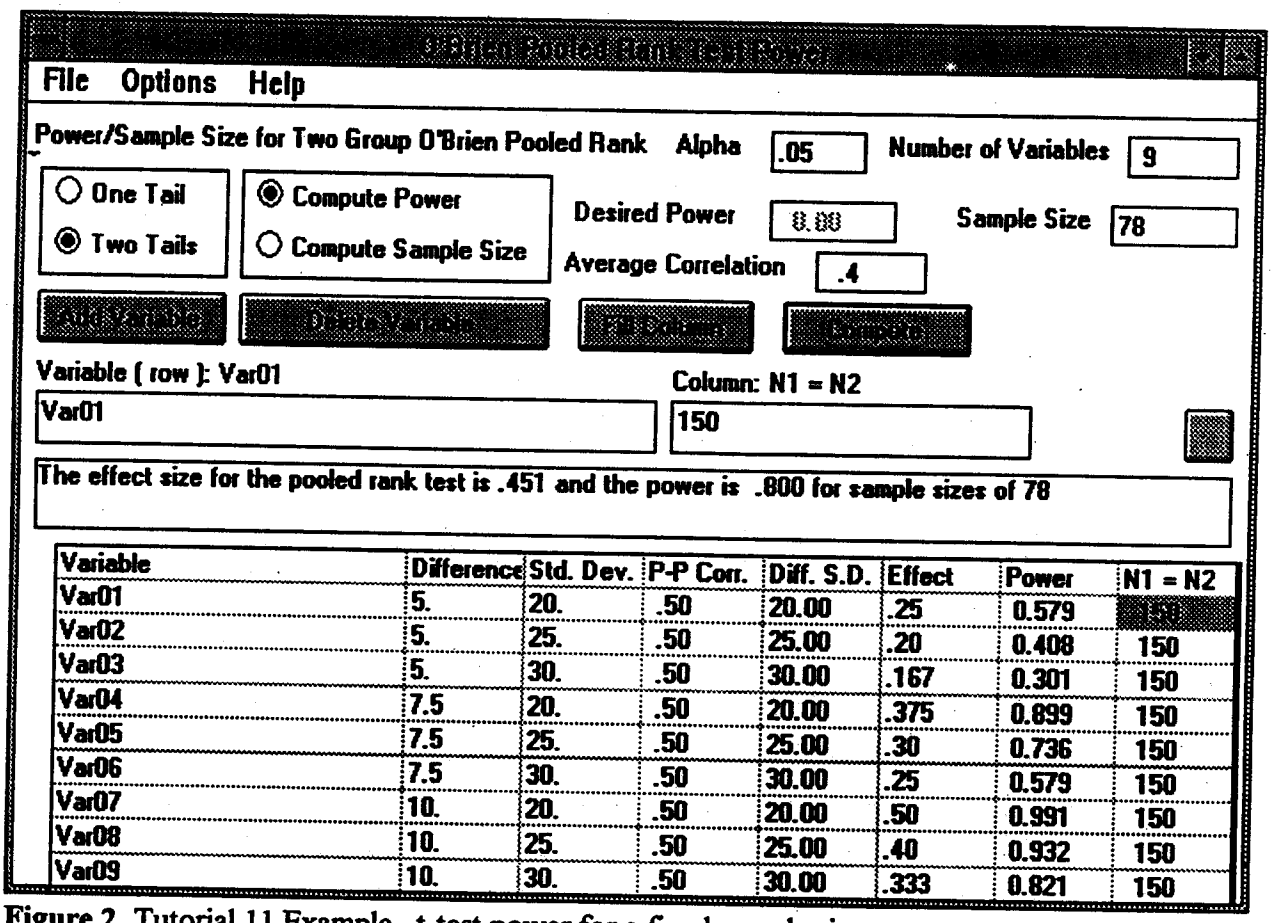
Tutorial 9 Lets see what adding a really bad, no treatment effect, variable does.

Change the difference on Var07 in example 9 to zero and compute, That ran the sample size up to 149 from 108 from Tutorial 1.
Set up the calculation for only one variable or you can enter the same variable multiple times with different Std. Dev. of Differences and get the results from the Power or N1=N2 columns. If the design is not a pre-post difference set the P-P Corr. column to .5. Figure 2 shows a sample size calculation for three assumed Differences, 5, 7.5 and 10 and three assumed variances 20, 25 and 30.

Tutorial 11 Compute the power of a number of t tests

The set up is essentially the same as the problem in Tutorial 10 except that we fill the N1=N2 column with the proposed sample size and the results are in the Power column.

Tutorial 10 I just want the sample size for a plain old two group t test. No pooled ranks, etc.



Variable	Difference	Std. Dev.	P-P Corr.	Diff. S.D.	Effect	Power	N1 = N2
Var01	5.	20.	.50	20.00	.25	0.579	150
Var02	5.	25.	.50	25.00	.20	0.408	150
Var03	5.	30.	.50	30.00	.167	0.301	150
Var04	7.5	20.	.50	20.00	.375	0.899	150
Var05	7.5	25.	.50	25.00	.30	0.736	150
Var06	7.5	30.	.50	30.00	.25	0.579	150
Var07	10.	20.	.50	20.00	.50	0.991	150
Var08	10.	25.	.50	25.00	.40	0.932	150
Var09	10.	30.	.50	30.00	.333	0.821	150

Figure 2 Tutorial 11 Example. t-test power for a fixed sample size.

Set up the calculation for only one variable or you can enter the same variable multiple times with different Std. Dev. of Differences and get the results from the Power or N1=N2 columns. If the design is not a pre-post difference set the P-P Corr. column to .5. Figure 2 shows a sample size calculation for three assumed Differences, 5, 7.5 and 10 and three assumed variances 20, 25 and 30.

Tutorial 11 Compute the power of a number of t tests

The set up is essentially the same as the problem in Tutorial 10 except that we fill the N1=N2 column with the proposed sample size and the results are in the Power column. The set up and results are shown in Figure 3.

File Options Help

Power/Sample Size for Two Group O'Brien Pooled Rank Alpha .05 Number of Variables 9

One Tail / Two Tails | Compute Power / Compute Sample Size | Desired Power 0.80 | Sample Size 78 | Average Correlation .4

Variable (row): Var01 Column: N1 = N2

Var01 150

The effect size for the pooled rank test is .451 and the power is .800 for sample sizes of 78

Variable	Difference	Std. Dev.	P-P Corr.	Diff. S.D.	Effect	Power	N1 = N2
Var01	5.	20.	.50	20.00	.25	0.579	150
Var02	5.	25.	.50	25.00	.20	0.408	150
Var03	5.	30.	.50	30.00	.167	0.301	150
Var04	7.5	20.	.50	20.00	.375	0.899	150
Var05	7.5	25.	.50	25.00	.30	0.736	150
Var06	7.5	30.	.50	30.00	.25	0.579	150
Var07	10.	20.	.50	20.00	.50	0.991	150
Var08	10.	25.	.50	25.00	.40	0.932	150
Var09	10.	30.	.50	30.00	.333	0.821	150

Figure 3 Tutorial 11, t-test power for a sample size of 150 in each group.