Power And Precision™

Michael Borenstein
Hillside Hospital, Albert Einstein College of Medicine,
and Biostatistical Programming Associates

Hannah Rothstein
Baruch College, City University of New York

Jacob Cohen
New York University

For information please visit our our web site at *www.PowerAnalysis.com* or contact

Biostat, Inc.
14 North Dean Street
Englewood, NJ 07631
Tel: (201) 541-5688
Toll free in USA 1-877-Biostat
Fax: (201) 541-5526

ISBN 0-9709662-0-2

Preface

Power And Precision™ is a stand-alone software program that can be used by itself or as a tool to enhance any other statistical package, such as SPSS or SYSTAT. Power And Precision helps find an appropriate balance among effect size, sample size, the criterion required for significance (alpha), and power. Typically, when a study is being planned, either the effect size is known from previous research or an effect size of practical significance is specified. In addition, the user enters the desired alpha and power. The analysis then indicates the number of cases needed to attain the desired statistical power.

This program can also be used to obtain the number of cases needed for desired precision. In general, precision is a function of the confidence level required, the sample size, and the variance of the effect size.

In studies of either power or precision, the program produces graphs and tables, as well as written reports that are compatible with many word processing programs.

Compatibility

Power And Precision is designed to operate on computer systems running Windows 2000, Windows XP, Windows Vista, or Windows 7.

Technical Support

For technical support please contact us by e-mail at *Support@PowerAnalysis.com* or by phone at (201) 541-5688.

Tell Us Your Thoughts

Please contact us with any comments or suggestions.

E-mail *MichaelB@PowerAnalysis.com*, phone (201) 541-5688, Fax (201) 541 or phone (201) 541-5688, Fax (201) 541-5688 or mail to Michael Borenstein, Biostat, Inc., 14 North Dean Street, Englewood, NJ 07631.

Acknowledgments

We gratefully acknowledge the support of the National Institute of Mental Health (NIMH), which made the development of this software possible through the Small Business Innovation Research Program. In particular, we thank the following people at NIMH for their enthusiasm, help, and encouragement: Gloria Levin, scientific review administrator; Michael Huerta, associate director, Division of Neuroscience and Behavioral Research; Mary Curvey, grants program specialist, Division of Neuroscience and Behavioral Research; James Moynihan, former associate director, Kenneth Lutterman, former health scientist administrator, and Harold Goldstein, former health scientist administrator.

Don Rubin, Larry Hedges, John Barnard, Kim Meier, David Schoenfeld, Ed Lakatos, and Jesse Berlin helped in the development of specific algorithms.

The algorithm for non-central *F* was developed by Barry W. Brown, James Lovato, and Kathy Russell of the Department of Biomathematics, University of Texas, M.D. Anderson Cancer Center, Houston, Texas (through a grant from the National Cancer Institute) and is used with their kind permission. Some of the program's procedures for exact tests of proportions were adapted from code originally written by Gerard E. Dallal and used with his kind permission. The program also employs algorithms under license from IMSL, Inc.

John Kane generously allowed the senior author to take a partial sabbatical from his work at Hillside Hospital to work on this program.

Finally, thanks to the people at SPSS who worked with us on this and previous versions of the product: Louise Rehling, Mark Battaglia, Jing Shyr, Nancy Dobrozdravic, Alice Culbert, Yvonne Smith, JoAnn Ziebarth, Mary Campbell, and Bonnie Shapiro. Special thanks to Tom Chiu and Hiroaki Minato, whose excellent counsel substantially enhanced the program, and to Tony Babinec and Kyle Weeks, who worked closely with us over several years to help plan the program and oversee its development.

Michael Borenstein
Hillside Hospital,
Albert Einstein College of Medicine,
and Biostatistical Programming Associates

Hannah Rothstein
Baruch College, City University of New York

Jacob Cohen
New York University

Contents

22 Equivalence Tests (Means) 207

23 Equivalence Tests (Proportions) 215

Appendix A

Appendix B

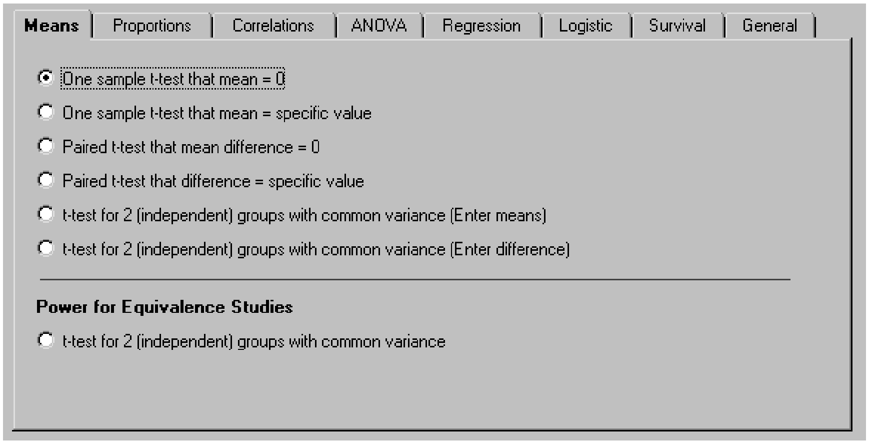
Appendix C

Appendix E

1 The Sixty-Second Tour

Selecting a Procedure

To display the available procedures, select *New analysis* from the File menu.



● One sample t-test that mean = 0
○ One sample t-test that mean = specific value
○ Paired t-test that mean difference = 0
○ Paired t-test that difference = specific value
○ t-test for 2 (independent) groups with common variance (Enter means)
○ t-test for 2 (independent) groups with common variance (Enter difference)

Power for Equivalence Studies

○ t-test for 2 (independent) groups with common variance

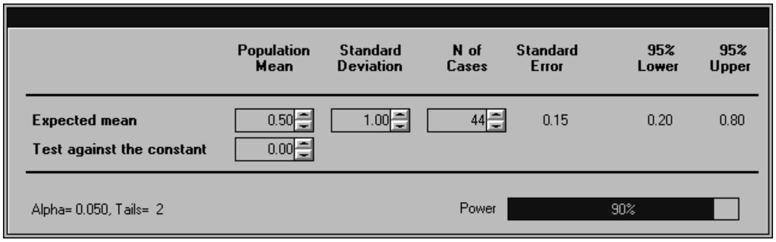
Procedures are grouped into the following categories:

- Means (t-tests and z-tests for one group and for two groups)
- Proportions (tests of single proportions, of two proportions for independent or matched groups, and for a $K \times C$ crosstabulation
- Correlations (one- and two-group correlations)
- ANOVA (analysis of variance and covariance—oneway or factorial)
- Multiple regression (for any number of sets, for increment or cumulative R^2)
- General case (allows user to specify the non-centrality parameter directly)

To see an example, click *Example*. To activate a procedure, click *OK.*

Navigating within the Program

- Enter data on main panel.



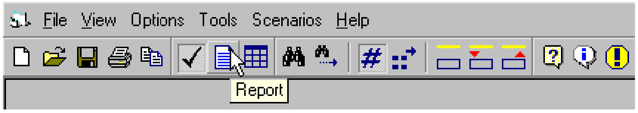
	Population Mean	Standard Deviation	N of Cases	Standard Error	95% Lower	95% Upper
Expected mean	0.50	1.00	44	0.15	0.20	0.80
Test against the constant	0.00					

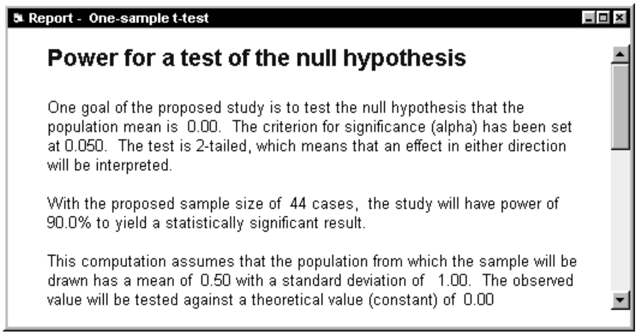
Alpha= 0.050, Tails= 2

Power 90%

Enter basic data for effect size, alpha, confidence level, and sample size. The program will display power and precision and will update them continuously as study parameters are set and modified.

- Click the *Report* icon to generate a report incorporating the data from the main screen.





Report - One-sample t-test

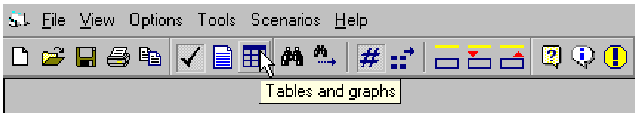
Power for a test of the null hypothesis

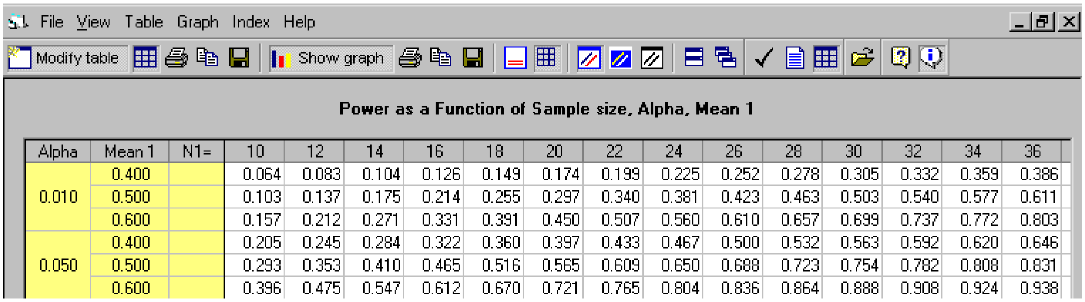
One goal of the proposed study is to test the null hypothesis that the population mean is 0.00. The criterion for significance (alpha) has been set at 0.050. The test is 2-tailed, which means that an effect in either direction will be interpreted.

With the proposed sample size of 44 cases, the study will have power of 90.0% to yield a statistically significant result.

This computation assumes that the population from which the sample will be drawn has a mean of 0.50 with a standard deviation of 1.00. The observed value will be tested against a theoretical value (constant) of 0.00

- Click the *Tables and graphs* icon to generate a table showing how power or precision will vary as a function of sample size. The table is based on the data from the main screen.

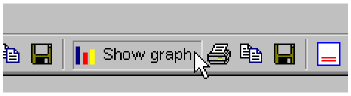


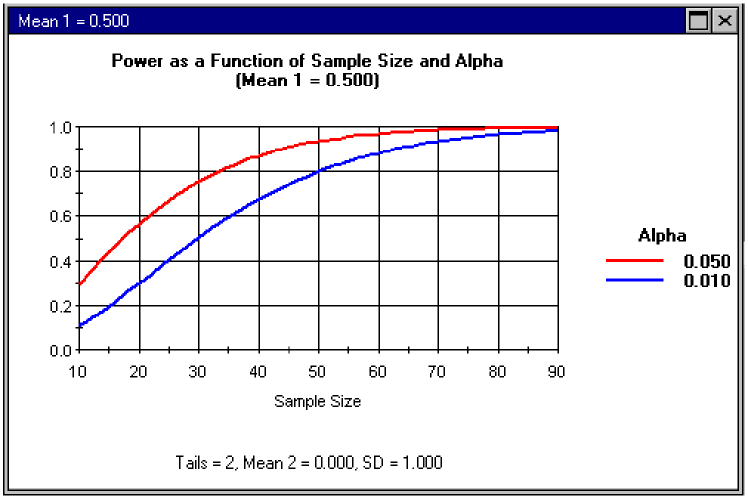


Power as a Function of Sample size, Alpha, Mean 1

Alpha	Mean 1	N1=	10	12	14	16	18	20	22	24	26	28	30	32	34	36
0.010	0.400		0.064	0.083	0.104	0.126	0.149	0.174	0.199	0.225	0.252	0.278	0.305	0.332	0.359	0.386
	0.500		0.103	0.137	0.175	0.214	0.255	0.297	0.340	0.381	0.423	0.463	0.503	0.540	0.577	0.611
	0.600		0.157	0.212	0.271	0.331	0.391	0.450	0.507	0.560	0.610	0.657	0.699	0.737	0.772	0.803
0.050	0.400		0.205	0.245	0.284	0.322	0.360	0.397	0.433	0.467	0.500	0.532	0.563	0.592	0.620	0.646
	0.500		0.293	0.353	0.410	0.465	0.516	0.565	0.609	0.650	0.688	0.723	0.754	0.782	0.808	0.831
	0.600		0.396	0.475	0.547	0.612	0.670	0.721	0.765	0.804	0.836	0.864	0.888	0.908	0.924	0.938

- ❖ Click the *Display Graph* icon to display a graph corresponding to the table.





- ❖ The program includes various tools that assist the user to set an effect size or to find the sample size required for power. Use these tools to find an appropriate balance between alpha and beta and/or to ensure adequate precision.

2 Overview of Power and Precision

Power Analysis

Traditionally, data collected in a research study is submitted to a significance test to assess the viability of the null hypothesis. The p-value, provided by the significance test and used to reject the null hypothesis, is a function of three factors: size of the observed effect, sample size, and the criterion required for significance (alpha).

A power analysis, executed when the study is being planned, is used to anticipate the likelihood that the study will yield a significant effect and is based on the same factors as the significance test itself. Specifically, the larger the effect size used in the power analysis, the larger the sample size; the larger (more liberal) the criterion required for significance (alpha), the higher the expectation that the study will yield a statistically significant effect.

These three factors, together with power, form a closed system—once any three are established, the fourth is completely determined. The goal of a power analysis is to find an appropriate balance among these factors by taking into account the substantive goals of the study, and the resources available to the researcher.

Effect Size

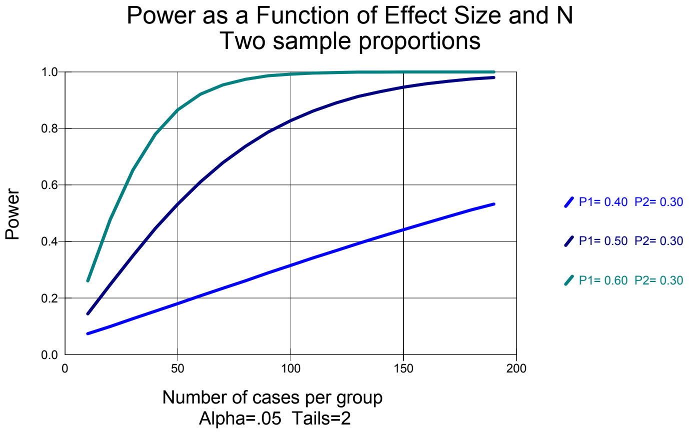
The term *effect size* refers to the magnitude of the effect under the alternate hypothesis. The nature of the effect size will vary from one statistical procedure to the next (it could be the difference in cure rates, or a standardized mean difference, or a correlation coefficient), but its function in power analysis is the same in all procedures.

The effect size should represent the smallest effect that would be of clinical or substantive significance, and for this reason, it will vary from one study to the next. In clinical trials, for example, the selection of an effect size might take into account the severity of the illness being treated (a treatment effect that reduces mortality by 1% might be clinically important, while a treatment effect that reduces transient asthma by 20% may be of little interest). It might take into account the existence of alternate treatments. (If alternate treatments exist, a new treatment would need to surpass these other treatments to be important.) It might also take into account the treatment's cost and side effects. (A treatment that carried these burdens would be adopted only if the treatment effect was very substantial.)

Power analysis gives power for a specific effect size. For example, the researcher might report that if the treatment increases the recovery rate by 15 percentage points, the study will have power of 80% to yield a significant effect. For the same sample size and alpha, if the treatment effect is less than 15 percentage points, then the power will be less than 80%. If the true effect size exceeds 15 percentage points, then power will exceed 80%.

While one might be tempted to set the "clinically significant effect" at a small value to ensure high power for even a small effect, this determination cannot be made in isolation. The selection of an effect size reflects the need for balance between the size of the effect that we can detect and resources available for the study.

Figure 2.1 Power as a function of effect size and N



Small effects will require a larger investment of resources than large effects. Figure 2.1 shows power as a function of sample size for three levels of effect size (assuming that alpha, two-tailed, is set at 0.05). For the smallest effect (30% versus 40%), we would need a sample of 356 per group to yield power of 80% (not shown on the graph). For the intermediate effect (30% versus 50%), we would need a sample of 93 per group to yield this level of power. For the largest effect size (30% versus 60%), we would need a sample of 42 per group to yield power of 80%. We may decide that for our purposes, it would make sense to enroll 93 per group to detect the intermediate effect but inappropriate to enroll 356 patients per group to detect the smallest effect.

The *true* (population) effect size is not known. While the effect size used for the power analysis is assumed to reflect the population effect size, the power analysis is more appropriately expressed as, "*If* the true effect is this large, power would be …," rather than, "The true effect is this large, and therefore power is …."

This distinction is an important one. Researchers sometimes assume that a power analysis cannot be performed in the absence of pilot data. In fact, it is usually possible to perform a power analysis based entirely on a logical assessment of what constitutes a clinically (or theoretically) important effect. Indeed, while the effect observed in prior studies might help to provide an *estimate* of the true effect, it is not likely to be the *true* effect in the population—if we knew that the effect size in these studies was accurate, there would be no need to run the new study.

Since the effect size used in power analysis is not the *true* population value, the researcher may decide to present a range of power estimates. For example (assuming that $N = 93$ per group and alpha $= 0.05$, two-tailed), the researcher may state that the study will have power of 80% to detect a treatment effect of 20 points (30% versus 50%) and power of 99% to detect a treatment effect of 30 points (30% versus 60%).

Cohen has suggested conventional values for small, medium, and large effects in the social sciences. The researcher may want to use these values as a kind of reality check to ensure that the values that he or she has specified make sense relative to these anchors. The program also allows the user to work directly with one of the conventional values rather than specifying an effect size, but it is preferable to specify an effect based on the criteria outlined above, rather than relying on conventions.

Alpha

The significance test yields a computed p-value that gives the likelihood of the study effect, given that the null hypothesis is true. For example, a p-value of 0.02 means that, assuming that the treatment has a null effect, and given the sample size, an effect as large as the observed effect would be seen in only 2% of studies.

The p-value obtained in the study is evaluated against the criterion, alpha. If alpha is set at 0.05, then a p-value of 0.05 or less is required to reject the null hypothesis and establish statistical significance.

If a treatment really is effective and the study succeeds in rejecting the nil hypothesis, or if a treatment really has no effect and the study fails to reject the nil hypothesis, the study's result is correct. A type 1 error is said to occur if there is a nil effect but we mistakenly reject the null. A type 2 error is said to occur if the treatment is effective but we fail to reject the nil hypothesis.

Note: The *null hypothesis* is the hypothesis to be nullified. When the null hypothesis posits a nil effect (for example, a mean difference of 0), the term *nil hypothesis* is used.

Assuming that the null hypothesis is true and alpha is set at 0.05, we would expect a type I error to occur in 5% of all studies—the type I error rate is equal to alpha. Assuming that the null hypothesis is false (and the true effect is given by the effect size used in computing power), we would expect a type 2 error to occur in the proportion of studies denoted by one minus power, and this error rate is known as beta.

If our only concern in study design were to prevent a type 1 error, it would make sense to set alpha as conservatively as possible (for example, at 0.001). However, alpha does not operate in isolation. For a given effect size and sample size, as alpha is decreased, power is also decreased. By moving alpha from, say, 0.10 toward 0.01, we reduce the likelihood of a type 1 error but increase the likelihood of a type 2 error.

Figure 2.2 Power as a function of alpha and N

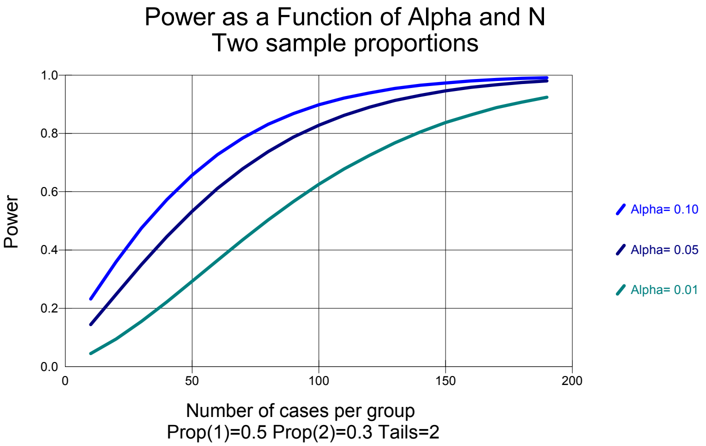


Figure 2.2 shows power as a function of sample size for three levels of alpha (assuming an effect size of 30% versus 50%, which is the intermediate effect size in the previous figure). For the most stringent alpha (0.01), an N of 139 per group is required for power of 0.80. For alpha of 0.05, an N of 93 per group is required. For alpha of 0.10, an N of 74 per group is required.

Traditionally, researchers in some fields have accepted the notion that alpha should be set at 0.05 and power at 80% (corresponding to a type 2 error rate and beta of 0.20). This notion implies that a type 1 error is four times as harmful as a type 2 error (the ratio of alpha to beta is 0.05 to 0.20), which provides a *general* standard in a specific application. However, the researcher must strike a balance between alpha and beta appropriate to the specific issues. For example, if the study will be used to screen a new drug for further testing, we might want to set alpha at 0.20 and power at 95% to ensure that a potentially useful drug is not overlooked. On the other hand, if we were working with a drug that carried the risk of side effects and the study goal was to obtain FDA approval for use, we might want to set alpha at 0.01, while keeping power at 95%.

Tails

The significance test is always defined as either one-tailed or two-tailed. A two-tailed test is a test that will be interpreted if the effect meets the criterion for significance and falls in either direction. A two-tailed test is appropriate for the vast majority of research studies. A one-tailed test is a test that will be interpreted only if the effect meets the criterion for significance and falls in the observed direction (that is, the treatment *improves* the cure rate) and is appropriate only for a specific type of research question.

Cohen gives the following example of a one-tailed test. An assembly line is currently using a particular process (A). We are planning to evaluate an alternate process (B), which would be expensive to implement but could yield substantial savings if it works as expected. The test has three possible outcomes: process A is better, there is no difference between the two, or process B is better. However, for our purposes, outcomes 1 and 2 are functionally equivalent, since either would lead us to maintain the status quo. In other words, we have no need to distinguish between outcomes 1 and 2.

A one-tailed test should be used only in a study in which, as in this example, an effect in the unexpected direction is functionally equivalent to no effect. It is *not* appropriate to use a one-tailed test simply because one is able to specify the expected direction of the effect prior to running the study. In medicine, for example, we typically expect that the new procedure will *improve* the cure rate, but a finding that it decreases the cure rate would still be important, since it would demonstrate a possible flaw in the underlying theory.

For a given effect size, sample size, and alpha, a one-tailed test is more powerful than a two-tailed test (a one-tailed test with alpha set at 0.05 has approximately the same power as a two-tailed test with alpha set at 0.10). However, the number of tails should be set based on the substantive issue of whether an effect in the reverse direction will be meaningful. In general, it would not be appropriate to run a test as one-tailed rather than two-tailed as a means of increasing power. (Power is higher for the one-tailed test only under the assumption that the observed effect falls in the expected direction. When the test is one-tailed, power for an effect in the reverse direction is nil).

Sample Size

For any given effect size and alpha, increasing the sample size will increase the power (ignoring for the moment the special case where power for a test of proportions is computed using exact methods). As is true of effect size and alpha, sample size cannot be viewed in isolation but rather as one element in a complex balancing act. In some studies, it might be important to detect even a small effect while maintaining high power. In such a case, it might be appropriate to enroll many thousands of patients (as was done in the physicians' study that found a relationship between aspirin use and cardiovascular events).

Typically, though, the number of available cases is limited. The researcher might need to find the largest N that can be enrolled and work backward from there to find an appropriate balance between alpha and beta. She may need to forgo the possibility of finding a small effect and acknowledge that power will be adequate for a large effect only.

Note. For studies that involve two groups, power is generally maximized when the total number of subjects is divided equally between two groups. When the number of cases in the two groups is not equal, the "effective N" for computing power falls closer to the smaller sample size than the larger one.

Power

Power is the fourth element in this closed system. Given an effect size, alpha, and sample size, power is determined. As a general standard, power should be set at 80%. However, for any given research, the appropriate level of power should be decided on a case-by-case basis, taking into account the potential harm of a type 1 error, the determination of a clinically important effect, and the potential sample size, as well as the importance of identifying an effect, should one exist.

Ethical Issues

Some studies involve putting patients at risk. At one extreme, the risk might be limited to loss of time spent completing a questionnaire. At the other extreme, the risk might involve the use of an ineffective treatment for a potentially fatal disease. These issues are clearly beyond the scope of this discussion, but one point should be made here.

Ethical issues play a role in power analysis. If a study to test a new drug will have adequate power with a sample of 100 patients, then it would be inappropriate to use a sample of 200 patients, since the second 100 are being put at risk unnecessarily. At the same time, if the study requires 200 patients in order to yield adequate power, it would be inappropriate to use only 100. These 100 patients may consent to take part in the study on the assumption that the study will yield useful results. If the study is underpowered, then the 100 patients have been put at risk for no reason.

Of course, the actual decision-making process is complex. One can argue about whether adequate power for the study is 80%, 90%, or 99%. One can argue about whether power should be set based on an improvement of 10 points, 20 points, or 30 points. One can argue about the appropriate balance between alpha and beta. In addition, the sample size should take account of precision as well as power (see "Precision," below). The point here is that these kinds of issues need to be addressed explicitly as part of the decision-making process.

The Null Hypothesis versus the Nil Hypothesis

Power analysis focuses on the study's potential for rejecting the null hypothesis. In most cases, the null hypothesis is the null hypothesis of no effect (also known as the nil hypothesis). For example, the researcher is testing a null hypothesis that the change in score from time 1 to time 2 is 0. In some studies, however, the researcher might attempt to disprove a null hypothesis other than the nil. For example, if the researcher claims that the intervention boosts the scores by 20 points or more, the impact of this claim is to change the effect size.

Additional Reading

The bibliography includes a number of references that offer a more comprehensive treatment of power. In particular, see Borenstein (1994B, 1997), Cohen (1965, 1988, 1992, 1994), Feinstein (1975), and Kraemer (1987).

Precision

The discussion to this point has focused on power analysis, which is an appropriate precursor to a test of significance. If the researcher is designing a study to test the null hypothesis, then the study design should ensure, to a high degree of certainty, that the study will be able to provide an adequate (that is, powerful) test of the null hypothesis.

The study may be designed with another goal as well. In addition to (or instead of) testing the null hypothesis, the researcher might use the study to estimate the magnitude of the effect—to report, for example, that the treatment increases the cure rate by 10 points, 20 points, or 30 points. In this case, study planning would focus not on the study's ability to reject the null hypothesis but rather on the precision with which it will allow us to estimate the magnitude of the effect.

Assume, for example, that we are planning to compare the response rates for treatments and anticipate that these rates will differ from each other by about 20 percentage points. We would like to be able to report the rate difference with a precision of plus or minus 10 percentage points.

The precision with which we will be able to report the rate difference is a function of the confidence level required, the sample size, and the variance of the outcome index. Except in the indirect manner discussed below, it is not affected by the effect size.

Sample Size

The confidence interval represents the precision with which we are able to report the effect size, and the larger the sample, the more precise the estimate. As a practical matter, sample size is the dominant factor in determining the precision.

Figure 2.3 Precision for a rate difference (95% confidence interval), effect size 20 points

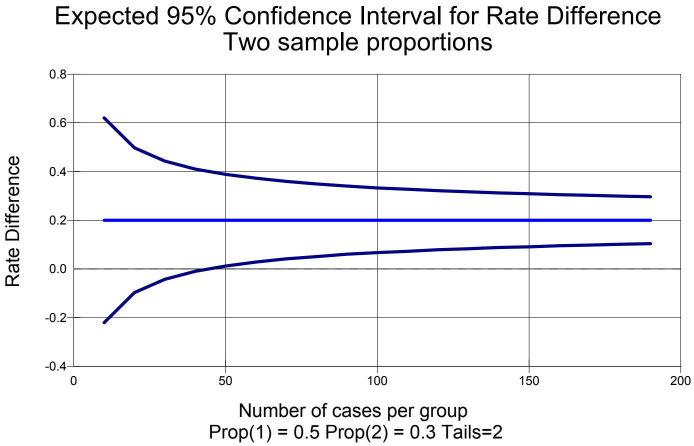


Figure 2.3 shows precision for a rate difference as a function of sample size. This figure is based on the same rates used in the power analysis (30% versus 50%). With N = 50 per group, the effect would be reported as 20 points, with a 95% confidence interval of plus or minus 19 points (01 to 39 points). With N= 100 per group, the effect would be reported as 20 points, with a 95% confidence interval of plus or minus 13 points (7 to 33). With N = 200 per group, the effect would be reported as 20 points, with a 95% confidence interval of plus or minus 9 points (11 to 29).

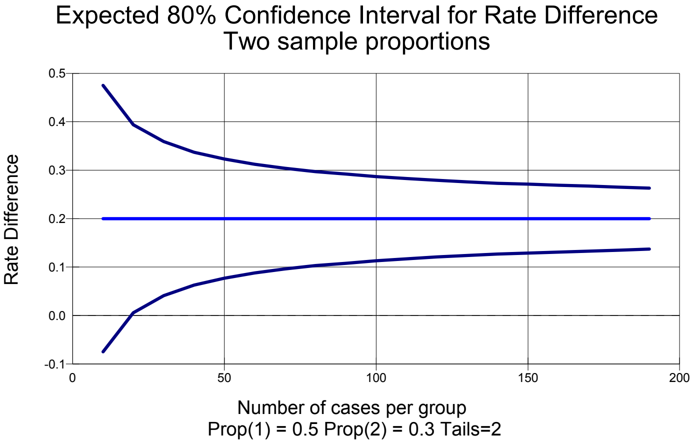
Note. For studies that involve two groups, precision is maximized when the subjects are divided equally between the two groups (this statement applies to the procedures included in this program). When the number of cases in the two groups is uneven, the "effective N" for computing precision falls closer to the smaller sample size than the larger one.

Confidence Level

The confidence level is an index of certainty. For example, with N= 93 per group, we might report that the treatment improves the response rate by 20 percentage points, with a 95% confidence interval of plus or minus 13 points (7 to 33). This means that in 95%

of all possible studies, the confidence interval computed in this manner will include the true effect. The confidence level is typically set in the range of 99% to 80%.

Figure 2.4 Precision for a rate difference (80% confidence interval)



The 95% confidence interval will be wider than the 90% interval, which in turn will be wider than the 80% interval. For example, compare Figure 2.4, which shows the expected value of the 80% confidence interval, with Figure 2.3, which is based on the 95% confidence interval. With a sample of 100 cases per group, the 80% confidence interval is plus or minus 9 points (11 to 29), while the 95% confidence interval is plus or minus 13 points (7 to 34).

The researcher may decide to report the confidence interval for more than one level of confidence. For example, he may report that the treatment improves the cure rate by 10 points (80% confidence interval 11 to 29, and 95% confidence interval 7 to 34). It has also been suggested that the researcher use a graph to report the full continuum of confidence intervals as a function of confidence levels. (See Poole, 1987a,b,c; Walker, 1986a,b.)

Tails

The researcher may decide to compute two-tailed or one-tailed bounds for the confidence interval. A two-tailed confidence interval extends from some finite value below the observed effect to another finite value above the observed effect. A one-tailed confidence "interval" extends from minus infinity to some value above the observed effect, or from some value below the observed effect to plus infinity (the logic of the procedure may impose a limit other than infinity, such as 0 and 1, for proportions). A one-tailed confidence interval might be used if we were concerned with effects in only one direction. For example, we might report that a drug increases the remission rate by 20 points, with a 95% lower limit of 15 points (the upper limit is of no interest).

For any given sample size, dispersion, and confidence level, a one-tailed confidence interval is narrower than a two-tailed interval in the sense that the distance from the observed effect to the computed boundary is smaller for the one-tailed interval (the one-tailed case is not really an interval, since it has only one boundary). As was the case with power analysis, however, the decision to work with a one-tailed procedure rather than a two-tailed procedure should be made on substantive grounds rather than as a means for yielding a more precise estimate of the effect size.

Variance of the Effect Size

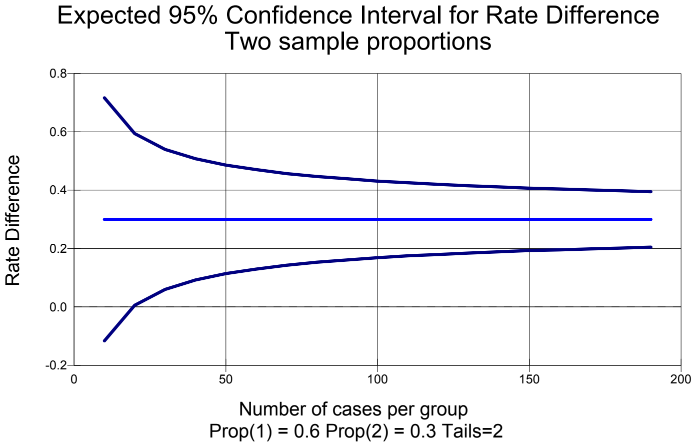
The third element determining precision is the dispersion of the effect size index. For t-tests, dispersion is indexed by the standard deviation of the group means. If we will be reporting precision using the metric of the original scores, then precision will vary as a function of the standard deviation. (If we will be reporting precision using a standard index, then the standard deviation is assumed to be 1.0, thus the standard deviation of the original metric is irrelevant.) For tests of proportions, the variance of the index is a function of the proportions. Variance is highest for proportions near 0.50 and lower for proportions near 0.0 or 1.0. As a practical matter, variance is fairly stable until proportions fall below 0.10 or above 0.90. For tests of correlations, the variance of the index is a function of the correlation. Variance is highest when the correlation is 0.

Effect Size

Effect size, which is a primary factor in computation of power, has little, if any, impact in determining precision. In the example, we would report a 20-point effect, with a 95% confidence interval of plus or minus 13 points. A 30-point effect would similarly be reported, with a 95% confidence interval of plus or minus 13 points.

Compare Figure 2.5, which is based on an effect size of 30 points, with Figure 2.3, which is based on an affect size of 20 points. The width of the interval is virtually identical in the two figures; in Figure 2.5, the interval is simply shifted north by 10 percentage points.

Figure 2.5 Precision for a rate difference (95% confidence interval), effect size 30 points

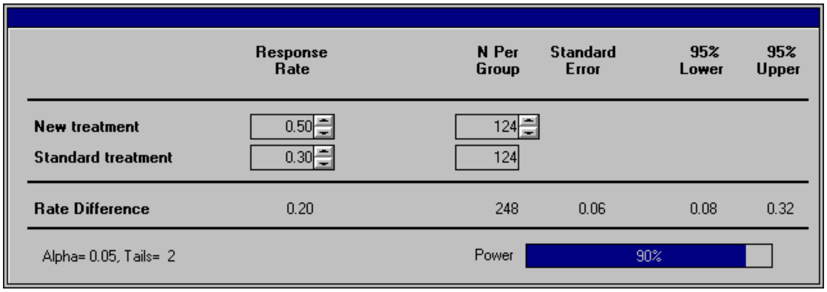


While effect size plays no direct role in precision, it may be related to precision indirectly. Specifically, for procedures that work with mean differences, the effect size is a function of the mean difference and also the standard deviation within groups. The former has no impact on precision; the latter affects both effect size and precision (a smaller standard deviation yields higher power and better precision in the raw metric). For procedures that work with proportions or correlations, the absolute value of the proportion or correlation affects the index's variance, which in turn may have an impact on precision.

Planning for Precision

The process of planning for precision has some obvious parallels to planning for power, but the two processes are not identical and, in most cases, will lead to very different estimates for sample size. The program displays the expected value of the precision for a given sample size and confidence level. In Screen 1, the user has entered data for effect size and found that a sample of 124 per group will yield power of 90% and precision (95% confidence interval) of plus or minus 12 points.

Figure 2.6 Proportions, 2 X 2 independent samples



	Response Rate	N Per Group	Standard Error	95% Lower	95% Upper
New treatment	0.50	124			
Standard treatment	0.30	124			
Rate Difference	0.20	248	0.06	0.08	0.32

Alpha= 0.05, Tails= 2

Power 90%

Typically, the user will enter data for effect size and sample size. The program immediately displays both power and precision for the given values. Changes to the effect size will affect power (and may have an incidental effect on precision). Changes to sample size will affect both power and precision. Changes to alpha will affect power, while changes to the confidence level will affect precision. Defining the test as one-tailed or two-tailed will affect both power and precision.

Tolerance Intervals

The confidence interval width displayed for t-tests is the median interval width. (Assuming the population standard deviation is correct, the confidence interval will be narrower than the displayed value in half of the samples and wider in half of the samples). The width displayed for exact tests of exact proportions is the expected value (that is, the mean width expected over an infinite number of samples). For other procedures where the program displays a confidence interval, the width shown is an approximate value. (It is the value that would be computed if the sample proportions in the sample correlation precisely matched the population values).

For many applications, especially when the sample size is large, these values will prove accurate enough for planning purposes. Note, however, that for any single study, the precision will vary somewhat from the displayed value. For t-tests, on the assumption that the population standard deviation is 10, the sample standard deviation will typically be smaller or greater than 10, yielding a narrower or wider confidence interval. Analogous issues exist for tests of proportions or correlations.

For t-tests, the researcher who requires more definitive information about the confidence interval may want to compute tolerance intervals—that is, the likelihood that the confidence interval will be no wider than some specific value. In this program, the 50% tolerance interval (corresponding to the median value) is displayed as a matter of course. The 80% (or other user-specified) tolerance interval is an option enabled from the View menu. For example, the researcher might report that in 50% of all studies, the mean would be reported with a 95% confidence interval no wider than 9 points, and in 80%

of all studies, the mean would be reported with a 95% confidence interval no wider than 10 points.

Note. The confidence interval displayed by the program is intended for anticipating the width of the confidence interval while planning a study, and not for computing the confidence interval after a study is completed. The computational algorithm used for t-tests includes an adjustment for the sampling distribution of the standard deviation that is appropriate for planning but not for analysis. The computational algorithms used for tests of proportions or a single correlation may be used for analysis as well.

Additional Reading

The bibliography includes a number of references that offer a more comprehensive treatment of precision. In particular, see Borenstein, 1994; Cohen, 1994; Bristol, 1989; Gardner and Altman, 1986; and Hahn and Meeker, 1991.

Significance Testing versus Effect Size Estimation

The two approaches outlined here—testing the null hypothesis of no effect and estimating the size of the effect—are closely connected. A study that yields a p-value of precisely 0.05 will yield a 95% confidence interval that begins (or ends) precisely at 0. A study that yields a p-value of precisely 0.01 will yield a 99% confidence interval that begins (or ends) precisely at 0. In this sense, reporting an effect size with corresponding confidence intervals can serve as a surrogate for tests of significance (if the confidence interval does not include the nil effect, the study is statistically significant) with the effect size approach focusing attention on the relevant issue. However, by shifting the focus of a report away from significance tests and toward the effect size estimate, we ensure a number of important advantages.

First, effect size focuses attention on the key issue. Usually, researchers and clinicians care about the size of the effect; the issue of whether or not the effect is nil is of relatively minor interest. For example, the clinician might recommend a drug, despite its potential for side effects, if he felt comfortable that it increased the remission rate by some specific amount, such as 20%, 30%, or 40%. Merely knowing that it increased the rate by *some* unspecified amount exceeding 0 is of little import. The effect size with confidence intervals focuses attention on the key index (how large the effect is), while providing likely boundaries for the lower and upper limits of the true effect size in the population.

Second, the focus on effect size, rather than on statistical significance, helps the researcher and the reader to avoid some mistakes that are common in the interpretation of significance tests. Since researchers care primarily about the size of the effect (and not whether the effect is nil), they tend to interpret the results of a significance test as though these results were an indication of effect size. For example, a p-value of 0.001 is assumed to reflect a large effect, while a p-value of 0.05 is assumed to reflect a moderate effect. This is inappropriate because the p-value is a function of sample size as well as effect size. Often, the non-significant p-value is assumed to indicate that the treatment has been proven ineffective. In fact, a non-significant p-value could reflect the fact that the treatment is not effective, but it could just as easily reflect the fact that the study was under-powered.

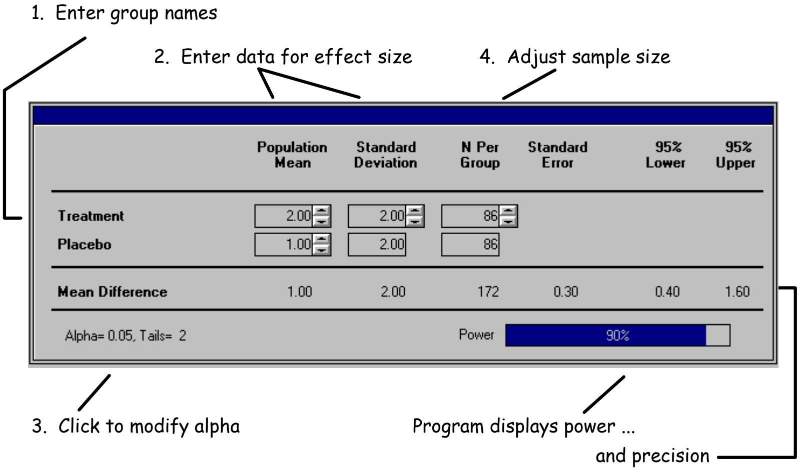
If power analysis is the appropriate precursor to a study that will test the null hypothesis, then precision analysis is the appropriate precursor to a study that will be used to estimate the size of a treatment effect. This program allows the researcher to take account of both.

Additional Reading

Suggested readings include the following: Borenstein, 1994; Cohen, 1992, 1994; Braitman, 1988; Bristol, 1989; Bulpitt, 1987; Detsky and Sackett, 1985; Feinstein, 1976; Fleiss, 1986a,b; Freiman et. al, 1978; Gardner and Altman, 1986; Gore, 1981; Makuch and Johnson, 1986; McHugh, 1984; Morgan, 1989; Rothman, 1978, 1986a,c; Simon, 1986; Smith and Bates, 1992.

3 The Main Screen

Computing Power Interactively



	Population Mean	Standard Deviation	N Per Group	Standard Error	95% Lower	95% Upper
Treatment	2.00	2.00	86			
Placebo	1.00	2.00	86			
Mean Difference	1.00	2.00	172	0.30	0.40	1.60

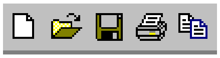
Alpha= 0.05, Tails= 2

Power 90%

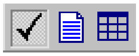
The precise format of the main screen will vary somewhat from one procedure to the next. Following are the key steps for most procedures:

- Optionally, enter names for the group(s).
- Optionally, modify alpha, confidence levels, and tails (choose *Alpha/CI/Tails* from the Options menu or click on the current value).
- Enter data for the effect size.
- Modify the sample size until power and/or precision reach the desired level (or click the *Find N* icon).
- Optionally, save various sets of parameters to the sensitivity analysis list.
- Optionally, click the *Report* icon to create a report.
- Optionally, click the *Make table* or *Make graph* icon to create a table or graph.

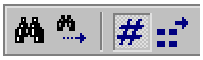
Toolbar



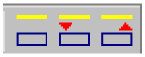
Icons on the left side of the toolbar are used to open, save, and print files and copy data to the clipboard. These functions are also available on the File menu.



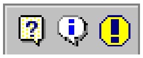
The next set of tools is used to navigate between screens. Click on the check mark to return to the main screen for a procedure. Click on the report, table, or graph tools to create and/or display the corresponding item. These functions are also available on the View menu.



The next set of icons provides tools for working with the main screen. The first icon (binoculars) is used to find the sample size required to yield the default value of power, and the next (small binoculars) is used to find the sample size required for any other value of power. The third icon toggles the display of power between an analog and a digital display. The fourth opens a dialog box that allows the user to set the number of decimal places for data entry and display.



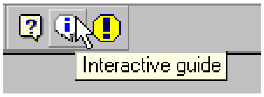
This set of icons is used to store and restore scenarios. The first icon displays the grid that is used to store scenarios. The second copies the current set of parameters from the screen into the grid, while the third restores a set of parameters from the grid back to the screen.



The last set of tools is used to activate various Help functions. Click on these tools to turn on or off the interactive guide, the interactive summary, and the Help system. These functions are also available on the Help menu.

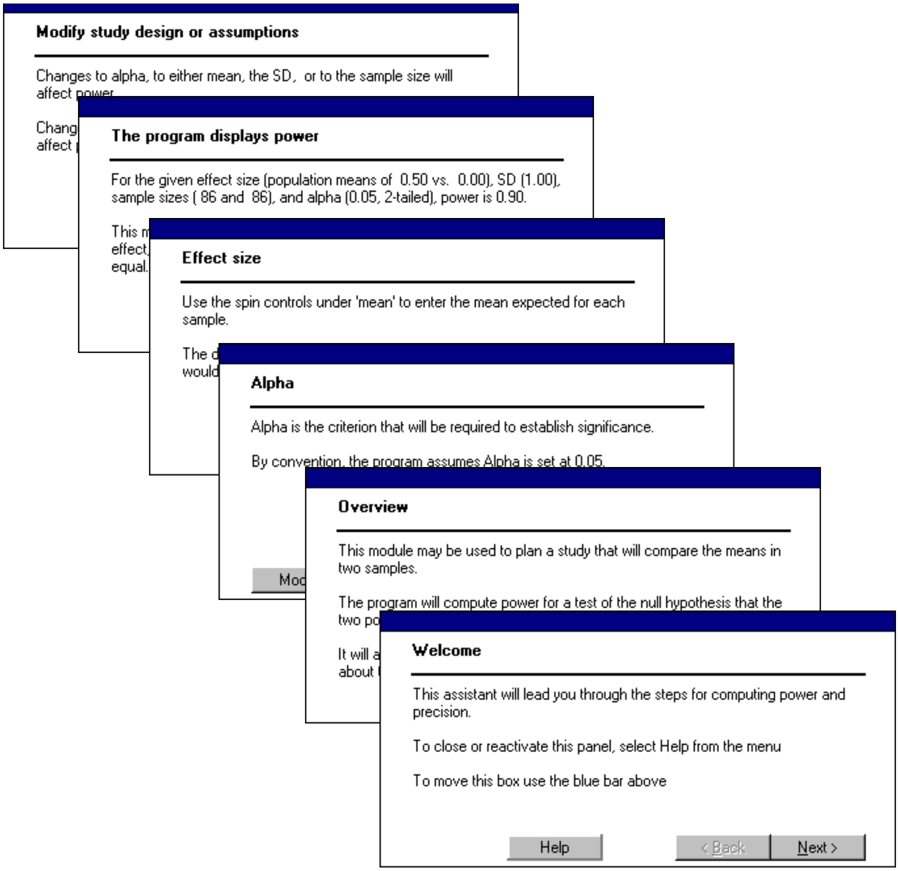
Interactive Guide

Each procedure includes an interactive guide that can be invoked by clicking the *Interactive Guide* icon or by choosing *Interactive guide* from the Help menu.



Click the *Interactive Guide* icon to display a series of panels that:

- Introduce the program's features for the current statistical procedure.
- Explain what type of data to enter.
- Provide a running commentary on the planning process.



Modify study design or assumptions

Changes to alpha, to either mean, the SD, or to the sample size will affect power.

Chang affect

The program displays power

For the given effect size (population means of 0.50 vs. 0.00), SD (1.00), sample sizes (86 and 86), and alpha (0.05, 2-tailed), power is 0.90.

This m effect equal.

Effect size

Use the spin controls under 'mean' to enter the mean expected for each sample.

The d would

Alpha

Alpha is the criterion that will be required to establish significance.

By convention, the program assumes Alpha is set at 0.05.

Overview

This module may be used to plan a study that will compare the means in two samples.

The program will compute power for a test of the null hypothesis that the two po

It will a about

Welcome

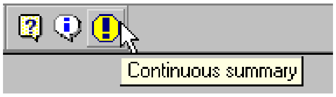
This assistant will lead you through the steps for computing power and precision.

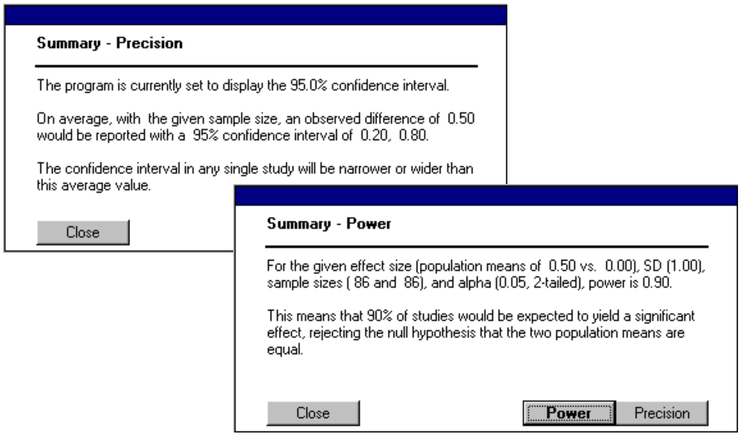
To close or reactivate this panel, select Help from the menu

To move this box use the blue bar above

Summary Panel

The program displays a summary panel that offers a text report for the current data. Click the *Display interactive summary* icon or choose *Continuous summary* from the Help menu.





Summary - Precision

The program is currently set to display the 95.0% confidence interval.

On average, with the given sample size, an observed difference of 0.50 would be reported with a 95% confidence interval of 0.20, 0.80.

The confidence interval in any single study will be narrower or wider than this average value.

Summary - Power

For the given effect size (population means of 0.50 vs. 0.00), SD (1.00), sample sizes (86 and 86), and alpha (0.05, 2-tailed), power is 0.90.

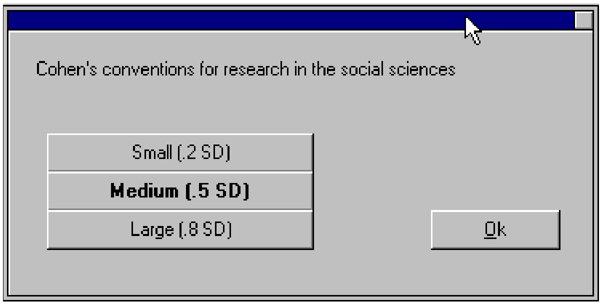
This means that 90% of studies would be expected to yield a significant effect, rejecting the null hypothesis that the two population means are equal.

The summary is an on-screen interactive panel. More extensive reports that can be annotated and exported to word processing programs are available in RTF format (see Chapter 5).

Effect Size Conventions

As a rule, the effect size should be based on the user's knowledge of the field and should reflect the smallest effect that would be important to detect in the planned study.

For cases in which the user is not able to specify an effect size in this way, Cohen has suggested specific values that might be used to represent *Small*, *Medium*, and *Large* effects in social science research.



Cohen's conventions for research in the social sciences

Click *Small*, *Medium*, or *Large* and the program will insert the corresponding effect size into the main screen.

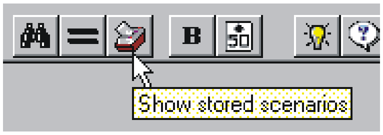
In this program, Cohen's conventions are available for one- and two-sample t-tests, one-sample and two-independent-samples tests of proportions, one- and two-sample correlations, and ANOVA.

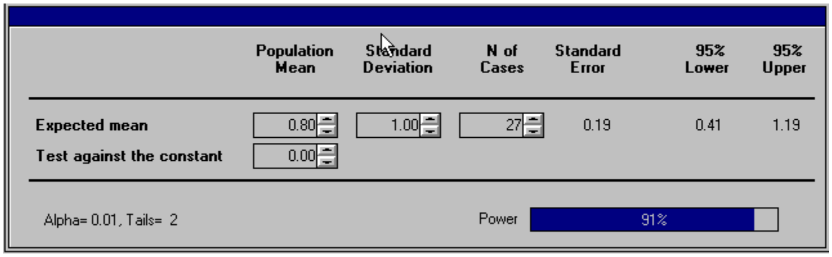
Sensitivity Analysis

The user may want to determine how power and precision vary if certain assumptions are modified. For example, how is power affected if we work with an effect (d) of 0.40, 0.45, or 0.50? How is power affected if we work with alpha of 0.05 or 0.01? This process is sometimes referred to as a sensitivity analysis.

The program features a tool to facilitate this process—any set of parameters entered into the main screen can be saved to a list. The list can be used to display power for various sets of parameters—the user can vary alpha, tails, sample size, effect size, etc., with power and precision displayed for each set. The list can be scrolled and sorted. Additionally, any line in the list can be restored to the main screen.

To display the sensitivity analysis list, click the *Show stored scenarios* icon or choose *Display stored scenarios* from the Tools menu.

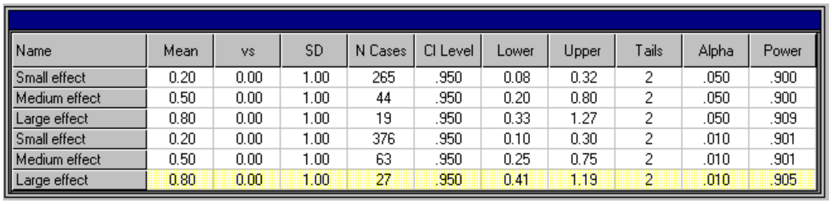




	Population Mean	Standard Deviation	N of Cases	Standard Error	95% Lower	95% Upper
Expected mean	0.80	1.00	27	0.19	0.41	1.19
Test against the constant	0.00					

Alpha= 0.01, Tails= 2

Power 91%



Name	Mean	vs	SD	N Cases	CI Level	Lower	Upper	Tails	Alpha	Power
Small effect	0.20	0.00	1.00	265	.950	0.08	0.32	2	.050	.900
Medium effect	0.50	0.00	1.00	44	.950	0.20	0.80	2	.050	.900
Large effect	0.80	0.00	1.00	19	.950	0.33	1.27	2	.050	.909
Small effect	0.20	0.00	1.00	376	.950	0.10	0.30	2	.010	.901
Medium effect	0.50	0.00	1.00	63	.950	0.25	0.75	2	.010	.901
Large effect	0.80	0.00	1.00	27	.950	0.41	1.19	2	.010	.905

Click the *Store main panel into list* icon to copy the current panel into the next line on the list. This includes all data on effect size, sample size, alpha, confidence intervals, and power. The user can store any number of these lines and then browse them. The adjacent tool will restore the selected line into the main panel, replacing any data currently in the panel.

Note. The program does not store information on computational options, such as the use of z rather than t for t-tests. When a study is restored from the list into the main panel, the options currently in effect will remain in effect.

Click the title at the top of any column to sort the list by that column.

The sensitivity analysis method is available for one- and two-sample tests of means, correlations, and independent proportions. For other tests (paired proportions, sign test, K × C proportions, ANOVA, and regression), the program does not feature this sensitivity box but does allow the creation of tables using the tables module and allows any scenario to be saved to disk for later retrieval.

Printing Sensitivity Analysis

To print the sensitivity analysis, right-click in the table of scenarios.

Copying to the Clipboard

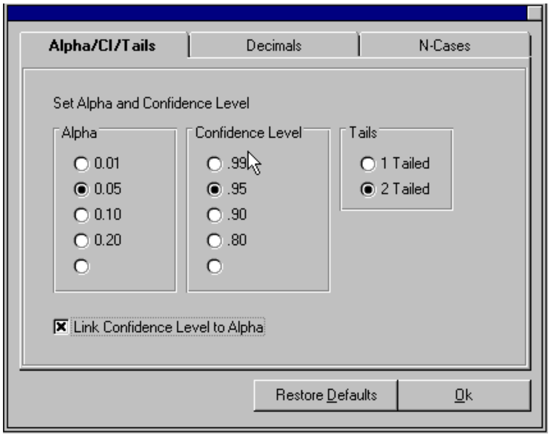
To copy the sensitivity analysis to the clipboard, right-click in the table of scenarios.

Saving Sensitivity Analysis

If the study is saved to disk, the sensitivity analysis is saved as well.

Note. The program can *automatically* generate tables and graphs in which sample size, alpha, and effect size are varied systematically (see Chapter 4).

Alpha, Confidence Level, and Tails



Set Alpha and Confidence Level

Alpha: ○ 0.01 ● 0.05 ○ 0.10 ○ 0.20 ○

Confidence Level: ○ .99 ● .95 ○ .90 ○ .80 ○

Tails: ○ 1 Tailed ● 2 Tailed

☒ Link Confidence Level to Alpha

Alpha and Confidence Level

Alpha, confidence intervals, and tails are set from the panel shown above. To activate this panel, click on the value currently shown for alpha, confidence level, or tails on the main screen. Or, choose *Alpha/CI/Tails* from the Options menu.

To set alpha, select one of the standard values or click *Other* and enter a value between 0.001 and 0.499. To set the confidence level, select any of the values offered or click *Other* and enter a value between 0.501 and 0.999.

The check box at bottom of this panel allows you to link the confidence level with alpha. When this option is selected, setting the confidence level at 95% implies that alpha will be set at 0.05; setting the confidence level at 99% implies that alpha will be set at 0.01, and so on. Similarly, any change to alpha is reflected in the confidence level.

Tails

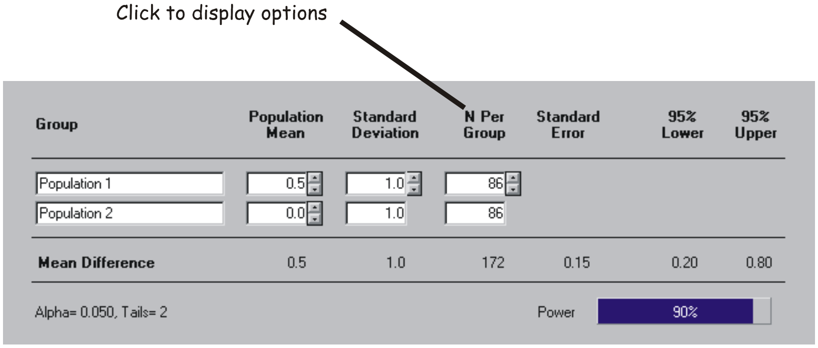
The value set for tails applies to both power and confidence level. Allowed values are 1 and 2. For procedures that are nondirectional (ANOVA, regression, or K × C crosstabs), the tails option is not displayed.

When the test is two-tailed, power is computed for an effect in either direction. When the test is one-tailed, power is computed for a one-tailed test in the direction that will yield higher power.

When the test is two-tailed, the program displays a two-tailed confidence interval (for example, based on a z-value of 1.96 for the 95% interval). When the test is one-tailed, the program displays a one-tailed interval (for example, using a z-value of 1.64). For a one-tailed interval, the program will display both the lower and upper limits, but the user should interpret only one of these. The "interval" extends either from infinity to the upper value shown or from the lower value shown to infinity.

Setting Options for the Number of Cases

Click to display options



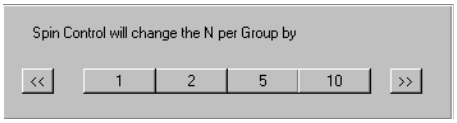
Group	Population Mean	Standard Deviation	N Per Group	Standard Error	95% Lower	95% Upper
Population 1	0.5	1.0	86			
Population 2	0.0	1.0	86			
Mean Difference	0.5	1.0	172	0.15	0.20	0.80

Alpha= 0.050, Tails= 2

Power 90%

To display the options for setting the sample size, click *N Per Group* or choose *N-Cases* from the Options menu.

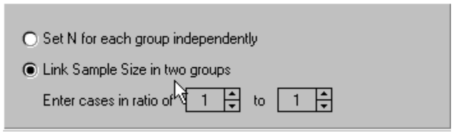
Setting the Increment for Spin Control



Spin Control will change the N per Group by

The program allows the user to modify the sample size by using a spin control. The panel shown above allows the user to customize the spin control by setting the size of the increment.

Linking the Number of Cases in Two Groups



○ Set N for each group independently

● Link Sample Size in two groups

Enter cases in ratio of 1 to 1

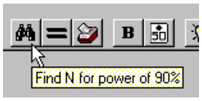
Initially, the program assumes that cases will be allocated to two groups in a ratio of 1:1. This panel allows the user to set a different ratio (for example, 1:2 or 3:5). (Note that power is based on the harmonic mean of the two sample sizes, which means that it is controlled primarily by the lower of the two sample sizes.) As long as the sample sizes are linked, the program will expect (and accept) a sample size for the first group only.

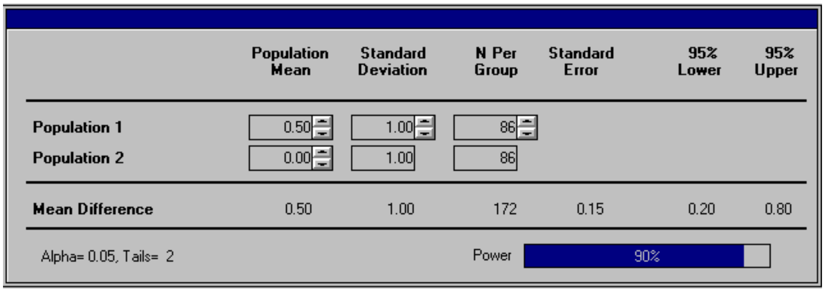
The user can also set the number of cases for each group independently of the other.

Note. Some of the program’s features (Find N, Table, and Graph) will not operate unless the sample sizes in the two groups are linked, but these features do not require that cases be assigned to the two groups in even numbers.

Finding the Number of Cases Automatically

1. Enter effect size
2. Modify alpha
3. Click to find required N





	Population Mean	Standard Deviation	N Per Group	Standard Error	95% Lower	95% Upper
Population 1	0.50	1.00	86			
Population 2	0.00	1.00	86			
Mean Difference	0.50	1.00	172	0.15	0.20	0.80

Alpha= 0.05, Tails= 2

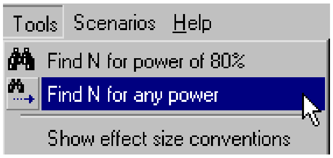
Power 90%

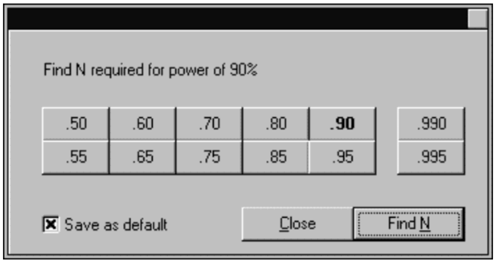
The program will find the number of cases required for the default value of power. Enter an effect size, modify alpha, and click the *Find N* icon. By default, the program assumes that the user wants power of 80%, but this can be modified (see "Modifying the Default Value of Power" below). Note that the number of cases required for precision is usually different from the number of cases required for power. The number of cases required for precision can be found by using the spin control to modify the number until precision is appropriate.

In situations in which the user has specified that cases should be allocated unevenly to the two groups (for example, in a ratio of 1:2), the program will honor this allocation scheme in finding the required sample size.

Modifying the Default Value of Power

To temporarily modify the default value for required power, press choose *Find N for any power* from the Tools menu.




Find N required for power of 90%

☒ Save as default

- To find the number of case required for a specific power, click the power. For example, to find the number of cases for power of 95%, click 0.95. The required number is shown immediately.
- To change the required power to 95%, select *Save as default* and then click 0.95. This new power will remain in effect until changed. (The *Find N* icon and the *Find N* menu option will read "95%" as well.) Power will revert to the default when you leave the module.

Printing the Main Screen

To print the data on the main screen, click the *Print* icon or choose *Print* from the File menu.

Copying to the Clipboard

To copy the data from the main screen to the clipboard, click the *Copy to clipboard* icon or choose *Clipboard* from the File menu. Then switch to a word processing program and choose *Paste* from the Edit menu. The data from the main screen cannot be pasted into the program's report screen.

Saving and Restoring Files

Any file can be saved to disk and later retrieved. Click the *Save file* icon or choose *Save study data* from the File menu.

The program saves the basic data (effect size, alpha, sample size, and confidence level), as well as any sets of data that had been stored, to the sensitivity analysis box.

Note. The program does not store computation options; the options in effect when the file is later retrieved will be used to recompute power at that time.

The *Save study data* menu command does not save tables, graphs, or reports. These can be saved in their own formats. They can be:

- Copied to a word processing program and saved with the word processing document
- Recreated at a later point by retrieving the basic data and clicking the icon for a report, table, or graph

The file must be saved from within a module (for example, from within the t-test module or the ANOVA module). Files can be retrieved from any point in the program, including the opening screen.

Data files are stored with an extension of *.pow*. Double-clicking on one of these files in the Windows Explorer will cause Windows to launch Power And Precision and load the data file.

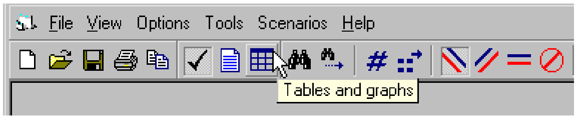
4 Tables and Graphs

Overview

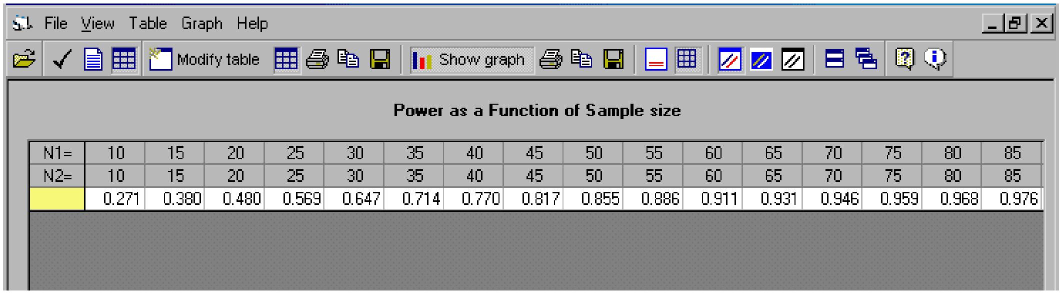
The program will create tables and graphs that show how power varies as a function of sample size, effect size, alpha, and other factors. A basic table can be created with a single click, and factors can be added in a matter of seconds. The graphs mirror the tables and, as such, are created and updated automatically as changes are made to a table.

Table Basics

To create a table, enter values on the main (interactive) screen, and then click the *Tables and graphs* icon.



The program automatically creates a table with the effect size and alpha taken from the main screen. Additionally, the sample size (start, increment, and end values) are set automatically, based on the effect size and alpha.

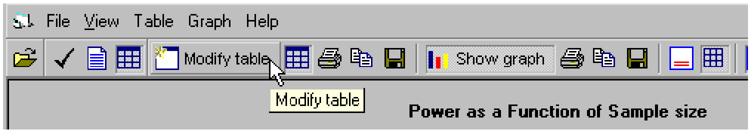


Power as a Function of Sample size

N1=	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85
N2=	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85
	0.271	0.380	0.480	0.569	0.647	0.714	0.770	0.817	0.855	0.886	0.911	0.931	0.946	0.959	0.968	0.976

Adding Factors to Tables

A basic table shows power as a function of sample size. To add a factor (for example, alpha) click *Modify table* on the toolbar.

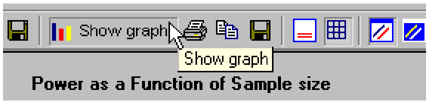


When you include multiple factors (for example, alpha and effect size), the program automatically nests one inside the other.

Note. To change the sequence of the nesting, simply drag and drop the columns to the desired position.

Displaying Graphs

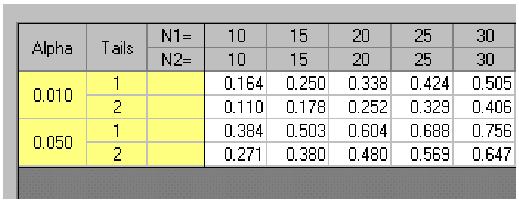
To display a graph corresponding to the table, click *Show graph* on the toolbar. The graph is created automatically.



Multiple Factors in Graphs

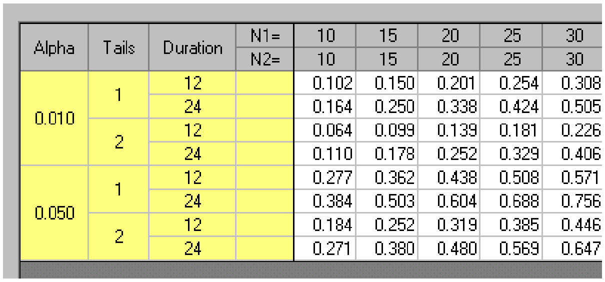
- When a table includes no factors other than sample size, the graph has only a single line.
- When a table has one factor (for example, alpha) in addition to sample size, the graph has multiple lines (for example, one for each level of alpha) and a legend is added.
- When a table has two additional factors (for example, alpha and effect size), the program will create multiple graphs.

❖ Determine which column should be used for the lines within a graph and drag that column to the right-most position (*Tails* in the following figure).



Alpha	Tails	N1=	10	15	20	25	30
		N2=	10	15	20	25	30
0.010	1		0.164	0.250	0.338	0.424	0.505
	2		0.110	0.178	0.252	0.329	0.406
0.050	1		0.384	0.503	0.604	0.688	0.756
	2		0.271	0.380	0.480	0.569	0.647

- ❖ Decide which column should be used to define the graph and drag that column to the adjacent position (*Alpha* in the following figure).
  - When a table has three or more additional factors (for example, alpha, tails, and study duration), follow the same procedure as for two factors, and then double-click in the third column from the right to select the level of the other factors to be plotted (*Alpha* in the following figure).



Alpha	Tails	Duration	N1=	10	15	20	25	30
			N2=	10	15	20	25	30
0.010	1	12		0.102	0.150	0.201	0.254	0.308
		24		0.164	0.250	0.338	0.424	0.505
	2	12		0.064	0.099	0.139	0.181	0.226
		24		0.110	0.178	0.252	0.329	0.406
0.050	1	12		0.277	0.362	0.438	0.508	0.571
		24		0.384	0.503	0.604	0.688	0.756
	2	12		0.184	0.252	0.319	0.385	0.446
		24		0.271	0.380	0.480	0.569	0.647

Saving, Printing, or Exporting Tables

- ❖ Print a table directly to a printer.
- ❖ Save a table to disk for later retrieval by this program.
- ❖ Copy a table to the clipboard as a bitmap and paste it into a word processing program, such as Word.
- ❖ Copy a table to the clipboard as a data file and paste it into a spreadsheet program, such as Excel.

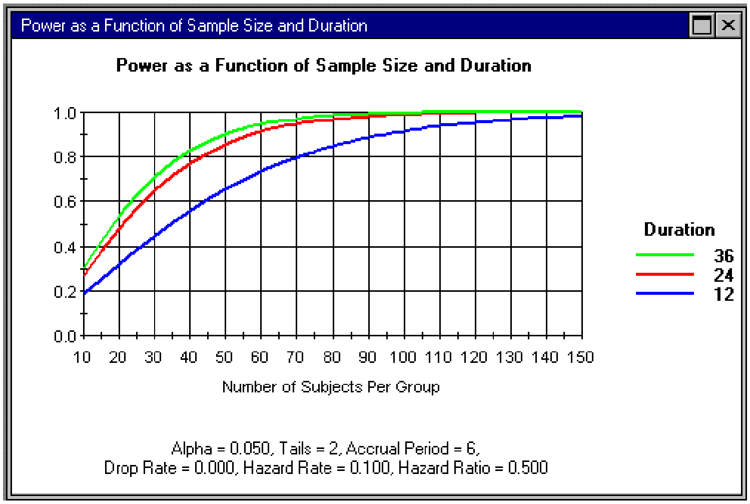
Saving, Printing, or Exporting Graphs

- ❖ Print a graph directly to the printer.
- ❖ Save a graph to disk in *.bmp*, *.wmf*, or *.emf* format (and import it into other programs, such as Word or PowerPoint).
- ❖ Copy a graph to the clipboard in *.bmp* format and paste it into another program.

  When there are multiple graphs, you can print, copy, or save them as a group or individually.

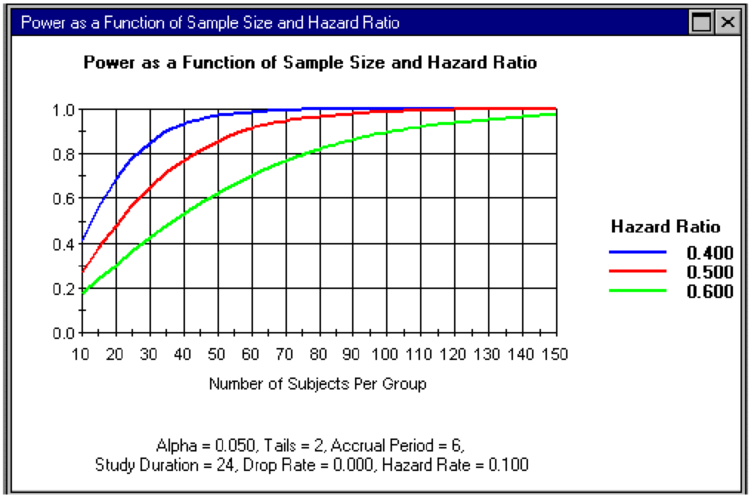
Role of Tables and Graphs in Power Analysis

Graphs provide a comprehensive picture of the available options for the study design. For example, if you are planning a survival study, you may be considering several options for the study duration. A graph of power as a function of sample size and study duration (shown below) can play a key role in the analysis. The graph shows that for power of 80%, you would need a sample size of about 45 per group for 24 months, compared with 38 per group for 36 months or 70 per group for 12 months. On this basis, you might decide that a 24-month study makes more sense than either of the other options. While it is possible to get the same information from the interactive screen, the graph is much easier to interpret.



Alpha = 0.050, Tails = 2, Accrual Period = 6,
Drop Rate = 0.000, Hazard Rate = 0.100, Hazard Ratio = 0.500

Additionally, a power analysis requires that you report more than a single value for power. Rather than report, for example, that "the hazard ratio is 0.50 and, therefore, power is 80% with 45 patients per group," it is more informative to report that "with 45 patients per group, we have power of 60% if the hazard ratio is 0.60; power of 80% if the hazard ratio is 0.50; and power of 95% if the hazard ratio is 0.40." Again, this type of information is better presented graphically.

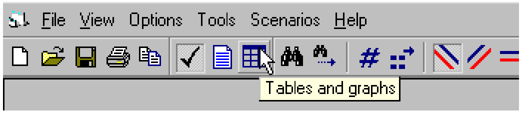


Alpha = 0.050, Tails = 2, Accrual Period = 6,
Study Duration = 24, Drop Rate = 0.000, Hazard Rate = 0.100

Creating Tables

The program will create a table and graph in which any number of factors, such as sample size, alpha, and effect size, are varied systematically.

To create a table and graph, enter values on the main (interactive) screen and then click the *Tables and graphs* icon. The program automatically creates a table and graph. The initial values for effect size and alpha are taken from the main screen. The initial values for sample size (start, increment, and end) are assigned automatically as well.



Setting Sample Size Automatically

When the program creates a table, the sample size (start, increment, and end values) for the table and graph are set automatically, based on the effect size and alpha. For example, if the effect size is large, the sample size might run from 10 to 50 in increments of 2. If the effect size is small, the sample size might run from 50 to 1000 in increments of 10. The sample size will be set to yield a graph in which power extends into the range of 95% to 99%. If you want to focus on one part of the range (or to extend the range of sample sizes), you can set the sample size manually.

Setting Sample Size Manually

- Click *Modify table*.
- Click the *Sample size* tab.
- Enter the new values for Start, Increment, and Final.
- Click *Apply* or *OK*.
- To return to automatic mode, select *Set automatically*.

Adding Factors to Tables

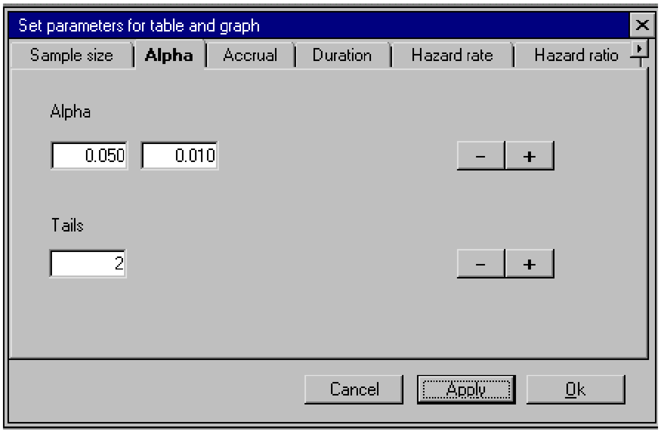
Initially, the program creates a table in which sample size varies but all other factors, such as effect size and alpha, are constant. You have the option of adding factors to a table.

To add factors to a table:

- Create the initial table.
- Click *Modify table* on the toolbar or on the Table menu.
- Click the desired tab (for example, *Alpha*, *Accrual*, *Duration*).

The initial value of any parameter is taken from the interactive screen. To display multiple values for any parameter, click the plus sign (+) and enter multiple values. Then click *Apply* or *OK*. Here, for example, the program will add *Alpha* as a factor with values of 0.01 and 0.05.

The table will now include a column labeled *Alpha* with values of 0.01 and 0.05. (In the table, the values are sorted automatically in order of increasing power.)

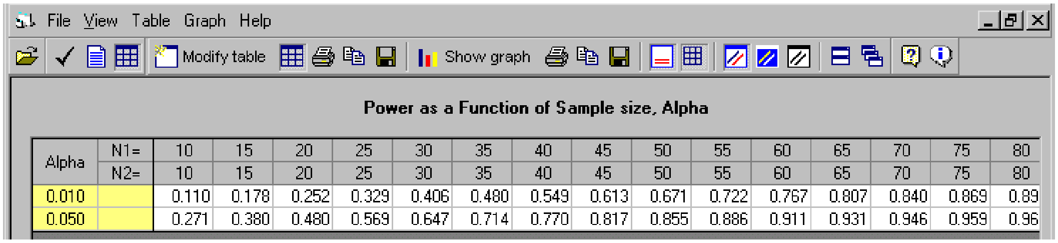


Alpha

0.050 0.010

Tails

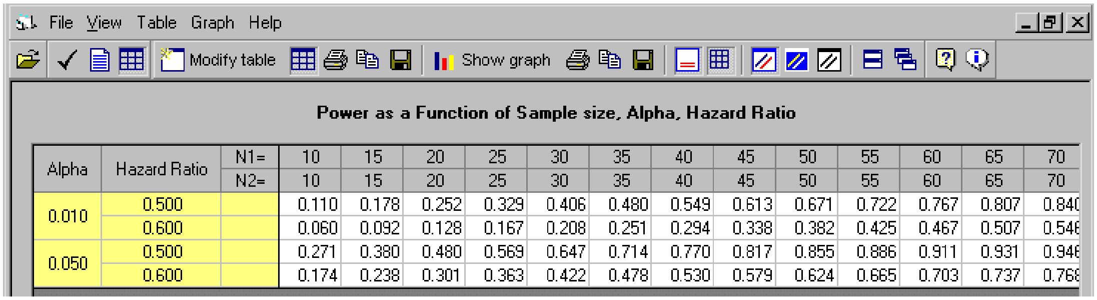
2



Power as a Function of Sample size, Alpha

Alpha	N1=	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80
	N2=	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80
0.010		0.110	0.178	0.252	0.329	0.406	0.480	0.549	0.613	0.671	0.722	0.767	0.807	0.840	0.869	0.89
0.050		0.271	0.380	0.480	0.569	0.647	0.714	0.770	0.817	0.855	0.886	0.911	0.931	0.946	0.959	0.96

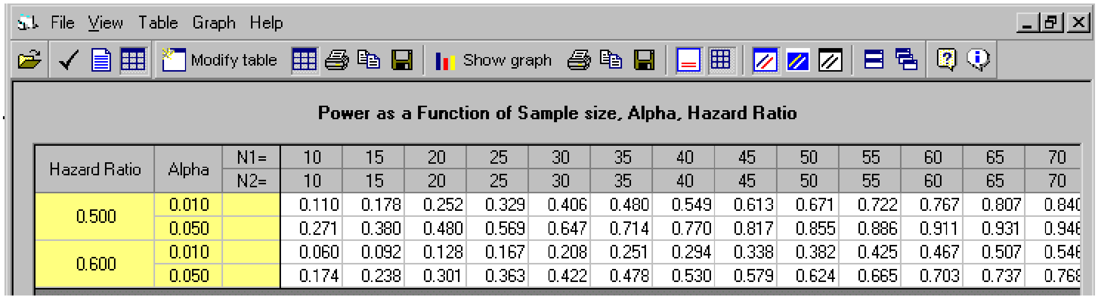
If you specify two factors, the program will automatically nest each factor inside the others. For example, if you specify that *Alpha = 0.01, 0.05* and *Hazard Ratio = 0.50, 0.60, 0.70*, the program will create six rows corresponding to the six possible combinations, as shown in the table below.



Power as a Function of Sample size, Alpha, Hazard Ratio

Alpha	Hazard Ratio	N1=	10	15	20	25	30	35	40	45	50	55	60	65	70
		N2=	10	15	20	25	30	35	40	45	50	55	60	65	70
0.010	0.500		0.110	0.178	0.252	0.329	0.406	0.480	0.549	0.613	0.671	0.722	0.767	0.807	0.840
	0.600		0.060	0.092	0.128	0.167	0.208	0.251	0.294	0.338	0.382	0.425	0.467	0.507	0.546
0.050	0.500		0.271	0.380	0.480	0.569	0.647	0.714	0.770	0.817	0.855	0.886	0.911	0.931	0.946
	0.600		0.174	0.238	0.301	0.363	0.422	0.478	0.530	0.579	0.624	0.665	0.703	0.737	0.768

Important. To switch the order of the factors on the table, drag and drop the columns. Columns must be dropped inside the shaded range on the left. The table above shows *Alpha* as the first factor and *Hazard Ratio* as the second. Drag the *Alpha* column to the right by one column, and the table is displayed with *Hazard Ratio* as the first factor and *Alpha* as the second.



Power as a Function of Sample size, Alpha, Hazard Ratio

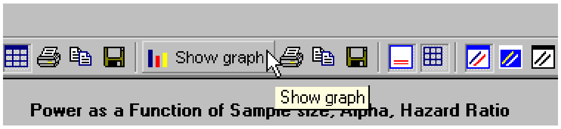
Hazard Ratio	Alpha	N1=	10	15	20	25	30	35	40	45	50	55	60	65	70
		N2=	10	15	20	25	30	35	40	45	50	55	60	65	70
0.500	0.010		0.110	0.178	0.252	0.329	0.406	0.480	0.549	0.613	0.671	0.722	0.767	0.807	0.840
	0.050		0.271	0.380	0.480	0.569	0.647	0.714	0.770	0.817	0.855	0.886	0.911	0.931	0.946
0.600	0.010		0.060	0.092	0.128	0.167	0.208	0.251	0.294	0.338	0.382	0.425	0.467	0.507	0.546
	0.050		0.174	0.238	0.301	0.363	0.422	0.478	0.530	0.579	0.624	0.665	0.703	0.737	0.768

To remove factors from the table:

- ❖ Click *Modify table* on the toolbar or on the Table menu.
- ❖ Click the desired tab (for example, *Alpha*, *Accrual*, *Duration*).
- ❖ Click the minus sign (–) until only one value remains for that factor.

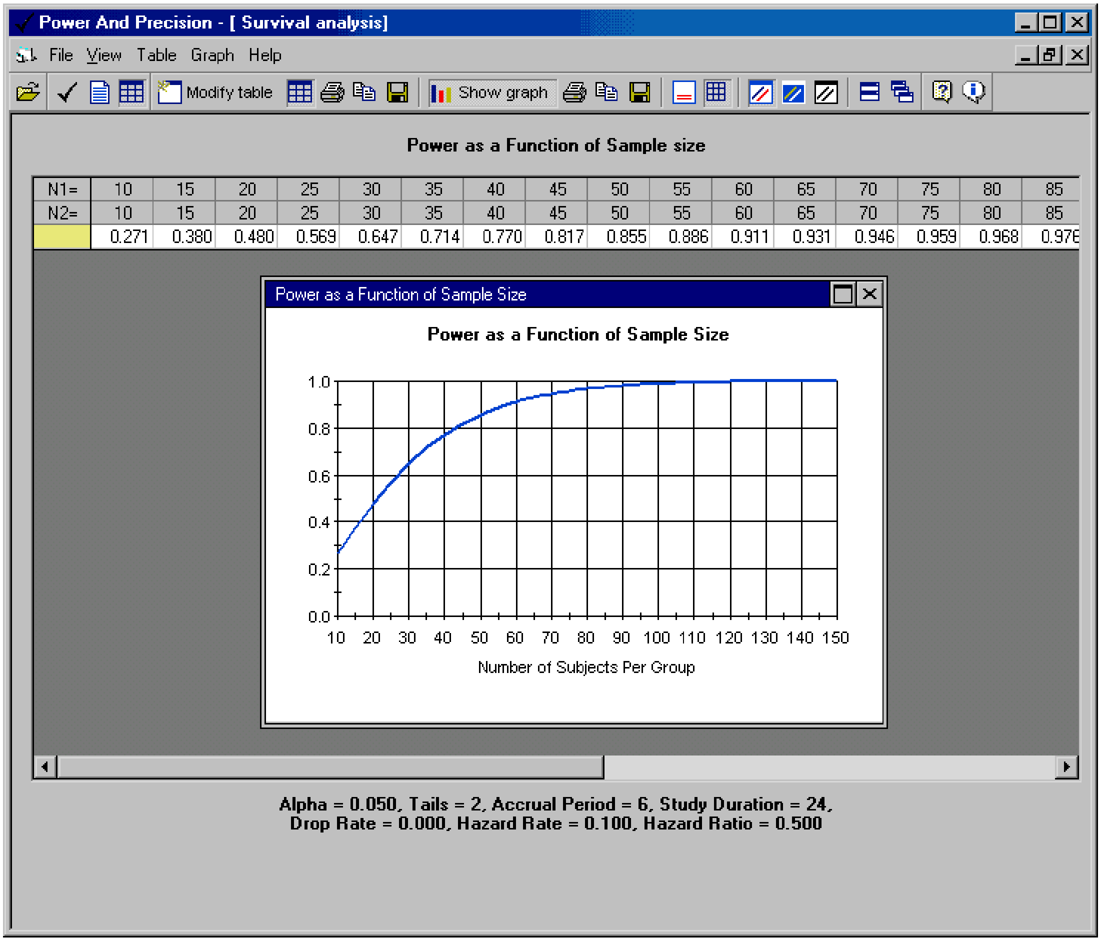
Displaying Graphs

To display a graph corresponding to the table, click *Show graph*. The graph is created automatically.



Power as a Function of Sample size, Alpha, Hazard Ratio

Both the table and the graph show power as a function of sample size in the range of 20 to 150 subjects per group.


Power And Precision - [Survival analysis]
File
View
Table
Graph
Help
Modify table
Show graph
Power as a Function of Sample Size
Power as a Function of Sample Size
1.0
0.8
0.6
0.4
0.2
0.0
10
20
30
40
50
60
70
80
90
100
110
120
130
140
150
Number of Subjects Per Group

Power as a Function of Sample size

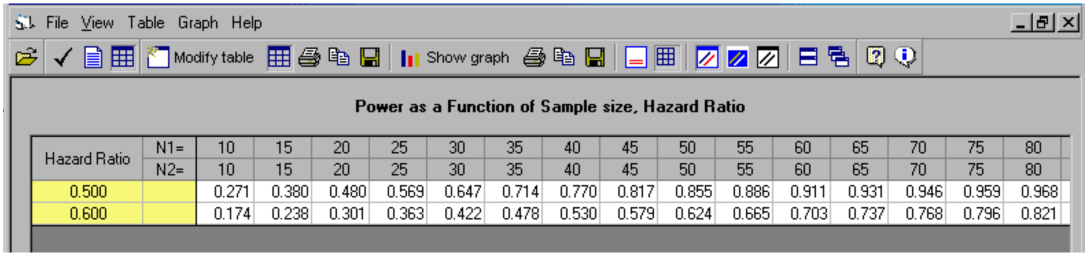
N1=	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85
N2=	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85
	0.271	0.380	0.480	0.569	0.647	0.714	0.770	0.817	0.855	0.886	0.911	0.931	0.946	0.959	0.968	0.976

Alpha = 0.050, Tails = 2, Accrual Period = 6, Study Duration = 24,
Drop Rate = 0.000, Hazard Rate = 0.100, Hazard Ratio = 0.500

When the table includes only one factor in addition to sample size:

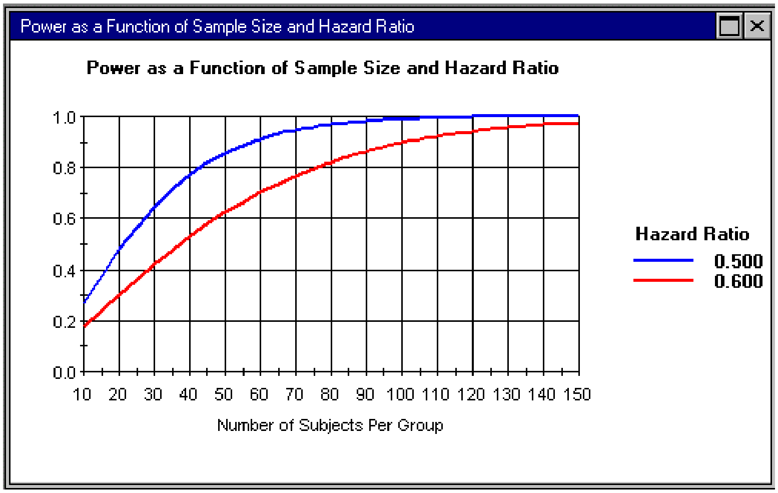
If there is only one factor in the table aside from sample size, the program displays one graph, and each level in this factor is a line in the graph.

The next table shows power as a function of sample size and hazard ratio. The graph shows sample size on the *x* axis and includes a line for each value of the hazard ratio that appears in the table.



Power as a Function of Sample size, Hazard Ratio

Hazard Ratio	N1=	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80
	N2=	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80
0.500		0.271	0.380	0.480	0.569	0.647	0.714	0.770	0.817	0.855	0.886	0.911	0.931	0.946	0.959	0.968
0.600		0.174	0.238	0.301	0.363	0.422	0.478	0.530	0.579	0.624	0.665	0.703	0.737	0.768	0.796	0.821



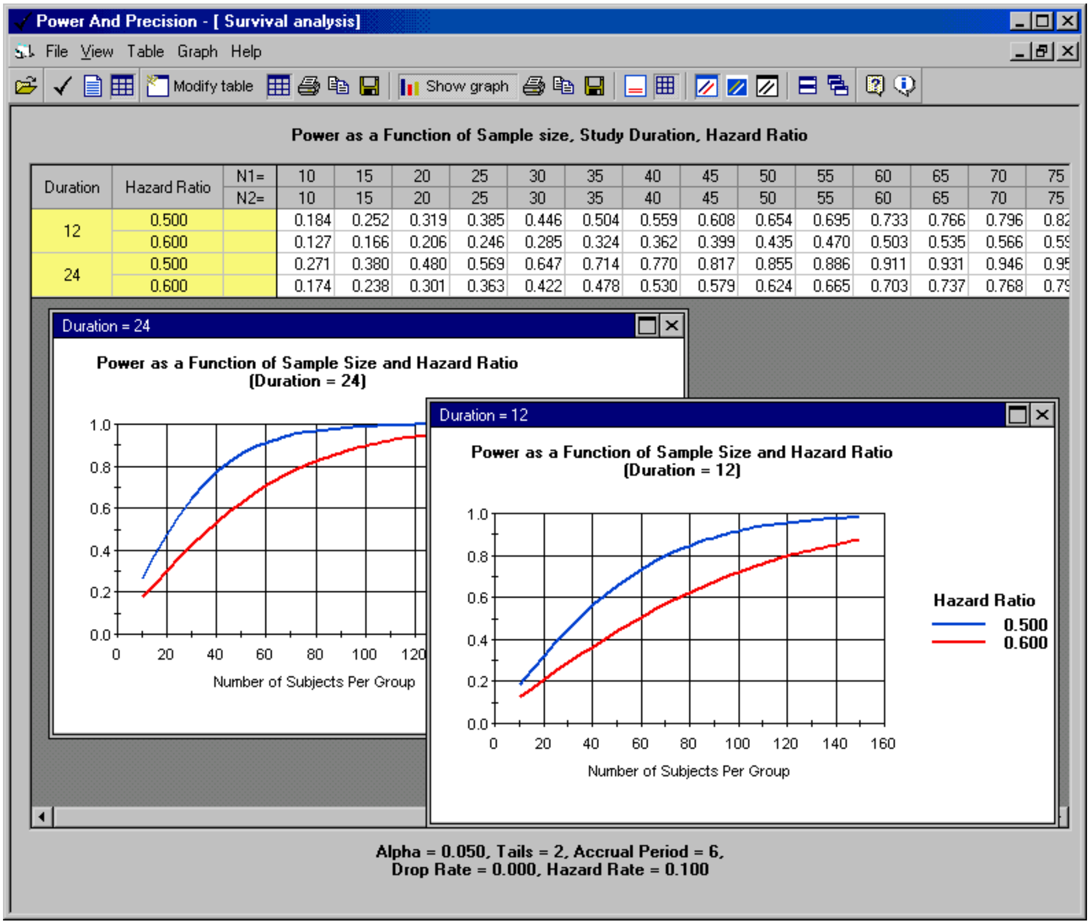
When the table includes two factors in addition to sample size:

❖ To modify the graphs, drag and drop the columns in the table.

- A separate graph is created for each level in the first column.
- The second column is used as the factor inside each graph.

The columns in the table are *Duration* followed by *Hazard Ratio*.

- The program creates one graph for each level of *Duration*.
- Within each graph, *Hazard Ratio* is the factor.


Power And Precision - [Survival analysis]
File View Table Graph Help
Modify table
Show graph

Power as a Function of Sample size, Study Duration, Hazard Ratio

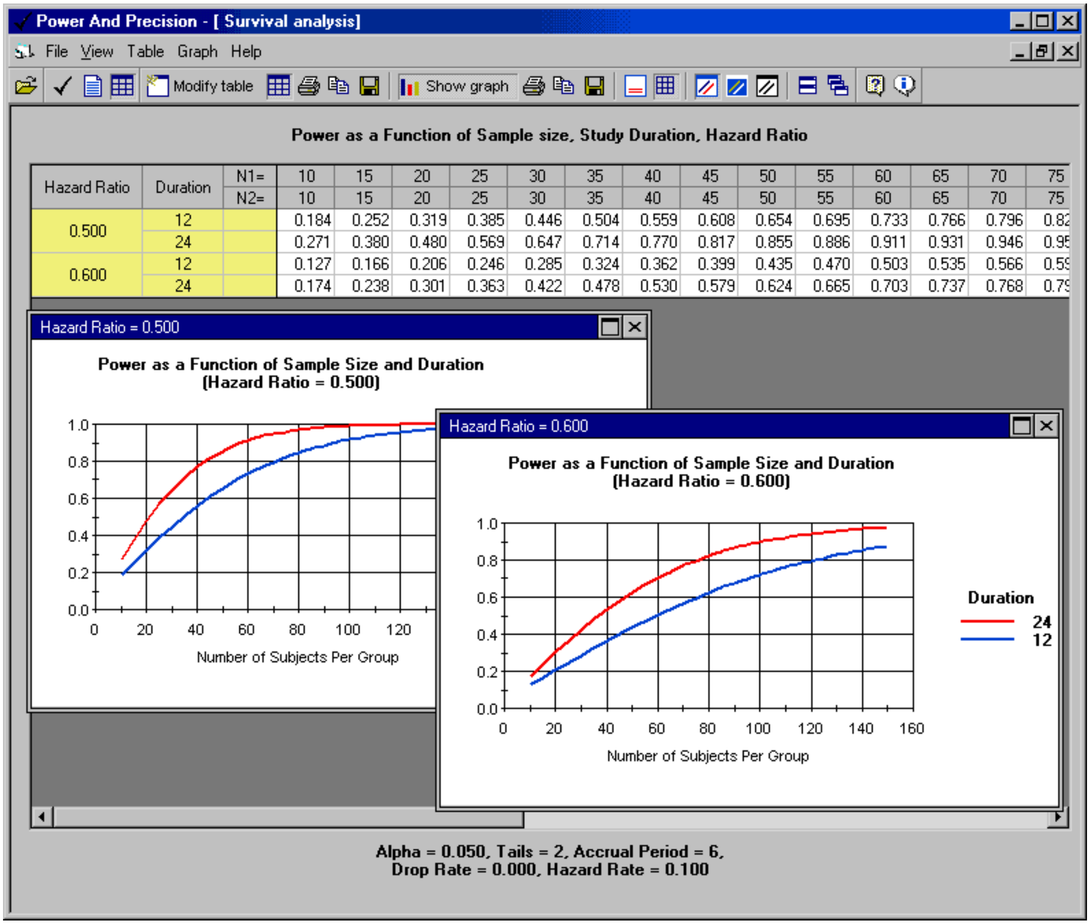
Duration	Hazard Ratio	N1=	10	15	20	25	30	35	40	45	50	55	60	65	70	75
		N2=	10	15	20	25	30	35	40	45	50	55	60	65	70	75
12	0.500		0.184	0.252	0.319	0.385	0.446	0.504	0.559	0.608	0.654	0.695	0.733	0.766	0.796	0.82
	0.600		0.127	0.166	0.206	0.246	0.285	0.324	0.362	0.399	0.435	0.470	0.503	0.535	0.566	0.59
24	0.500		0.271	0.380	0.480	0.569	0.647	0.714	0.770	0.817	0.855	0.886	0.911	0.931	0.946	0.95
	0.600		0.174	0.238	0.301	0.363	0.422	0.478	0.530	0.579	0.624	0.665	0.703	0.737	0.768	0.79

Duration = 24
Power as a Function of Sample Size and Hazard Ratio
(Duration = 24)
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
120
Number of Subjects Per Group
Duration = 12
Power as a Function of Sample Size and Hazard Ratio
(Duration = 12)
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
120
140
160
Number of Subjects Per Group
Hazard Ratio
0.500
0.600

Alpha = 0.050, Tails = 2, Accrual Period = 6,
Drop Rate = 0.000, Hazard Rate = 0.100

❖ To switch the factors in the graphs, drag and drop the columns in the table.

- The position of the two columns has been switched to *Hazard Ratio* followed by *Duration*.
- The program creates one graph for each level of *Hazard Ratio*.
- Within each graph, *Duration* is the factor.


Power And Precision - [Survival analysis]
File View Table Graph Help
Modify table
Show graph
Hazard Ratio = 0.500
Power as a Function of Sample Size and Duration
(Hazard Ratio = 0.500)
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
120
Number of Subjects Per Group
Hazard Ratio = 0.600
Power as a Function of Sample Size and Duration
(Hazard Ratio = 0.600)
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
120
140
160
Number of Subjects Per Group
Duration
24
12

Power as a Function of Sample size, Study Duration, Hazard Ratio

Hazard Ratio	Duration	N1=	10	15	20	25	30	35	40	45	50	55	60	65	70	75
		N2=	10	15	20	25	30	35	40	45	50	55	60	65	70	75
0.500	12		0.184	0.252	0.319	0.385	0.446	0.504	0.559	0.608	0.654	0.695	0.733	0.766	0.796	0.82
	24		0.271	0.380	0.480	0.569	0.647	0.714	0.770	0.817	0.855	0.886	0.911	0.931	0.946	0.95
0.600	12		0.127	0.166	0.206	0.246	0.285	0.324	0.362	0.399	0.435	0.470	0.503	0.535	0.566	0.59
	24		0.174	0.238	0.301	0.363	0.422	0.478	0.530	0.579	0.624	0.665	0.703	0.737	0.768	0.79

Alpha = 0.050, Tails = 2, Accrual Period = 6,
Drop Rate = 0.000, Hazard Rate = 0.100

Tip. The factor of interest should be moved to the right-most position in the table (*Duration* in this example). This factor will then serve as the factor within graphs, which makes it easier to study its impact on power.

When the table includes three factors in addition to sample size:

To create graphs when the table includes three factors in addition to sample size:

- ❖ Determine which column should be used for the lines *within a graph* and drag that column to the right-most position (*Hazard Ratio* in this example).

- ❖ Decide which column should be used to *define the graph* and drag that column to the adjacent position (*Duration* in this example).

  The lines within the graph will correspond to the hazard ratio, and you will have a separate graph for each study duration.

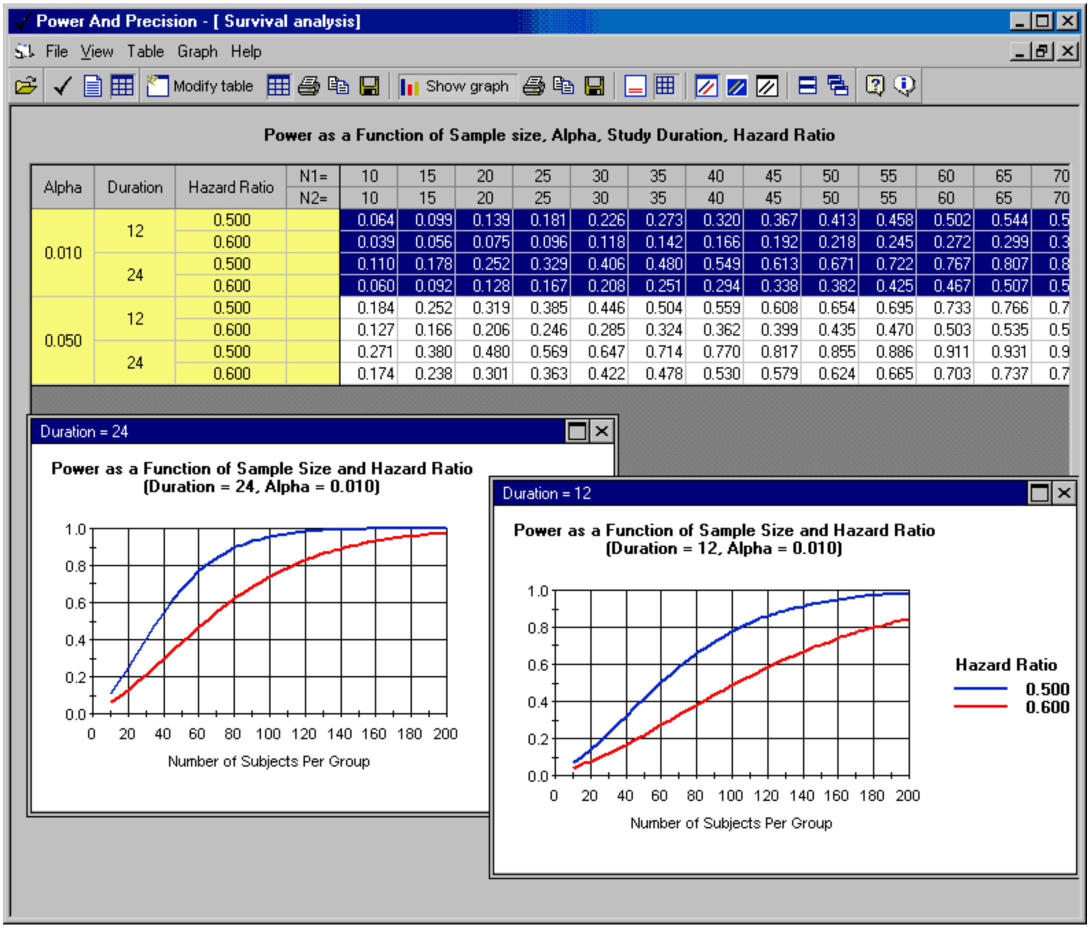
- ❖ The next factor to consider is the one in the next column to the left (*Alpha* in this example). Double-click on either value of alpha (0.01 or 0.05). The graphs will be based on this value.

- ❖ In this example (see the figure below), the user has double-clicked on *Alpha = 0.01*. The section of the table that corresponds to *Alpha = 0.01* has been highlighted and both graphs show *Alpha = 0.01* as part of the title. Click on another value in the *Alpha* column and the graphs will immediately change to reflect this value.

Synopsis:

- The first column (*Alpha*) is used to define a parameter that is constant in all graphs. Click on *Alpha = 0.01* and all graphs show *Alpha = 0.01*.
- The next column will vary from one graph to the next.
- The final (right-most) column is used to define the lines within each graph.

Tip. To create a separate graph for each level of alpha, switch the position of the *Alpha* and *Duration* columns. To use alpha as the factor within graphs, switch the position of the *Alpha* and *Hazard Ratio* columns.


Power And Precision - [Survival analysis]
File View Table Graph Help
Modify table
Show graph
Duration = 24
Power as a Function of Sample Size and Hazard Ratio
(Duration = 24, Alpha = 0.010)
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
120
140
160
180
200
Number of Subjects Per Group
Duration = 12
Power as a Function of Sample Size and Hazard Ratio
(Duration = 12, Alpha = 0.010)
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
120
140
160
180
200
Number of Subjects Per Group
Hazard Ratio
0.500
0.600

Power as a Function of Sample size, Alpha, Study Duration, Hazard Ratio

Alpha	Duration	Hazard Ratio	N1=	10	15	20	25	30	35	40	45	50	55	60	65	70
			N2=	10	15	20	25	30	35	40	45	50	55	60	65	70
0.010	12	0.500		0.064	0.099	0.139	0.181	0.226	0.273	0.320	0.367	0.413	0.458	0.502	0.544	0.5
		0.600		0.039	0.056	0.075	0.096	0.118	0.142	0.166	0.192	0.218	0.245	0.272	0.299	0.3
	24	0.500		0.110	0.178	0.252	0.329	0.406	0.480	0.549	0.613	0.671	0.722	0.767	0.807	0.8
		0.600		0.060	0.092	0.128	0.167	0.208	0.251	0.294	0.338	0.382	0.425	0.467	0.507	0.5
0.050	12	0.500		0.184	0.252	0.319	0.385	0.446	0.504	0.559	0.608	0.654	0.695	0.733	0.766	0.7
		0.600		0.127	0.166	0.206	0.246	0.285	0.324	0.362	0.399	0.435	0.470	0.503	0.535	0.5
	24	0.500		0.271	0.380	0.480	0.569	0.647	0.714	0.770	0.817	0.855	0.886	0.911	0.931	0.9
		0.600		0.174	0.238	0.301	0.363	0.422	0.478	0.530	0.579	0.624	0.665	0.703	0.737	0.7

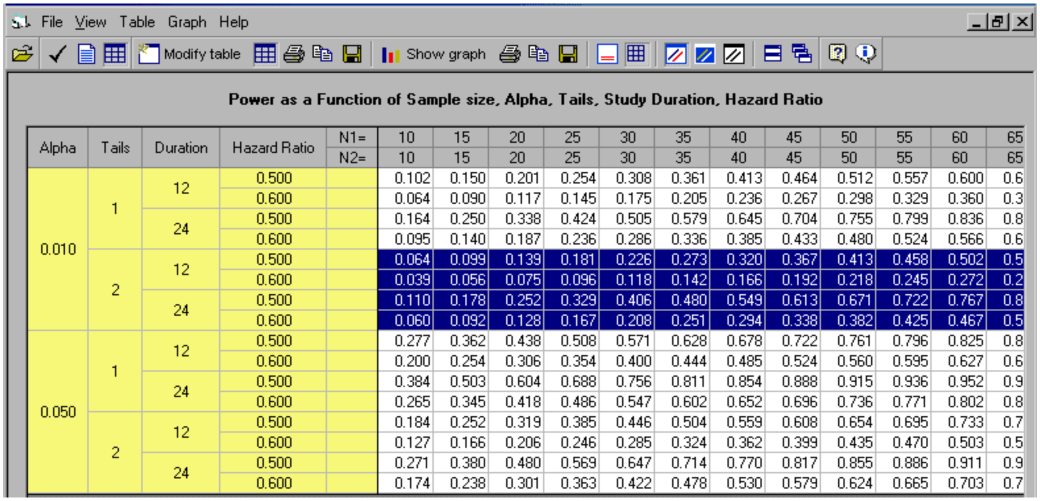
- ❖ Double click on 0.01 in the first column, and all graphs are created with *Alpha = 0.01*.
  - The next column is *Duration*, so there is a separate graph for each level of duration.
  - The right-most column is *Hazard Ratio*, so hazard ratio is the factor with graphs.

When the table includes four or more factors in addition to sample size:

In this example, we have added tails as an additional factor.

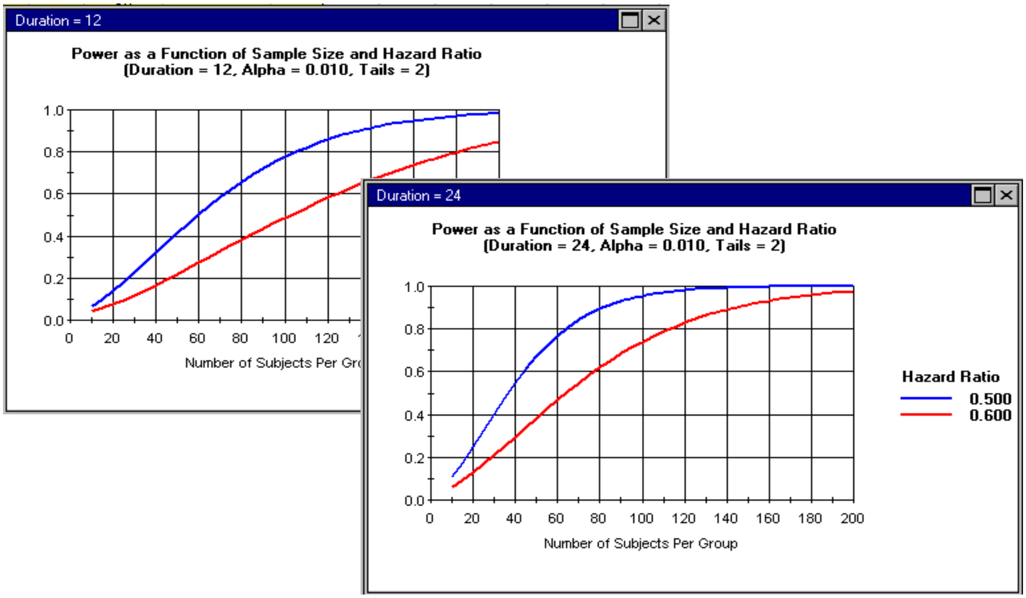
- ❖ Use drag and drop as above to modify the order of the columns.

- Double-click on the desired cell in the second column from the right (*Tails* in this example).
- With one click, you can select both *Alpha* and *Tails* (see below).
- Double-click *Tails = 2*. Note that this also selects *Alpha = 0.01* in the column on the left.



Power as a Function of Sample size, Alpha, Tails, Study Duration, Hazard Ratio

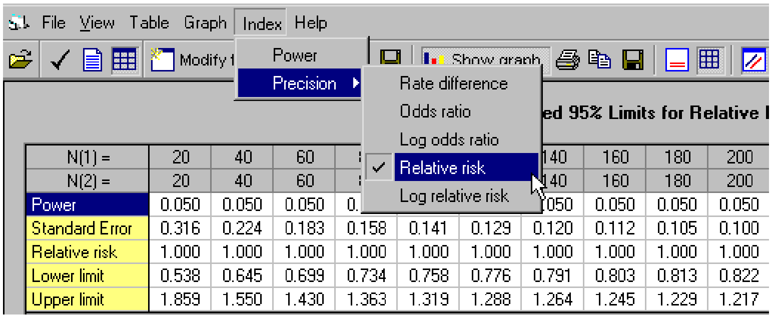
Alpha	Tails	Duration	Hazard Ratio	N1=	10	15	20	25	30	35	40	45	50	55	60	65
				N2=	10	15	20	25	30	35	40	45	50	55	60	65
0.010	1	12	0.500		0.102	0.150	0.201	0.254	0.308	0.361	0.413	0.464	0.512	0.557	0.600	0.6
			0.600		0.064	0.090	0.117	0.145	0.175	0.205	0.236	0.267	0.298	0.329	0.360	0.3
		24	0.500		0.164	0.250	0.338	0.424	0.505	0.579	0.645	0.704	0.755	0.799	0.836	0.8
			0.600		0.095	0.140	0.187	0.236	0.286	0.336	0.385	0.433	0.480	0.524	0.566	0.6
	2	12	0.500		0.064	0.099	0.139	0.181	0.226	0.273	0.320	0.367	0.413	0.458	0.502	0.5
			0.600		0.039	0.056	0.075	0.096	0.118	0.142	0.166	0.192	0.218	0.245	0.272	0.2
		24	0.500		0.110	0.178	0.252	0.329	0.406	0.480	0.549	0.613	0.671	0.722	0.767	0.8
			0.600		0.060	0.092	0.128	0.167	0.208	0.251	0.294	0.338	0.382	0.425	0.467	0.5
0.050	1	12	0.500		0.277	0.362	0.438	0.508	0.571	0.628	0.678	0.722	0.761	0.796	0.825	0.8
			0.600		0.200	0.254	0.306	0.354	0.400	0.444	0.485	0.524	0.560	0.595	0.627	0.6
		24	0.500		0.384	0.503	0.604	0.688	0.756	0.811	0.854	0.888	0.915	0.936	0.952	0.9
			0.600		0.265	0.345	0.418	0.486	0.547	0.602	0.652	0.696	0.736	0.771	0.802	0.8
	2	12	0.500		0.184	0.252	0.319	0.385	0.446	0.504	0.559	0.608	0.654	0.695	0.733	0.7
			0.600		0.127	0.166	0.206	0.246	0.285	0.324	0.362	0.399	0.435	0.470	0.503	0.5
		24	0.500		0.271	0.380	0.480	0.569	0.647	0.714	0.770	0.817	0.855	0.886	0.911	0.9
			0.600		0.174	0.238	0.301	0.363	0.422	0.478	0.530	0.579	0.624	0.665	0.703	0.7



Graphs of Precision

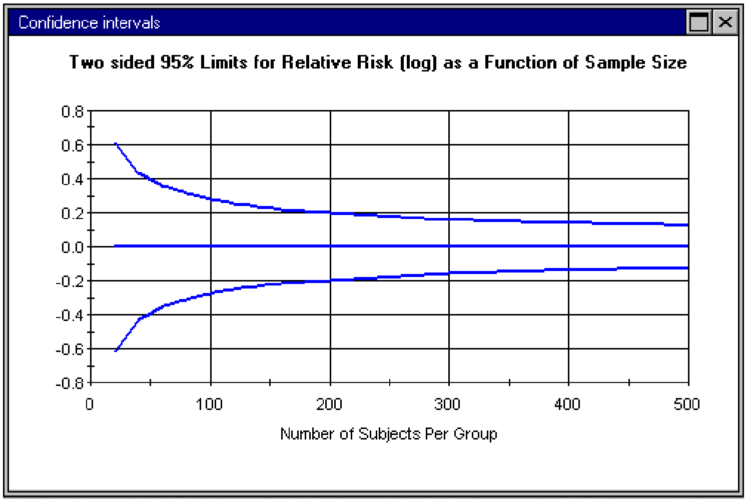
For certain procedures, the program will display precision (the expected width of the confidence interval) as well as power.

When this option is available, the program displays an additional button on the toolbar. Click this button to display the graph of precision.



N(1) =	20	40	60	!			140	160	180	200
N(2) =	20	40	60	!			140	160	180	200
Power	0.050	0.050	0.050	0.			050	0.050	0.050	0.050
Standard Error	0.316	0.224	0.183	0.158	0.141	0.129	0.120	0.112	0.105	0.100
Relative risk	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000	1.000
Lower limit	0.538	0.645	0.699	0.734	0.758	0.776	0.791	0.803	0.813	0.822
Upper limit	1.859	1.550	1.430	1.363	1.319	1.288	1.264	1.245	1.229	1.217

This example is for a study that will examine the rate difference in two independent groups.

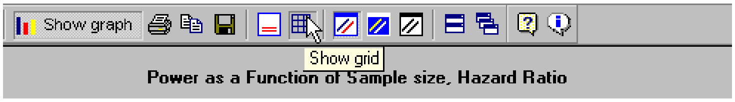


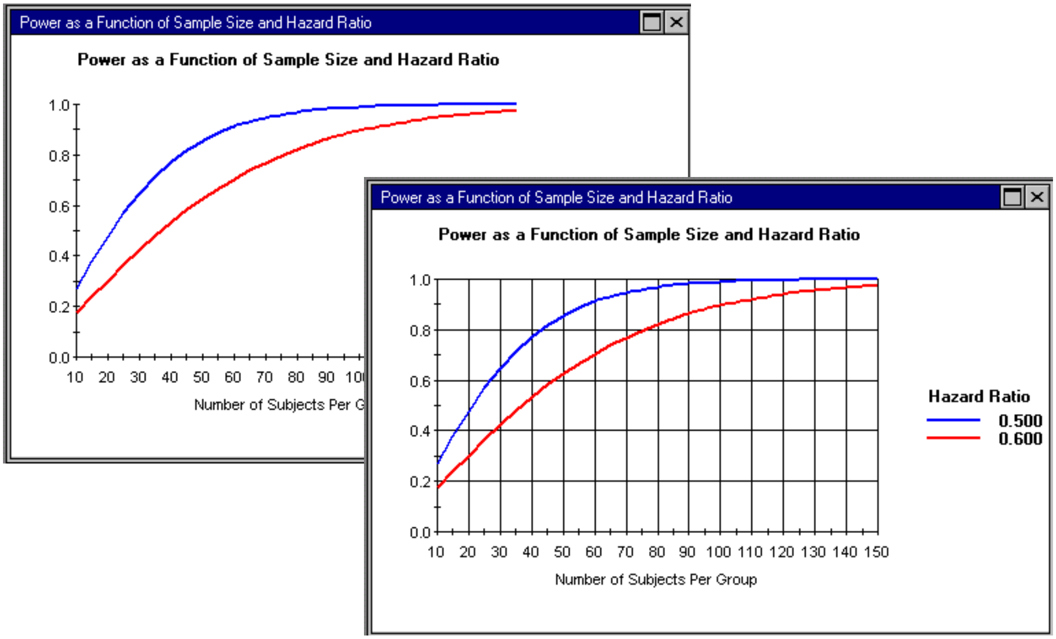
Customizing Graphs

Gridlines

- To add or remove gridlines from a graph, click *Show grid* on the toolbar or on the Graph menu.

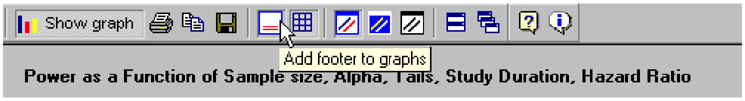
Figure 4.1 Graphs with the gridlines hidden (left) and visible (right).




Power as a Function of Sample Size and Hazard Ratio
Power as a Function of Sample Size and Hazard Ratio
1.0
0.8
0.6
0.4
0.2
0.0
10 20 30 40 50 60 70 80 90 10
Number of Subjects Per G
Power as a Function of Sample Size and Hazard Ratio
Power as a Function of Sample Size and Hazard Ratio
1.0
0.8
0.6
0.4
0.2
0.0
10 20 30 40 50 60 70 80 90 100 110 120 130 140 150
Number of Subjects Per Group
Hazard Ratio
0.500
0.600

Headers and Footers

- To add or remove footers from a graph, click *Add footer to graphs* on the toolbar or on the Graph menu.



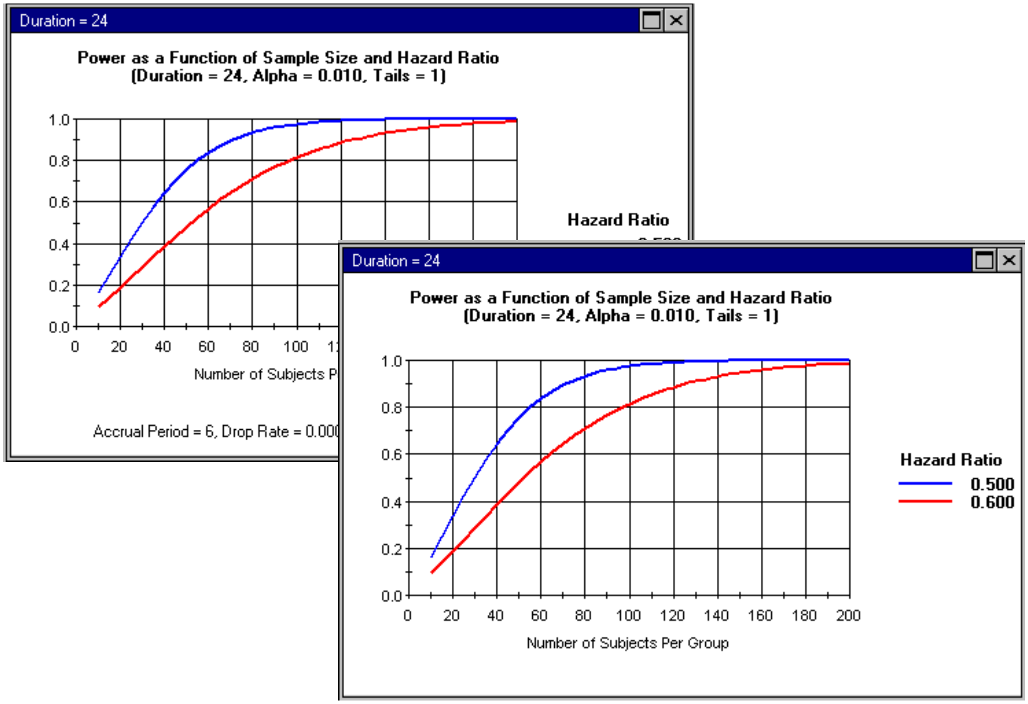
Power as a Function of Sample size, Alpha, Tails, Study Duration, Hazard Ratio

The first line in the graph title lists the factors that vary in the graph (sample size and hazard ratio).

The second line lists the factors that vary in the table but are constant in the graph (duration, alpha, and tails).

The graph footer lists the factors that are constant (accrual period, drop rate, and hazard rate).

Figure 4.2 Graphs with footer (left) and without (right)

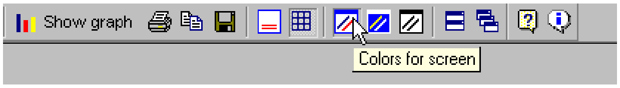

Duration = 24
Power as a Function of Sample Size and Hazard Ratio
(Duration = 24, Alpha = 0.010, Tails = 1)
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
1
Number of Subjects P
Hazard Ratio
Accrual Period = 6, Drop Rate = 0.00
Duration = 24
Power as a Function of Sample Size and Hazard Ratio
(Duration = 24, Alpha = 0.010, Tails = 1)
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
120
140
160
180
200
Number of Subjects Per Group
Hazard Ratio
0.500
0.600

Tip

The footer will be identical for all graphs. For this reason, when studying graphs on-screen or printing a group of graphs on one page, it is best to exclude the footer from the graphs and print it once for the full page. When you copy all graphs to the clipboard, the footer is copied once for the full set and is pasted into the word processing program below the graphs. When you want to create a graph that will be used independently (such as on a slide), you should include the footer as a part of the graph itself.

Colors

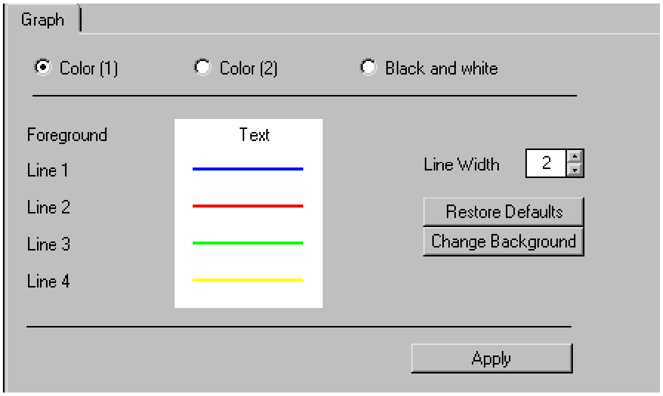
❖ To modify graph colors, click *Colors for screen* on the toolbar or on the Graph menu.



The program includes three sets of colors:

- Colors for screen (white background)
- Colors for slides (dark background)
- Black and white (for printing), with symbols used to distinguish between lines

Each of the schemes may be modified extensively. Click *Customize graph* on the Graph menu.



● Color (1) ○ Color (2) ○ Black and white

Foreground: Text
Line 1
Line 2
Line 3
Line 4

Line Width: 2

You can customize the following schemes:

- The color used for the background, foreground, and each line.
- The line width.
- The symbol width.

- The symbol size.
- To modify any of the colors, click on that color.
- Changes are saved automatically and applied whenever that scheme is selected.

Printing, Saving, and Exporting Tables

Printing Tables

❖ Click *Print table* on the toolbar or on the Table menu.

The program will print the entire table (and is not limited to the rows or columns that are displayed at the time). The panels that display above and below the table on the screen will be incorporated into the table automatically. If needed, the print will continue across multiple pages.

Copying Tables to the Clipboard As Bitmaps

❖ Click *Copy to clipboard as bitmap* on the toolbar or on the Table menu.

The program will copy the entire table (and is not limited to the rows or columns that are displayed at the time). The panels that display above and below the table on the screen will be incorporated into the table automatically.

If the table includes many columns, the bitmap will be difficult to read when pasted into a word processing program and compressed to fit the page size. Therefore, you may want to modify the sample size on the table to yield a smaller number of columns. Change the start, increment, or end values and then recreate the table.

Copying Tables to the Clipboard As Data

❖ Click *Copy to clipboard as data* on the toolbar or on the Table menu.

This data can then be pasted into a spreadsheet for additional manipulation.

The data includes many columns that are not visible on the screen. For example, if alpha is constant at 0.05, then onscreen alpha is excluded from the table and displayed below the table instead. When the data is copied to the clipboard, however, there will be a column labeled *Alpha*, with 0.05 shown for every row. In Excel or other programs, you may elect to hide these columns.

Saving Tables to Disk

- Choose *Save table* from the File menu.

The program saves the current table (not the underlying data). To open the table at a later time, you must be in the tables module for the same statistical procedure. When the table is opened, you will be able to redisplay the graphs and manipulate the columns immediately, but if you want to modify any values, you will need to reenter those values.

Tip

A table that was created for a particular statistical procedure should be retrieved only from the same statistical procedure. That is, if you saved a table for survival with attrition, then select this same procedure before proceeding to the Table window to retrieve that table.

Printing, Saving, and Exporting Graphs

Saving a Single Graph

To save a single graph to a file, right-click on the graph and choose *Save graph to disk*. To select the file type, use the drop-down list for file extensions. The options are bitmap (*.bmp*), Windows metafile (*.wmf*), and enhanced Windows metafile (*.emf*). In general, the Windows metafile (*.wmf*) format will provide the best results when the file is imported into other programs, such as Word or PowerPoint. When saving a single graph, you should include the footer (click *Add footer to graph*).

Saving Multiple Graphs

When there are multiple graphs, you can save them by right-clicking on each one and specifying a name. Alternatively, choose *Save all graphs* from the File menu and specify a name. The program will save all of the graphs, appending a suffix to the name. For example *myfile.wmf* will be saved as *myfile-01.wmf*, *myfile-02.wmf*, and so on. Note that this method will *not* check before overwriting files with the same name.

Tip

When the graph is saved to a file, only the graph itself is saved. The panel that appears at the bottom of the page is not saved with the file. To include this information on the graph, include the footer (click *Add footer to graph*).

To import the file into Word:

- ❖ If the graph was saved in *.wmf* format or in *.emf* format, choose *Insert > Picture > From File*.
- ❖ If the file was saved in *.bmp* format, choose *Insert > Object > Create from File*.

Copying Graphs to the Clipboard

Copying a Single Graph to the Clipboard

To copy a single graph to a file, right-click on that graph and choose *Copy graph to clipboard*. The file is copied as a bitmap. Paste the graph into the other program and (optionally) change its location on the page.

Copying Multiple Graphs to the Clipboard

When there are multiple graphs, you can copy them by right-clicking on each one, copying to the clipboard, and pasting into another program. Alternatively, choose *Copy all graphs to clipboard* from the File menu. The program will copy all of the graphs. Paste the graphs into the other program and (optionally) change their locations on the page.

Tips

- When one or more graphs are copied to the clipboard, the panel that appears at the bottom of the page is also copied.
- The bitmaps can be resized in the other program. However, if the quality of the image is important (such as on a slide), you should try saving the graph to disk as a *.wmf* file and then importing it into the other program.

Printing Graphs

Printing a Single Graph

To print a single graph to a file, right-click on that graph and choose *Print graph*. Before printing, you can modify the size of the graph on-screen by dragging the sides or corners of the graph itself, and the print will correspond to the on-screen size.

Printing Multiple Graphs

When there are multiple graphs, you can print them by right-clicking on each one and choosing *Print graph*. Alternatively, choose *Print all graphs* from the File menu. The program will print one to two graphs on the first page and will use additional pages as needed.

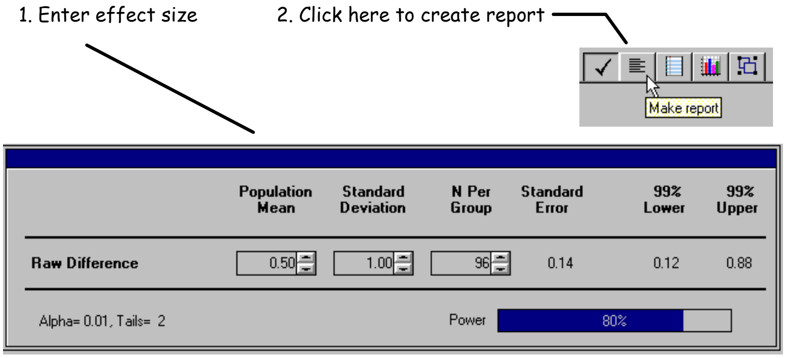
Tip

When one graph is printed, the panel that appears at the bottom of the page is also printed. When multiple graphs are printed, the panel is printed only once for the full set.

5 Reports

The program can generate a report that summarizes the results of the power and precision analysis in a format that is compatible with most Windows word processing programs.

Figure 5.1 Creating a report



1. Enter effect size

2. Click here to create report

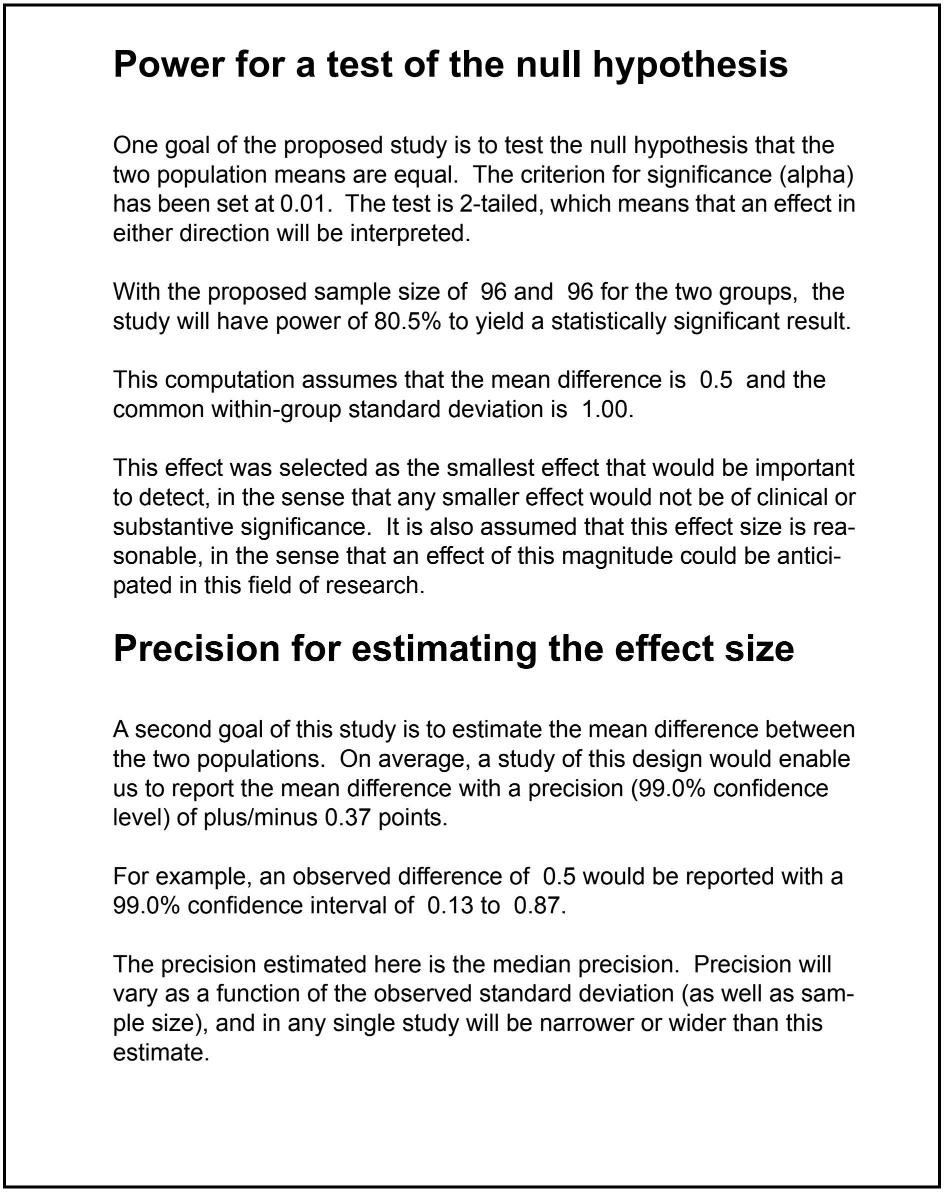
	Population Mean	Standard Deviation	N Per Group	Standard Error	99% Lower	99% Upper
Raw Difference	0.50	1.00	96	0.14	0.12	0.88

Alpha= 0.01, Tails= 2

Power 80%

- On the main screen, enter an effect size.
- Optionally, modify the group names, alpha, tails, confidence level, and sample size.
- Modify the study parameters until power and/or precision are acceptable.
- Click the *Make report* icon on the toolbar, or from the Tools menu choose *Generate text report*.
- You can edit the report using standard word processing procedures.
- The report is created in rich text format (RTF), which is compatible with most word processing programs.

If you generate additional reports, the new reports will be added to the end of the file. Figure 5.2 shows sample reports.



Power for a test of the null hypothesis

One goal of the proposed study is to test the null hypothesis that the two population means are equal. The criterion for significance (alpha) has been set at 0.01. The test is 2-tailed, which means that an effect in either direction will be interpreted.

With the proposed sample size of 96 and 96 for the two groups, the study will have power of 80.5% to yield a statistically significant result.

This computation assumes that the mean difference is 0.5 and the common within-group standard deviation is 1.00.

This effect was selected as the smallest effect that would be important to detect, in the sense that any smaller effect would not be of clinical or substantive significance. It is also assumed that this effect size is reasonable, in the sense that an effect of this magnitude could be anticipated in this field of research.

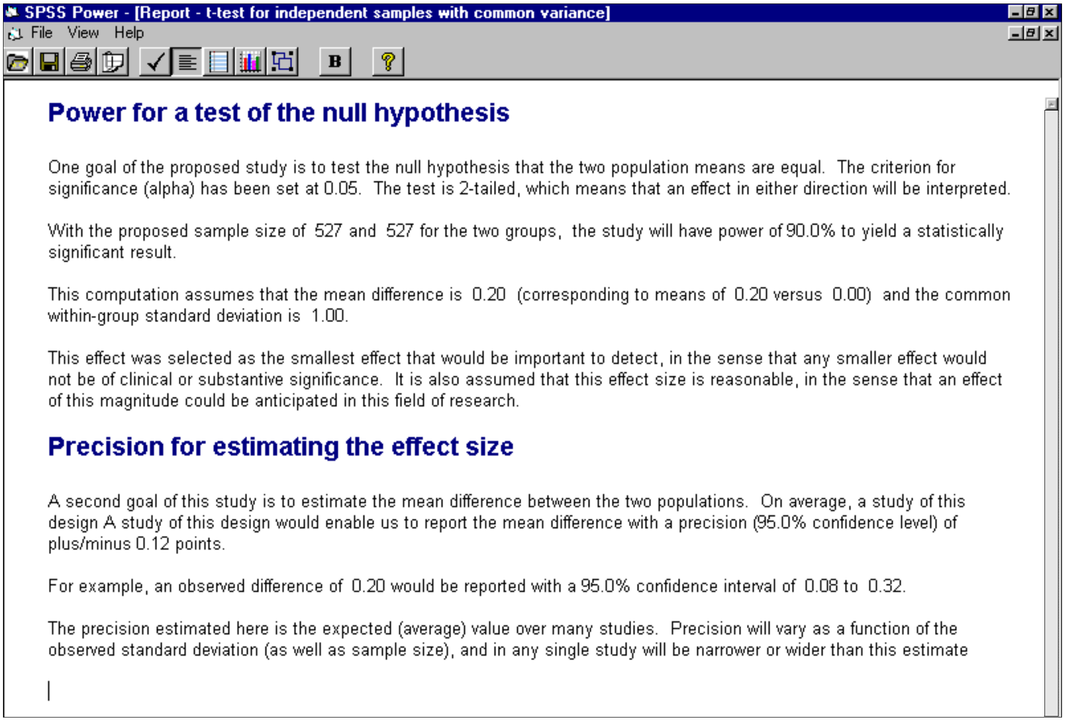
Precision for estimating the effect size

A second goal of this study is to estimate the mean difference between the two populations. On average, a study of this design would enable us to report the mean difference with a precision (99.0% confidence level) of plus/minus 0.37 points.

For example, an observed difference of 0.5 would be reported with a 99.0% confidence interval of 0.13 to 0.87.

The precision estimated here is the median precision. Precision will vary as a function of the observed standard deviation (as well as sample size), and in any single study will be narrower or wider than this estimate.

Figure 5.2 Sample reports



SPSS Power - [Report - t-test for independent samples with common variance]
File View Help
B

Power for a test of the null hypothesis

One goal of the proposed study is to test the null hypothesis that the two population means are equal. The criterion for significance (alpha) has been set at 0.05. The test is 2-tailed, which means that an effect in either direction will be interpreted.

With the proposed sample size of 527 and 527 for the two groups, the study will have power of 90.0% to yield a statistically significant result.

This computation assumes that the mean difference is 0.20 (corresponding to means of 0.20 versus 0.00) and the common within-group standard deviation is 1.00.

This effect was selected as the smallest effect that would be important to detect, in the sense that any smaller effect would not be of clinical or substantive significance. It is also assumed that this effect size is reasonable, in the sense that an effect of this magnitude could be anticipated in this field of research.

Precision for estimating the effect size

A second goal of this study is to estimate the mean difference between the two populations. On average, a study of this design A study of this design would enable us to report the mean difference with a precision (95.0% confidence level) of plus/minus 0.12 points.

For example, an observed difference of 0.20 would be reported with a 95.0% confidence interval of 0.08 to 0.32.

The precision estimated here is the expected (average) value over many studies. Precision will vary as a function of the observed standard deviation (as well as sample size), and in any single study will be narrower or wider than this estimate

Later chapters in this manual include sample reports generated for each statistical procedure.

Tip

The program allows the user to incorporate citations to all analyses referenced in the report (choose *Add references* from the Edit menu).

Printing Reports

Click the *Print* icon, or from the File menu choose *Print*. On the *Forms* tab, select *Report* (this option will be enabled only if a report has been created).

Copying Reports to the Clipboard

Highlight all or part of the report (if none of the report is selected, the entire report will be copied). Then click the *Copy to clipboard* icon, or from the File menu choose *Clipboard*. Switch to a word processing program and paste the report into a file.

Saving Reports

Choose *Save report to disk* from the File menu. The report will be saved in RTF format and can later be imported into a word processing program.

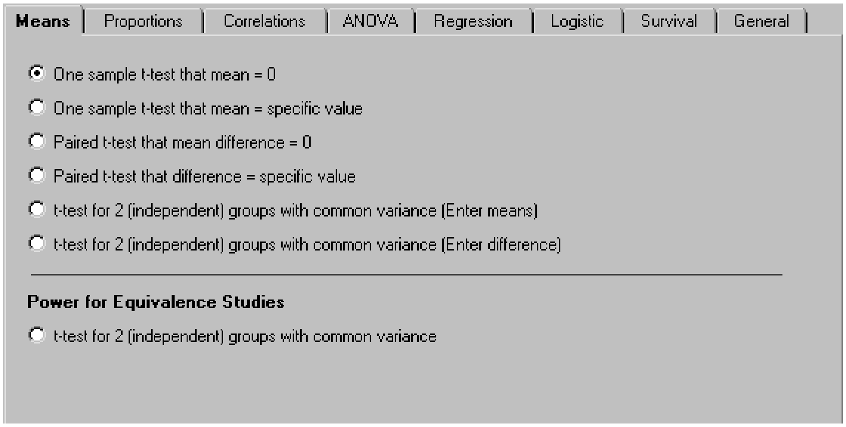
Tip

Do not save to disk using the *Save file* icon or the *Save study data* option on the File menu. These functions save the study parameters rather than the report.

6 T-Test for One Group

Selecting the Procedure

To display the available procedures, choose *New analysis* from the File menu.



● One sample t-test that mean = 0

○ One sample t-test that mean = specific value

○ Paired t-test that mean difference = 0

○ Paired t-test that difference = specific value

○ t-test for 2 (independent) groups with common variance (Enter means)

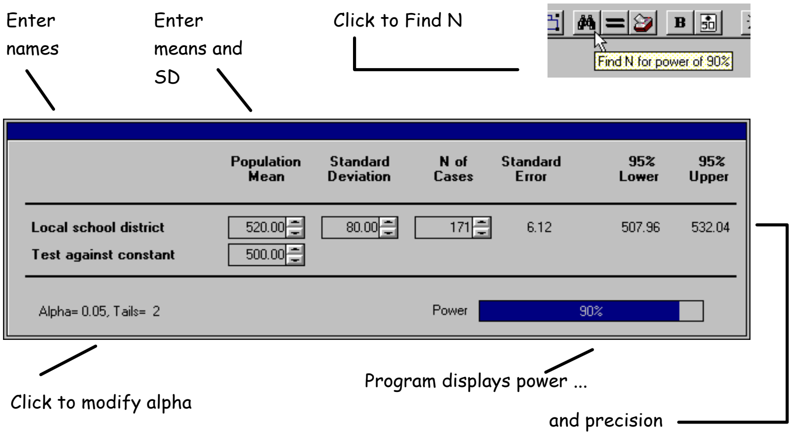
○ t-test for 2 (independent) groups with common variance (Enter difference)

Power for Equivalence Studies

○ t-test for 2 (independent) groups with common variance

Application

Figure 6.1 One sample t-test



Enter names

Enter means and SD

Click to Find N

	Population Mean	Standard Deviation	N of Cases	Standard Error	95% Lower	95% Upper
Local school district	520.00	80.00	171	6.12	507.96	532.04
Test against constant	500.00					

Alpha= 0.05, Tails= 2

Power 90%

Click to modify alpha

Program displays power ...

and precision

The program includes two versions of the one-group t-test. The first assumes that the mean will be tested against the nil value (0). The second allows the user to test the mean against any specific (null) value.

Effect Size

The effect size for t-tests is the standard difference, d, defined as the mean difference divided by the standard deviation.

In theory, the effect size index extends from 0 (indicating no effect) to infinity. In practice, of course, d is limited to a substantially smaller range, as reflected in the conventions suggested by Cohen for research in the social sciences—small (0.20), medium (0.50), and large (0.80).

The program requires the user to enter the mean and standard deviation (SD) for the population to be tested. If the mean will be tested against a mean of 0, no additional data are required for the effect size. If the mean will be tested against a specific (nonzero) value, the researcher must supply this value as well.

Tip

To enter the standard difference (d) directly, set the program to enter the raw difference and provide a value of 1.0 for the standard deviation. In this case, the raw difference is equal to d.

Alpha, Confidence Level, and Tails

Click *Alpha*, *Tails*, or the confidence level to modify these values. The value set for tails applies to both the power analysis and the confidence interval.

Sample Size

The spin control adjacent to the N of Cases value can be used to modify the sample size quickly. The size of the increment can be specified by the user (click *N of Cases*).

Click the *Find N* icon to have the program find the number of cases required for the default level of power. The default value for power is 80%, but this can be modified temporarily (Ctrl-F) or permanently (choose *Preferences* from the Options menu).

Tolerance Interval

The confidence interval displayed by the program is the median interval and is shown balanced about the expected mean. For example, assume that the user enters an expected mean of 100, with $\text{SD} = 10$ and $\text{N} = 40$. The program displays the 95% confidence interval as 96.83 to 103.17 (which is equal to the mean plus or minus 3.17).

Choose *Tolerance intervals* from the View menu. The program shows that the 50% tolerance interval for the 95% confidence interval is 6.34 (corresponding to the distance from 96.83 to 103.17), which means that in 50% of the studies, the 95% confidence interval will be no wider than 6.34 points.

Additionally, it shows that the 80% tolerance interval is 6.96, which means that in 80% of all studies, the 95% confidence interval will be no wider than 6.96.

To modify the confidence level, click on the value shown, or choose *Alpha/Tails* from the Options menu. To modify the tolerance level, change the value shown in the tolerance interval box.

Because a one-tailed “interval” actually extends from minus infinity to some value above the mean, or from some value below the mean to plus infinity, it is not useful here. We could, however, consider the distance from the observed mean to the single boundary. If the user selects a one-tailed confidence interval, the median distance from the observed mean to the single boundary will be no greater than 2.64 in 50% of samples and no greater than 2.90 in 80% of samples.

As the sample size increases, not only are we able to estimate the mean more precisely (which yields a narrower confidence interval), but the dispersion of the sample standard deviation about the expected standard deviation decreases; this means that the width of the confidence interval shows less variation from sample to sample. For example, with $\text{N} = 200$, the 50% and 80% tolerance intervals are quite similar, at 2.78 and 2.90, respectively (assuming that $\text{SD} = 10.0$ and using two tails).

When the user specifies that the variance is known (choose *Computational formulas* from the Options menu), the confidence interval is computed using the population variance specified by the user. In this case, the confidence interval will not vary from sample to sample, and the 80% (or any other) tolerance interval is identical to the 50% tolerance interval.

Computational Options for Power

By default, the program assumes that the population variance is unknown (will be estimated based on the observed variance), which is normally the case. The program allows the user to specify that the population variance is known (technically, this would be a z-test rather than a t-test). In this case, the program will work with the z-distribution rather than the t-distribution for computing power and confidence intervals. In practice, this distinction will have a substantial impact only if the number of cases is quite small (under 20–30) and the effect size is large. With larger sample sizes, the impact on computations may not be evident.

To switch between the t-test and the z-test, choose *Computational formulas* from the Options menu.

This module applies exclusively to the t-test for one group. The formulas used are identical to those used by the paired t-test module, but the reports and tools are optimized for each module. The present module should *not* be used for the two-group t-test because the computational formulas are different.

Computational Options for Precision

The computational options for precision are identical to those for power. By default, the program works with the t-distribution (for estimated variance), but the user can choose to work with the z-distribution (for known variance). The option set for power is applied to precision as well.

Options for Screen Display

The program will display the standard difference, d. Choose *Customize screen* from the Options menu.

Example

A school administrator wants to know whether students in his school district are scoring better or worse than the national norm of 500 on the SAT. He decides that a difference of 20 points or more from this normative value would be important to detect. Based on historical data, he anticipates that the standard deviation of scores in his district is about 80 points.

The researcher sets alpha at 0.05 and would like to obtain power of 90%. He would also like to report the group mean with a precision (90% confidence interval) of plus or minus 15 points.

- Choose *One sample t-test that mean = specified value*.

For the following steps, see Figure 6.1:

- Enter the group names—*Local school district* and *Test against constant*.

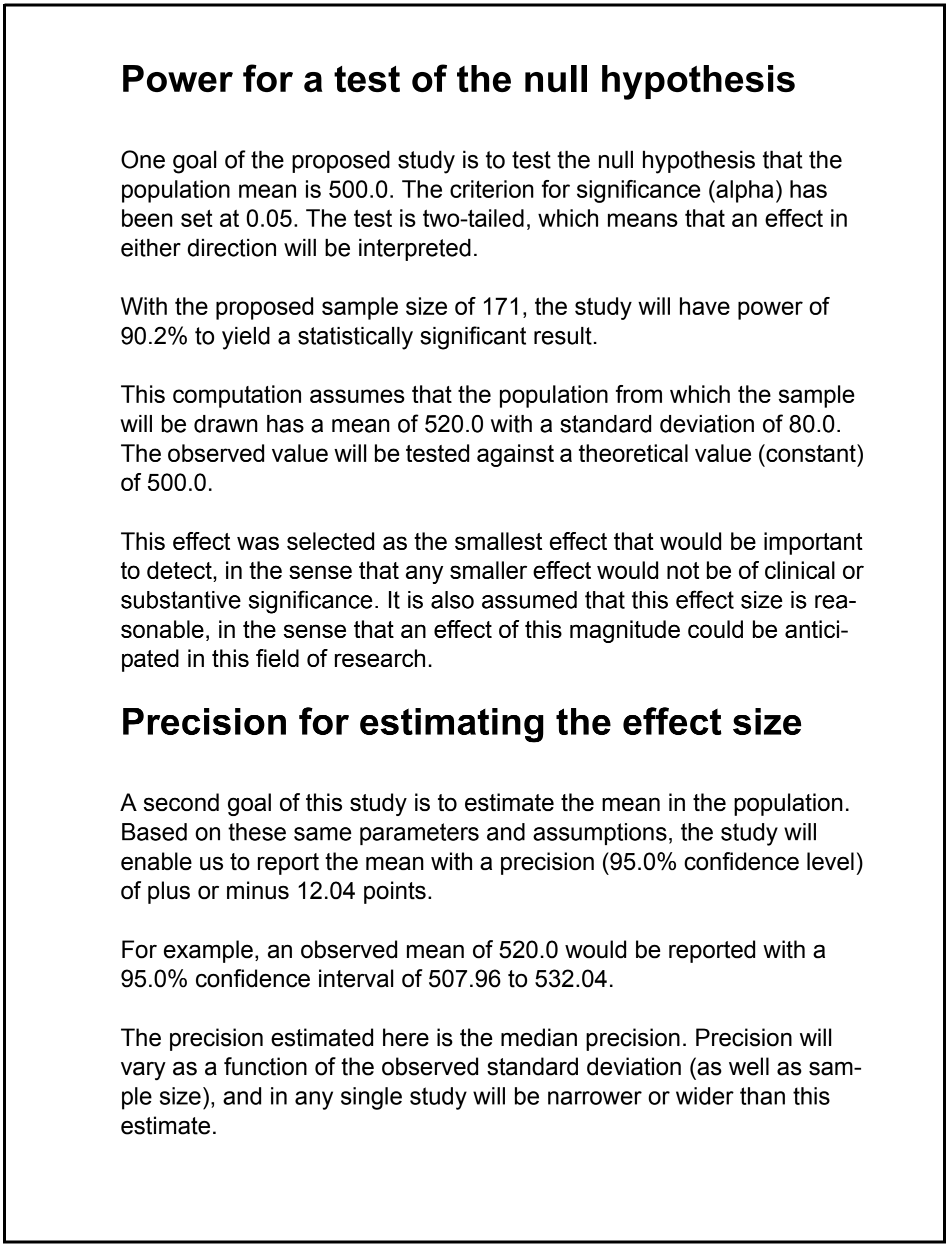
- Enter the effect size—a population mean of 520 versus a constant of 500, with a standard deviation of 80.

- Click *Alpha* and select the following values: *Alpha = 0.05*, *Confidence Level = 95%*, and *Tails = 2*.

- Press Ctrl-F (or click the *Find N* icon if the default power has already been set to 90). The program shows that a sample of 171 students will yield power of 90%.

- The program shows that an observed mean of 520 would be reported, with a 95% confidence interval of 507.95 to 532.05.

- Click the *Report* icon to generate the following report:



Power for a test of the null hypothesis

One goal of the proposed study is to test the null hypothesis that the population mean is 500.0. The criterion for significance (alpha) has been set at 0.05. The test is two-tailed, which means that an effect in either direction will be interpreted.

With the proposed sample size of 171, the study will have power of 90.2% to yield a statistically significant result.

This computation assumes that the population from which the sample will be drawn has a mean of 520.0 with a standard deviation of 80.0. The observed value will be tested against a theoretical value (constant) of 500.0.

This effect was selected as the smallest effect that would be important to detect, in the sense that any smaller effect would not be of clinical or substantive significance. It is also assumed that this effect size is reasonable, in the sense that an effect of this magnitude could be anticipated in this field of research.

Precision for estimating the effect size

A second goal of this study is to estimate the mean in the population. Based on these same parameters and assumptions, the study will enable us to report the mean with a precision (95.0% confidence level) of plus or minus 12.04 points.

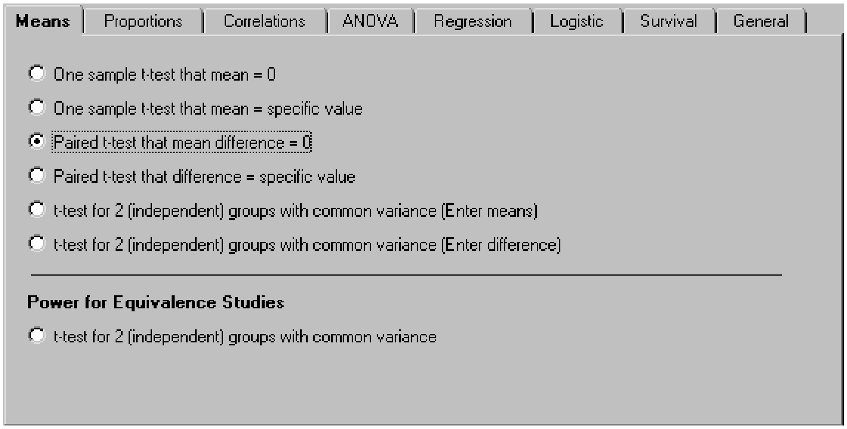
For example, an observed mean of 520.0 would be reported with a 95.0% confidence interval of 507.96 to 532.04.

The precision estimated here is the median precision. Precision will vary as a function of the observed standard deviation (as well as sample size), and in any single study will be narrower or wider than this estimate.

7 Paired T-Test

Selecting the Procedure

To display the available procedures, choose *New analysis* from the File menu.



○ One sample t-test that mean = 0

○ One sample t-test that mean = specific value

● Paired t-test that mean difference = 0

○ Paired t-test that difference = specific value

○ t-test for 2 (independent) groups with common variance (Enter means)

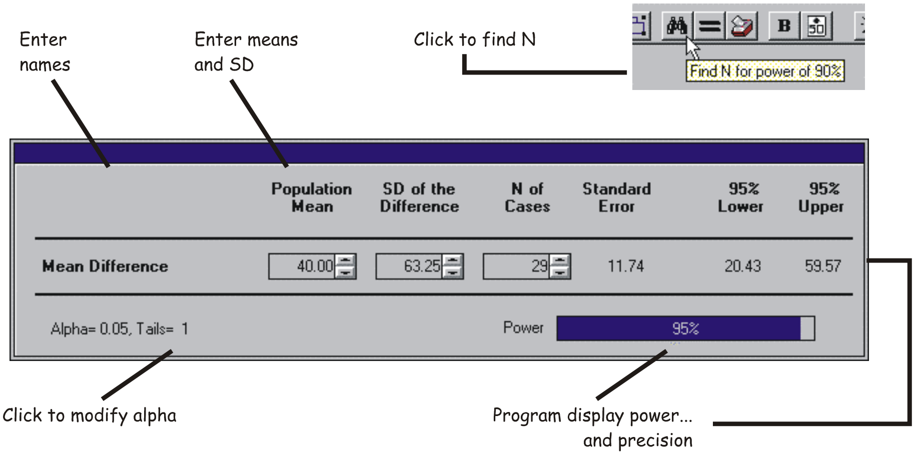
○ t-test for 2 (independent) groups with common variance (Enter difference)

Power for Equivalence Studies

○ t-test for 2 (independent) groups with common variance

Application

Figure 7.1 Paired t-test



Enter names

Enter means and SD

Click to find N

	Population Mean	SD of the Difference	N of Cases	Standard Error	95% Lower	95% Upper
Mean Difference	40.00	63.25	29	11.74	20.43	59.57

Alpha= 0.05, Tails= 1

Power 95%

Click to modify alpha

Program display power...
and precision

The paired t-test is used to compare two sets of scores when each score in one group is somehow linked with a score in the other group.

The test is commonly used to test a "difference" score, such as the change in a test score from one time point to the next. The paired t-test is also used in studies where persons in the two conditions are matched—for example, when a study enrolls pairs of siblings (one sibling assigned to each condition) or when patients are paired on the basis of disease severity before being assigned to either of two treatment plans.

It is always assumed that there will be a substantial positive correlation on the dependent variable between the two members of the pair. In other words, when we are testing a difference score, we assume that the pre-test and post-test are correlated. If this condition does not hold, it may be more appropriate (and powerful) to use a t-test for independent groups.

Effect Size

The effect size for t-tests is the standard difference, d, defined as the mean difference divided by the standard deviation of the difference.

In theory, the effect size index extends from 0 (indicating no effect) to infinity. In practice, of course, d is limited to a substantially smaller range, as reflected in the conventions suggested by Cohen for research in the social sciences—small (0.20), medium (0.50), and large (0.80).

Assume that patients are being entered into a study that will evaluate a treatment for cholesterol. Each patient's cholesterol level is assessed at baseline, the patient is put on a special regimen for six months, and the level is assessed again. The change from pre-treatment to post-treatment will be tested by a paired t-test. A clinically meaningful effect would be a drop of 40 points per patient over the treatment period.

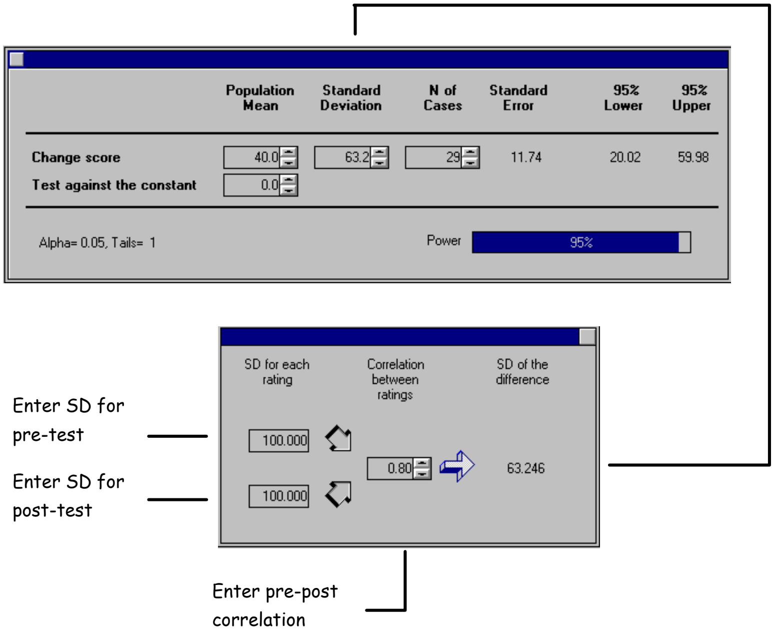
In the paired t-test, the effect size is computed using the standard deviation of the difference, rather than the standard deviation at the pre-test (or post-test).

This is an important distinction. Assume, for example, that the treatment works as expected—every patient's cholesterol level drops by some 40 points. John enters treatment with a level of 400 and leaves with a level of around 360. Paul enters treatment with a level of 280 and leaves with a level of 240. If we look at the group as a whole, the treatment effect is obscured by the overlap between the two time points. Specifically, John's level *after* treatment is substantially worse (360) than Paul's level *prior to* treatment (280). By contrast, if we look at the treatment's impact on a patient-by-patient basis, we see reductions that consistently fall within a fairly narrow range of 40 points. Clearly, the second perspective yields a more compelling case that the treatment is effective.

The distinction between these two perspectives is the distinction between the test of two independent groups (one assessed prior to treatment and the other assessed after treatment) and a single group followed over time (which is the subject of this chapter). This distinction is made by using the standard deviation of the difference, rather than the standard deviation of either time point, to compute the effect.

If the researcher is able to estimate the standard deviation of the difference score, the program will accept this value directly. In many cases, the user will be able to estimate the standard deviation of the individual scores and the correlation between pre-treatment and post-treatment scores but may not be able to estimate the standard deviation of the difference. The program will accept the two standard deviations and the correlation, and compute the standard deviation of the difference. In the current example, the researcher, based on historical data, assumes that the pre/post correlation is about 0.80.

Figure 7.2 Standard deviation assistant



	Population Mean	Standard Deviation	N of Cases	Standard Error	95% Lower	95% Upper
Change score	40.0	63.2	29	11.74	20.02	59.98
Test against the constant	0.0					

Alpha= 0.05, Tails= 1

Power 95%

SD for each rating	Correlation between ratings	SD of the difference
100.000 100.000	0.80	63.246

Enter SD for pre-test

Enter SD for post-test

Enter pre-post correlation

The program includes two versions of the paired t-test. The first assumes that the difference will be tested against the nil value (0). In the running example, the user would enter 40 (which will be tested against 0). The second allows the user to test the mean against any specific value. For the running example, the user could enter 260 versus 300.

Tip

To enter the standard deviation of the difference directly, close the assistant box. When the assistant box is open (the default), the program requires that you enter the standard deviation for each time point and the correlation between the two sets of scores. These data are used to compute the standard deviation of the difference, which is transferred to the main panel.

Alpha, Confidence Level, and Tails

Click *Alpha*, *Tails*, or the confidence level to modify these values. The value set for tails applies to both the power analysis and the confidence interval.

Sample Size

The spin control adjacent to the N of Cases value can be used to modify the sample size quickly. The size of the increment can be specified by the user (click *N of Cases*).

Click the *Find N* icon to have the program find the number of cases required for the default level of power. The default value for power is 80%, but this can be modified temporarily (Ctrl-F) or permanently (choose *Preferences* from the Options menu).

Tolerance Interval

The confidence interval displayed by the program is the median interval and is shown balanced about the expected mean. For example, assume that the user enters an expected mean difference of 0, with $\text{SD Difference} = 10$ (the standard deviation for each time point is 10 and the pre/post correlation is 0.50) and $\text{Number pairs} = 40$. The program displays the 95% confidence interval as –3.17 to +3.17.

Choose *Tolerance intervals* from the View menu. The program shows that the 50% tolerance interval for the 95% confidence interval is 6.34 (corresponding to the distance from –3.15 to +3.15), which means that in 50% of the studies, the 95% confidence interval will be no wider than 6.34 points.

Additionally, it shows that the 80% tolerance interval is 6.96, which means that in 80% of all studies, the 95% confidence interval will be no wider than 6.96.

To modify the confidence level, click on the value shown, or choose *Alpha/Tails* from the Options menu. To modify the tolerance level, change the value shown in the tolerance interval box.

Because a one-tailed "interval" actually extends from minus infinity to some value above the mean, or from some value below the mean to plus infinity, it is not useful here. We could, however, consider the distance from the observed mean to the single boundary. If the user selects a one-tailed confidence interval, the median distance from the observed mean to the single boundary will be no greater than 2.64 in 50% of samples and no greater than 2.90 in 80% of samples.

As the sample size increases, not only are we able to estimate the mean more precisely (which yield a narrower confidence interval), but the dispersion of the sample standard deviation about the expected standard deviation decreases; this means that the width of the confidence interval shows less variation from sample to sample. For example, with $\text{Number pairs} = 200$, the 50% and 80% tolerance intervals are quite similar, at 2.78 and 2.90, respectively (assuming that $\text{SD} = 10.0$).

When the user specifies that the variance is known (choose *Computational formulas* from the Options menu), the confidence interval is computed using the population variance specified by the user. In this case, the confidence interval will not vary from sample to sample, and the 80% (or any other) tolerance interval is identical to the 50% tolerance interval.

Computational Options for Power

By default, the program assumes that the population variance is unknown (will be estimated based on the observed variance), which is normally the case. The program allows the user to specify that the population variance is known (technically, this would be a z-test rather than a t-test). In this case, the program will work with the z-distribution rather than the t-distribution for computing power and confidence intervals. In practice, this distinction will have a substantial impact only if the number of cases is quite small (under 20–30) and the effect size is large. With larger sample sizes, the impact on computations may not be evident.

To switch between the t-test and the z-test, choose *Computational formulas* from the Options menu.

This module applies exclusively to the paired t-test. The formulas used are identical to those used by the one-sample t-test module, but the reports and tools are optimized for each module. The present module should *not* be used for the two-sample t-test because the computational formulas are different.

Computational Options for Precision

The computational options for precision are identical to those for power. By default, the program works with the t-distribution (for estimated variance), but the user can choose to work with the z-distribution (for known variance). The option set for power is applied to precision as well.

Options for Screen Display

The program will display the standard difference, d. Choose *Customize screen* from the Options menu.

Example

This illustration is a continuation of the example used for effect size. The researcher will enroll patients whose cholesterol levels fall in the range of 200–450, have them modify their diets for a year, and then test their cholesterol levels again.

There is no logical reason for the diet to *increase* cholesterol levels. More to the point, the researcher's interest is limited to finding a *decrease* in levels. A finding that the diet increases levels would have the same substantive impact as a finding of no effect because either result would mean that the diet should not be adopted. On this basis, the researcher decides to use a one-tailed test.

Alpha will be set at 0.05 and the researcher wants the study to have power of 90%. Additionally, she wants to know the precision with which she will be able to report the effect size.

- Choose *Paired t-test that the mean difference = 0.*

For the following steps, see Figure 7.2:

- Enter the mean change—40 points.
- Enter the standard deviation for each time point (100) and the pre/post correlation (0.80). The program inserts the standard error of the difference (63.2).
- Click *Alpha* and select the following values: *Alpha = 0.05*, *Confidence Level = 95%*, and *Tails = 1*.
- Press Ctrl-F (or click the *Find N* icon if the default power has already been set to 95). The program shows that a sample of 29 patients will yield power of 95%.
- The program shows that an observed change of 40 points would be reported, with a 95% confidence interval of 20.26 to 59.74.
- Click the *Report* icon to generate the following report:

Power for a test of the null hypothesis

One goal of the proposed study is to test the null hypothesis that the population mean change is 0.0. The criterion for significance (alpha) has been set at 0.05. The test is one-tailed, which means that only an effect in the expected direction will be interpreted.

With the proposed sample size of 29, the study will have power of 95.3% to yield a statistically significant result.

This computation assumes that the population from which the sample will be drawn has a mean change of 40.0 with a standard deviation of 63.2. The observed value will be tested against a theoretical value (constant) of 0.0.

This effect was selected as the smallest effect that would be important to detect, in the sense that any smaller effect would not be of clinical or substantive significance. It is also assumed that this effect size is reasonable, in the sense that an effect of this magnitude could be anticipated in this field of research.

Precision for estimating the effect size

A second goal of this study is to estimate the mean change in the population. Based on these same parameters and assumptions, the study will enable us to report the mean change with a precision (95.0% confidence level) of plus or minus 19.74 points.

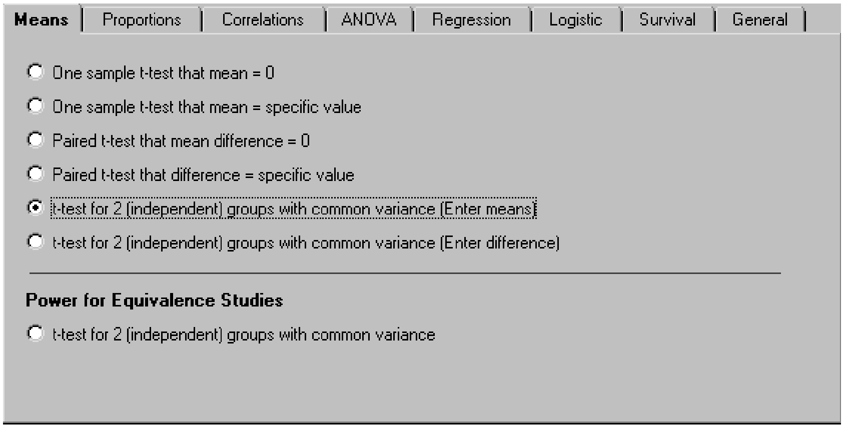
For example, an observed mean change of 40.0 would be reported with a 95.0% confidence interval of 20.43 to infinity or (alternatively, per the a priori hypothesis) minus infinity to 59.57. (Because the confidence interval has been defined as one-tailed, only one boundary is meaningful).

The precision estimated here is the median precision. Precision will vary as a function of the observed standard deviation (as well as sample size), and in any single study will be narrower or wider than this estimate.

8 T-Test for Independent Groups

Selecting the Procedure

To display the available procedures, choose *New analysis* from the File menu.



○ One sample t-test that mean = 0

○ One sample t-test that mean = specific value

○ Paired t-test that mean difference = 0

○ Paired t-test that difference = specific value

● t-test for 2 (independent) groups with common variance (Enter means)

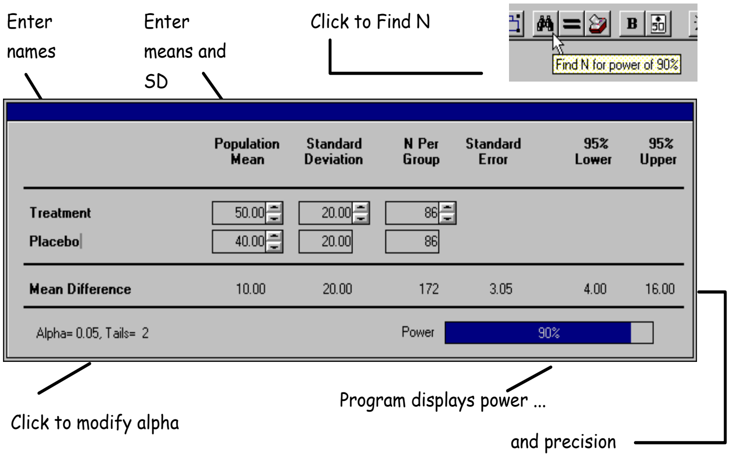
○ t-test for 2 (independent) groups with common variance (Enter difference)

Power for Equivalence Studies

○ t-test for 2 (independent) groups with common variance

Application

Figure 8.1 T-test for two independent groups



Enter names

Enter means and SD

Click to Find N

Find N for power of 90%

	Population Mean	Standard Deviation	N Per Group	Standard Error	95% Lower	95% Upper
Treatment	50.00	20.00	86			
Placebo	40.00	20.00	86			
Mean Difference	10.00	20.00	172	3.05	4.00	16.00

Alpha= 0.05, Tails= 2

Power 90%

Click to modify alpha

Program displays power ...

and precision

This procedure is used to test the mean difference in two independent groups. It is intended for a case in which the two groups share a common within-group standard deviation. The program allows the user to enter the mean difference directly or to enter the mean for each group (the program will display the mean difference).

Effect Size

The effect size for t-tests is the standard difference, d, defined as the mean difference between groups divided by the common within-group standard deviation.

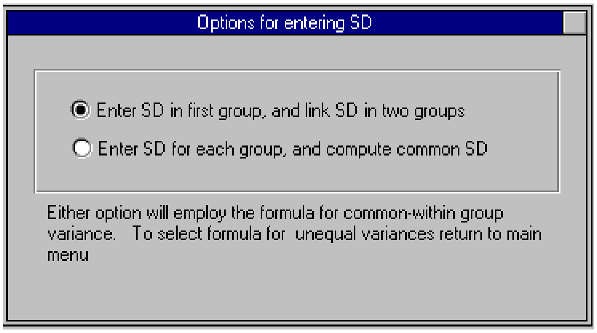
Assume, for example, that two populations have mean scores on the SAT of 550 versus 500 and that the standard deviation within either population is 100 points. The effect size $((550 - 500)/100\ )$ is 0.50.

In theory, the effect size index extends from 0 (indicating no effect) to infinity. In practice, of course, d is limited to a substantially smaller range, as reflected in the conventions suggested by Cohen for research in the social sciences—small (0.20), medium (0.50) and large (0.80).

The user can choose to enter the mean and standard deviation (SD) for each group, in which case the program will compute the mean difference (Mean1–Mean). Alternatively, the user can choose to enter the mean difference directly.

The program computes power for a t-test based on common within-groups standard deviations. Therefore, the user enters the standard deviation for the first group only, and this value is applied to (and displayed for) both groups.

The program will also allow the user to enter a separate standard deviation for each group (to activate the SD box for the second group, click on it). In this case, the program will compute the common within-group standard deviation using a weighted average of the two estimates and will display this value. This option is available only if the user is entering data for each group (rather than entering the raw difference).



● Enter SD in first group, and link SD in two groups

○ Enter SD for each group, and compute common SD

Either option will employ the formula for common-within group variance. To select formula for unequal variances return to main menu

Activate from
Options/
Data entry-
Study design

Note. These alternatives affect the mode of data entry only, and *not* the computational formula. Both alternatives assume a common within-group standard deviation. The program does not compute power for the t-test based on different population variances.

Tip

To enter the standard difference (d) directly, set the program to enter the raw difference and provide a value of 1.0 for the standard deviation. In this case, the raw difference is equal to d.

Alpha, Confidence Level, and Tails

Click *Alpha*, *Tails*, or the confidence level to modify these values. The value set for tails applies to both the power analysis and the confidence interval.

Sample Size

The program assumes that cases will be entered into the two groups in equal numbers. To modify this assumption, click *N Per Group*. The program allows the user to link the cases using some other assignment ratio (for example, 1:2 or 3:5). As long as the groups are linked, the user will enter a sample size for one group only, which facilitates the process of finding an appropriate sample size. The program will also allow the user to enter the number for each group independently of the other.

The spin control adjacent to the N Per Group value can be used to modify the sample size quickly. The size of the increment can be specified by the user (click *N Per Group*).

Click the *Find N* icon to have the program find the number per group required for the default level of power. The program will honor any assignment ratio (for example, 2:1) that has been specified by the user. The default value for power is 90%, but this can be modified temporarily (Ctrl-F) or permanently (choose *Preferences* from the Options menu).

Tolerance Interval

The confidence interval displayed by the program is the expected, or average, interval and is shown balanced about the expected mean. For example, assume that the user enters an expected mean difference of 0, with SD = 10 for each group and N = 40 per group. The program displays the 95% confidence interval as –4.42 to 4.42.

Choose *Tolerance intervals* from the View menu. The program shows that the 50% tolerance interval for the 95% confidence interval is 8.84 (corresponding to the distance from –4.42 to +4.42), which means that in 50% of the studies, the 95% confidence interval will be no wider than 8.84 points.

Additionally, it shows that the 80% tolerance interval is 9.44, which means that in 80% of all studies, the 95% confidence interval will be no wider than 9.44.

To modify the confidence level, click on the value shown, or select *Alpha/Tails* from the Options menu. To modify the tolerance level, change the value shown in the tolerance interval box.

Because a one-tailed “interval” actually extends from minus infinity to some value above the mean, or from some value below the mean to plus infinity, it is not useful here. We could, however, consider the distance from the observed mean to the single boundary. If the user selects a one-tailed confidence interval, the 50% interval is shown as 7.39 and the 80% tolerance level is shown as 7.89. The user could report that the expected distance from the observed mean to the single boundary will be no greater than 3.70 (that is, half of the 7.39 interval) in 50% of cases and no greater than 3.95 (that is, half of the 7.89 interval) in 80% of cases.

As the sample size increases, not only are we able to estimate the mean more precisely (which yields a narrower confidence interval), but the dispersion of the sample standard deviation about the expected standard deviation decreases; this means that the width of the confidence interval shows less variation from sample to sample. For example, with $N = 200$ per group, the 50% and 80% tolerance intervals are quite similar to each other, at 3.29 and 3.39, respectively (assuming that $SD = 10.0$).

When the user specifies that the variance is known (choose *Computational formulas* from the Options menu), the confidence interval is computed using the population variance specified by the user. In this case, the confidence interval will not vary from sample to sample, and the 80% (or any other) tolerance interval is identical to the 50% tolerance interval.

Computational Options for Power

By default, the program assumes that the population variance is unknown and will be estimated based on the observed variance, which is normally the case. The program allows the user to specify that the population variance is known; technically, this would be a z-test rather than a t-test. In this case, the program will work with the z-distribution rather than the t-distribution for computing power and confidence intervals. In practice, this distinction will have an impact only if the number per group is quite small (under 20–30) and the effect size is large. With larger sample sizes, the impact on computations may not be evident.

To switch between the t-test and the z-test, choose *Computational formulas* from the Options menu.

This program assumes that the populations share a common within-group standard deviation. The program does not compute power for the t-test based on different population variances.

Computational Options for Precision

The computational options for precision are identical to those for power. By default, the program works with the t-distribution (for estimated variance), but the user can choose to work with the z-distribution (for known variance). The option set for power is applied to precision as well.

Options for Screen Display

The program will display the standard difference, d. Choose *Customize screen* from the Options menu.

Example

Patients suffering from allergies will be assigned at random to one of two treatment conditions (treatment versus placebo) and asked to rate their comfort level on a scale of 0 to 100. The expected standard deviation within groups is 20 points. The researcher believes that a 10-point difference between groups is the smallest difference that would be important to detect.

The researcher sets alpha at 0.05 and would like to obtain power of 90%. She would also like to report the size of the effect with a precision (90% confidence interval) of plus or minus 5 points.

❖ Choose *t-test for 2 (independent) groups with common variance (Enter means).*

For the following steps, see Figure 8.1:

❖ Enter names for the two groups—*Treatment* and *Placebo*.

❖ Enter the mean for each group—50 and 40.

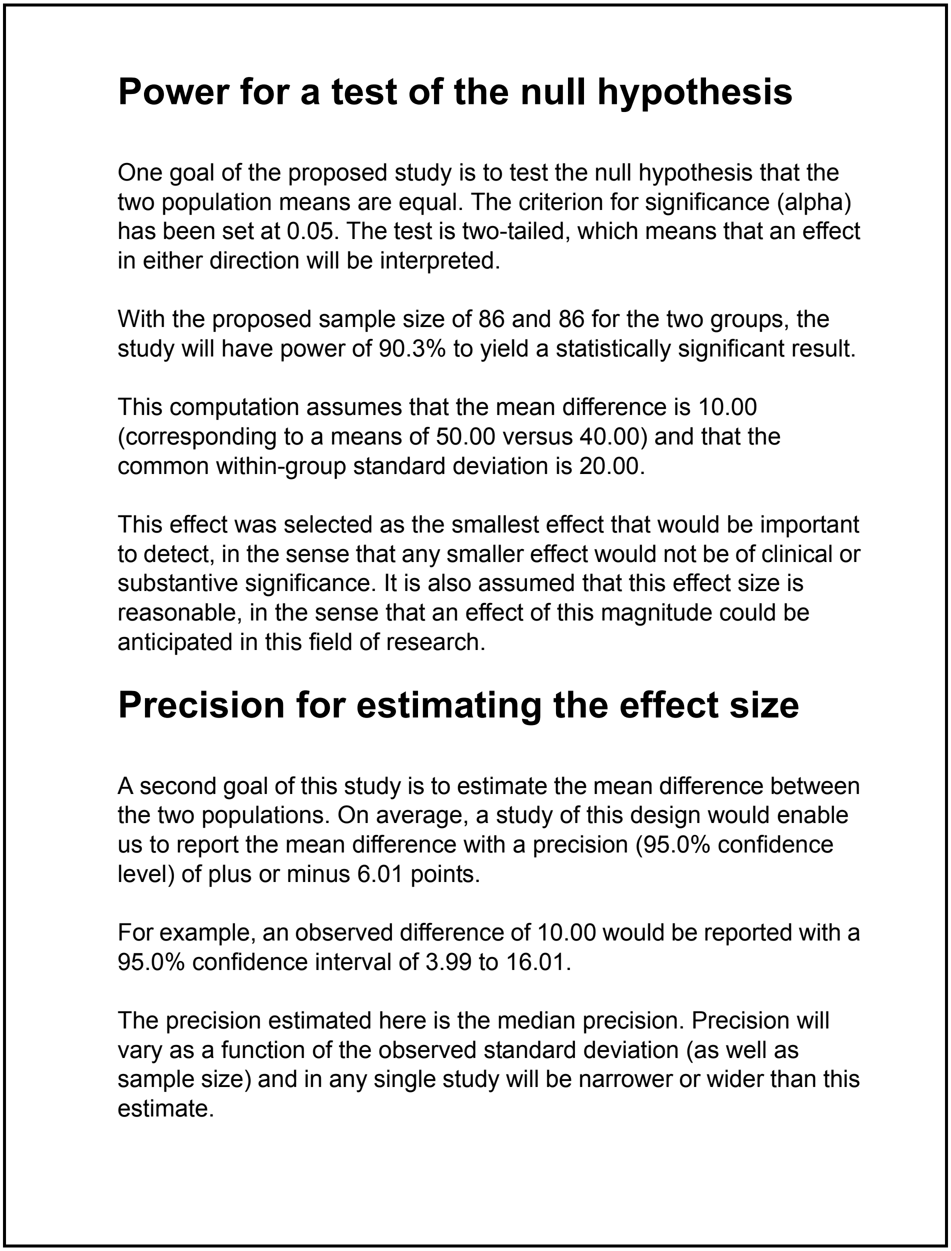
❖ Enter the standard deviation for the first group—20. This value is applied to the second group as well.

❖ Click *Alpha* and select the following values: *Alpha = 0.05*, *Confidence Level = 95%*, and *Tails = 2*.

❖ Press Ctrl-F (or click the *Find N* icon if the default power has already been set to 90). The program shows that a sample of 86 patients per group will yield power of 90%.

❖ The program shows that an observed difference of 10 points would be reported, with a 95% confidence interval of 3.99 to 16.01.

❖ Click the *Report* icon to generate the following report:



Power for a test of the null hypothesis

One goal of the proposed study is to test the null hypothesis that the two population means are equal. The criterion for significance (alpha) has been set at 0.05. The test is two-tailed, which means that an effect in either direction will be interpreted.

With the proposed sample size of 86 and 86 for the two groups, the study will have power of 90.3% to yield a statistically significant result.

This computation assumes that the mean difference is 10.00 (corresponding to a means of 50.00 versus 40.00) and that the common within-group standard deviation is 20.00.

This effect was selected as the smallest effect that would be important to detect, in the sense that any smaller effect would not be of clinical or substantive significance. It is also assumed that this effect size is reasonable, in the sense that an effect of this magnitude could be anticipated in this field of research.

Precision for estimating the effect size

A second goal of this study is to estimate the mean difference between the two populations. On average, a study of this design would enable us to report the mean difference with a precision (95.0% confidence level) of plus or minus 6.01 points.

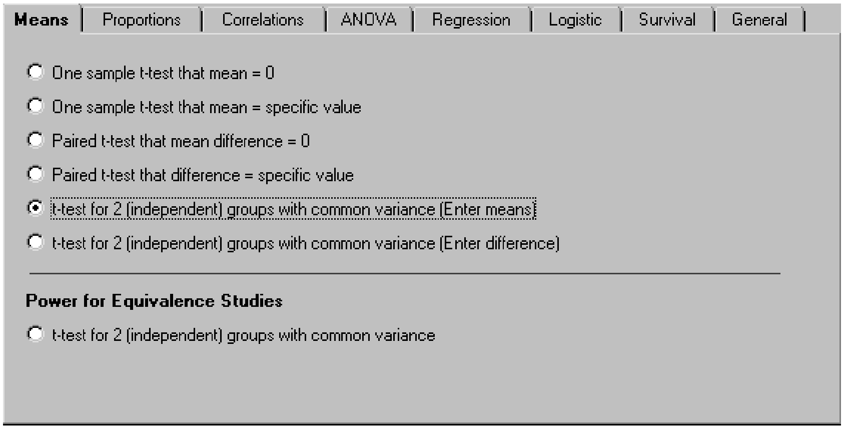
For example, an observed difference of 10.00 would be reported with a 95.0% confidence interval of 3.99 to 16.01.

The precision estimated here is the median precision. Precision will vary as a function of the observed standard deviation (as well as sample size) and in any single study will be narrower or wider than this estimate.

9 Proportions in One Sample

Selecting the Procedure

To display the available procedures, choose *New analysis* from the File menu.



○ One sample t-test that mean = 0

○ One sample t-test that mean = specific value

○ Paired t-test that mean difference = 0

○ Paired t-test that difference = specific value

● t-test for 2 (independent) groups with common variance (Enter means)

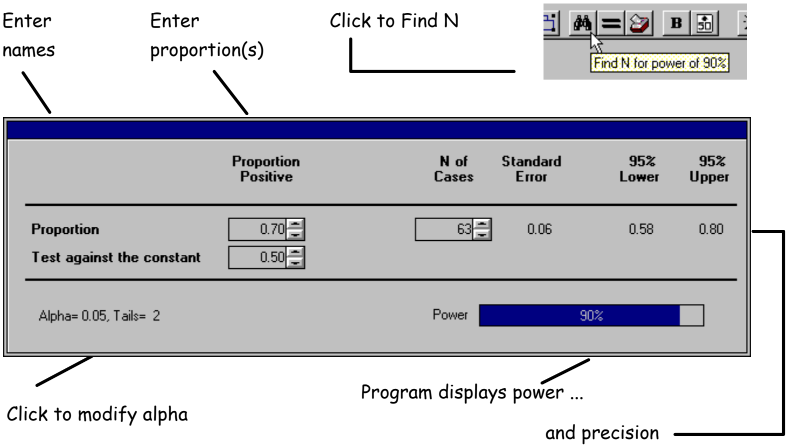
○ t-test for 2 (independent) groups with common variance (Enter difference)

Power for Equivalence Studies

○ t-test for 2 (independent) groups with common variance

Application

Figure 9.1 One-sample test that proportion equals 6



Enter names

Enter proportion(s)

Click to Find N

Find N for power of 90%

	Proportion Positive	N of Cases	Standard Error	95% Lower	95% Upper
Proportion	0.70	63	0.06	0.58	0.80
Test against the constant	0.50				

Alpha= 0.05, Tails= 2

Power 90%

Click to modify alpha

Program displays power ...

and precision

The program includes two versions of the one-sample test of proportion. The first version assumes that the difference will be tested against the null value of 50%. For example, patients are being treated with radiation, which carries the burden of serious side effects but the promise of a better long-term outcome. We want to survey patients after the fact and find out if the majority feel that the decision to have radiation was the correct one. (The null hypothesis is that 50% will respond positively and 50%, negatively.) A clinically important effect would be one that differed from 50% by 20 percentage points or more.

The second version allows the user to customize the null hypothesis. For example, we expect that 80% of patients are satisfied that they elected to have radiation, and we want to test the null hypothesis that 70% feel this way. In this case we would enter 0.80 for the proportion and 0.70 for the constant.

Effect Size

The effect size for the one-sample test of proportions is based on the difference between the two proportions. Unlike the t-test, where a difference of 10 versus 20 is equivalent to a difference of 40 versus 50 (that is, a 10-point difference in either case), when we work with proportions, the absolute values of the two proportions are relevant. Concretely, a difference of 10% versus 20% represents a more detectable effect than a difference of 40% versus 50%. (The variance of the effect size is larger, and the effect size is smaller, when we work with proportions near 0.50.)

For this reason, we refer to the effect by citing the rate difference followed by the actual values—for example, a 10-point difference (40% versus 50%).

The effect size reported in this way can range from 0 (indicating no effect) to 0.999. For research in the social sciences, Cohen has suggested the following conventional values—small (0.55 versus 0.50), medium (0.65 versus 0.50), and large (0.75 versus 0.50).

Alpha, Confidence Level, and Tails

Click *Alpha*, *Tails*, or the confidence level to modify these values. The value set for tails applies to both the power analysis and the confidence interval.

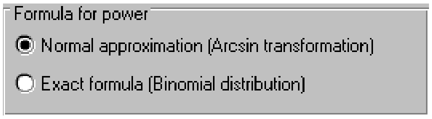
Sample Size

The spin control adjacent to the N of Cases value can be used to modify the sample size quickly. The size of the increment can be specified by the user (click *N of Cases*).

Click the *Find N* icon to have the program find the number of cases required for the default level of power. The default value for power is 90%, but this may be modified temporarily (Ctrl-F) or permanently (select *Preferences* from the Options menu).

Note. The Find N procedure is available only when the program is using the normal (arcsin) approximation to compute power.

Computational Options for Power



Formula for power

● Normal approximation (Arcsin transformation)

○ Exact formula (Binomial distribution)

The program offers two methods for computing power: the arcsin transformation, which is an approximation, and the exact formula, which employs the binomial distribution.

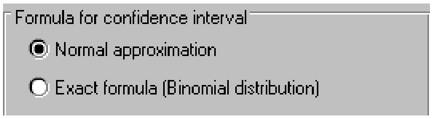
The binomial formula is recommended for small sample sizes (under 30), and the arcsin formula is recommended for larger samples (as the number of cases grows, the arcsin method approaches the accuracy of the exact method and has the advantage of being faster).

When the user chooses to work with the binomial test for power, the program displays the actual alpha for the test. This is always less than or equal to the alpha specified by the user, and it varies as a function of sample size.

To switch between computational methods, select *Computational formulas* from the Options menu.

Note. The Find N function is available only for the arcsin procedure. When the binomial formula is in effect, a small increase in sample size may result in a drop in power. This apparent anomaly is due to the fact that power is based on actual alpha rather than nominal alpha, and the increase in N may result in a more conservative actual alpha.

Computational Options for Precision



Formula for confidence interval

● Normal approximation

○ Exact formula (Binomial distribution)

The program offers two options for computing confidence intervals: a method based on the normal approximation and an exact method based on the binomial formula. To switch between computational methods, select *Computational formulas* from the Options menu.

Example

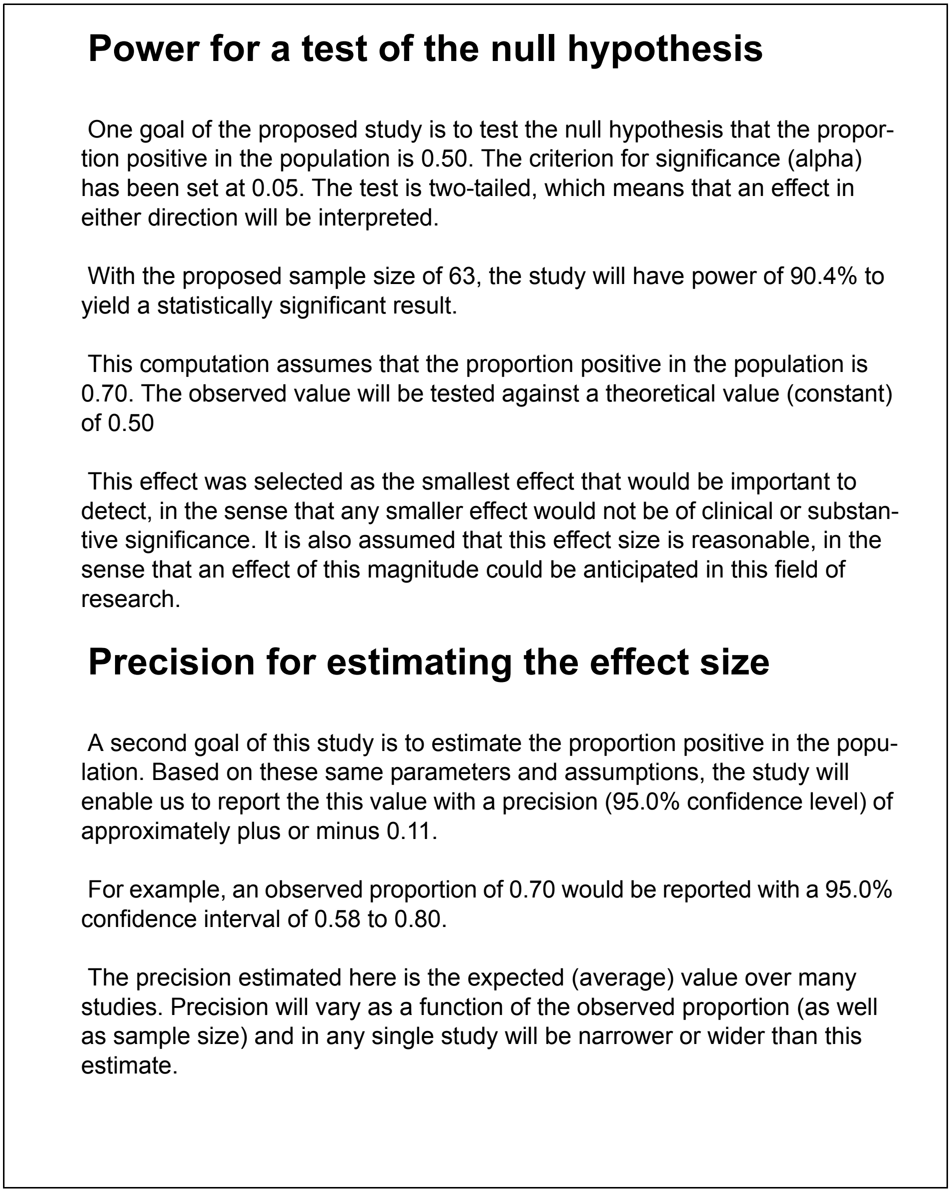
A researcher wants to test the null hypothesis that 50% of patients will report that their quality of life has improved as a result of their treatment. The smallest difference that would be important to detect is a difference of 70% versus this null value.

- Choose *One sample test that proportion = specific value*.

For the following steps, see Figure 9.1:

- Choose either version of the One Proportion test.
- For effect size, enter 0.70 (versus 0.50).
- Click *Alpha* and select the following values: *Alpha = 0.05*, *Confidence Level = 95%*, and *Tails = 2*.
- Press Ctrl-F (or click the *Find N* icon if the default power has already been set to 90). The program shows that a sample of 63 patients will yield power of 90%.
- The program shows that an observed proportion of 70% would be reported, with a 95% confidence interval of 58% to 80%.

- Click the *Report* icon to generate the following report:



Power for a test of the null hypothesis

One goal of the proposed study is to test the null hypothesis that the proportion positive in the population is 0.50. The criterion for significance (alpha) has been set at 0.05. The test is two-tailed, which means that an effect in either direction will be interpreted.

With the proposed sample size of 63, the study will have power of 90.4% to yield a statistically significant result.

This computation assumes that the proportion positive in the population is 0.70. The observed value will be tested against a theoretical value (constant) of 0.50

This effect was selected as the smallest effect that would be important to detect, in the sense that any smaller effect would not be of clinical or substantive significance. It is also assumed that this effect size is reasonable, in the sense that an effect of this magnitude could be anticipated in this field of research.

Precision for estimating the effect size

A second goal of this study is to estimate the proportion positive in the population. Based on these same parameters and assumptions, the study will enable us to report the this value with a precision (95.0% confidence level) of approximately plus or minus 0.11.

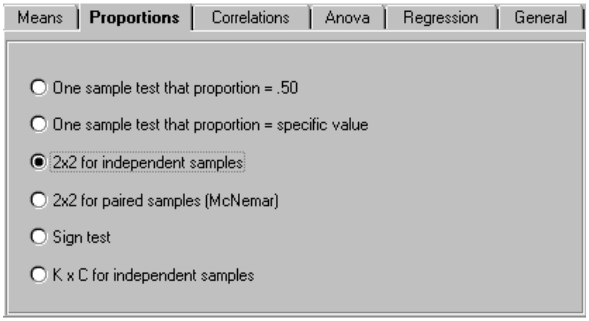
For example, an observed proportion of 0.70 would be reported with a 95.0% confidence interval of 0.58 to 0.80.

The precision estimated here is the expected (average) value over many studies. Precision will vary as a function of the observed proportion (as well as sample size) and in any single study will be narrower or wider than this estimate.

10 Proportions in Two Independent Groups

Selecting the Procedure

To display the available procedures, choose *New analysis* from the File menu.



Means | **Proportions** | Correlations | Anova | Regression | General

○ One sample test that proportion = .50

○ One sample test that proportion = specific value

● 2x2 for independent samples

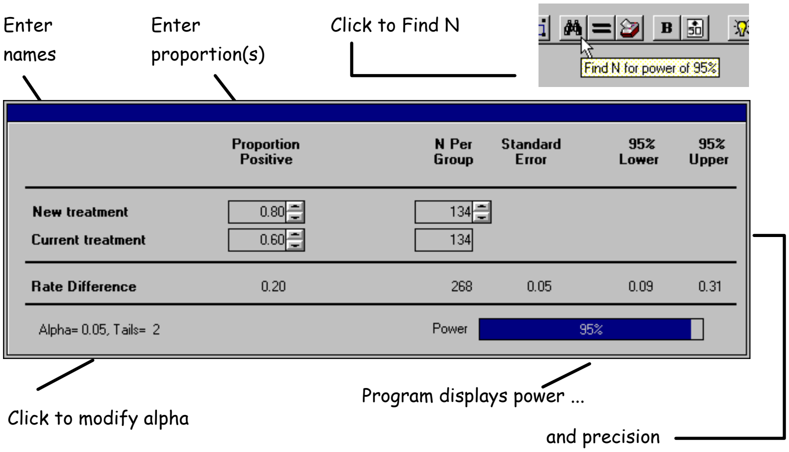
○ 2x2 for paired samples (McNemar)

○ Sign test

○ K x C for independent samples

Application

Figure 10.1 2x2 for independent samples



Enter names

Enter proportion(s)

Click to Find N

	Proportion Positive	N Per Group	Standard Error	95% Lower	95% Upper
New treatment	0.80	134			
Current treatment	0.60	134			
Rate Difference	0.20	268	0.05	0.09	0.31

Alpha= 0.05, Tails= 2

Power 95%

Click to modify alpha

Program displays power ...

and precision

The two-group test of proportions is used to test the hypothesis that the proportion of cases meeting some criterion is identical in the two groups.

For example, we are planning to assign patients to one of two treatment options and then test the null hypothesis that the treatments are equally effective (that is, that the proportion of patients cured will be identical in the two populations).

The user's attention is called to the fact that the program assumes that the row marginals are fixed. This assumption is not likely to be reasonable for certain types of studies, such as case control studies. This can have important implications for the computation of power as well as precision.

Effect Size

The effect size for the two-sample test of proportions is based on the difference between the two proportions. Unlike the t-test, where a difference of 10 versus 20 is equivalent to a difference of 40 versus 50 (that is, a 10-point difference in either case), when we work with proportions, the absolute values of the two proportions are relevant. Concretely, a difference of 10% versus 20% represents a more detectable effect than a difference of 40% versus 50%.

For this reason, we refer to the effect by citing the rate difference followed by the actual values—for example, a 10-point difference (40% versus 50%).

The effect size reported in this way can range from 0 (indicating no effect) to 0.999. For research in the social sciences, Cohen has suggested the following conventional values—small (40% versus 50%), medium (40% versus 65%), and large (40% versus 78%).

Alpha, Confidence Level, and Tails

Click *Alpha*, *Tails*, or the confidence level to modify these values. The value set for tails applies to both the power analysis and the confidence interval.

Sample Size

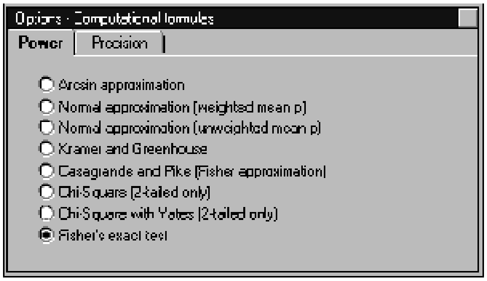
The program assumes that cases will be entered into the two groups in equal numbers. To modify, click *N of Cases*. The program allows the user to link the cases using some other assignment ratio (for example, 1:2 or 3:5). As long as the groups are linked, the user will enter a sample size for one group only. The program also allows the user to enter the number for each group independently of the other.

The spin control adjacent to the N of Cases value can be used to modify the sample size quickly. The size of the increment can be specified by the user (click *N of Cases*).

Click the *Find N* icon to have the program find the number of cases required for the default level of power. The program will honor any assignment ratio (for example, 2:1) that has been specified by the user. The default value for power is 90%, but this can be modified temporarily (Ctrl-F) or permanently (choose *Preferences* from the Options menu).

Computational Options for Power

The program allows the user to select from several computational formulas for power. To switch between options, choose *Computational formulas* from the Options menu.



○ Arcsin approximation
○ Normal approximation (weighted mean p)
○ Normal approximation (unweighted mean p)
○ Kramer and Greenhouse
○ Casagrande and Pike (Fisher approximation)
○ Chi-Square (2-tailed only)
○ Chi-Square with Yates (2-tailed only)
● Fisher's exact test

The most accurate estimate of power is given by Fisher's exact test (this is true even if the study will be analyzed by the chi-square test rather than the Fisher test), but this computation (unlike the others) involves an iterative procedure.

The user is advised to use the Casagrande and Pike option, which yields an excellent approximation to Fisher's exact test. This method requires no iterations and allows access to all of the program tools (Find N, tables, etc.). As a final step, for a small number of cases, you may want to select *Fisher's exact test* to get an exact computation, which will generally fall within 0.01 or 0.02 of the Casagrande estimate.

The other computational options are summarized here. These options became popular because they do not require a computer, but they have little to recommend them here. With large sample sizes, all formulas yield similar results.

Normal approximation. Two options for the normal approximation are offered: unweighted and weighted. Under the former, the null value is set at the midpoint between the two populations; under the latter, it is weighted by the number of cases in either group. When the two groups are of equal size, the two formulas are identical. Except under special circumstances, the unweighted option should be selected. The program also includes the arcsin, a variant on the normal approximation discussed by Cohen (1988).

Chi-square. The program allows the user to compute power using the noncentral chi-square distribution, with or without the Yates correction. These options are available for a two-tailed test only.

Approximations to Fisher's exact test. A number of approximations have been developed for computing power for Fisher's exact test. The Kramer-Greenhouse method is seen as overly conservative and generally yields the lowest estimates of power. The Casagrande and Pike method yields results that are very close to the exact computation of power by the Fisher method.

The Fisher exact method. The program allows the user to compute power for Fisher's exact test. This test gives the most accurate estimate of power, even when the post-study analysis will be performed using the chi-square test rather than the Fisher test.

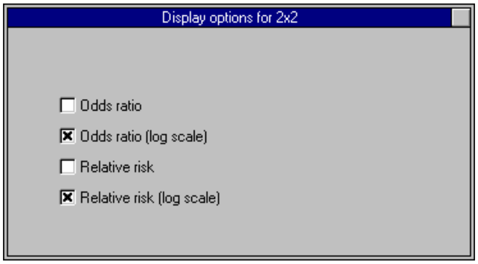
Computational Options for Precision

The program offers two options for computing confidence intervals: the log method or the Cornfield method.

Options for Screen Display

By default, the program displays the rate difference corresponding to the two proportions. The program will also compute the corresponding relative risk and odds ratio and display these together with the corresponding expected confidence intervals.

This feature allows the researcher to ensure that the effect size specified for power corresponds to a relative risk (or odds ratio) that is appropriate. Additionally, if the researcher intends to report the relative risk or odds ratio after the study is completed, this provides an estimate of the precision with which these parameters will be reported. To toggle the display of these indices, choose *Customize screen* from the Options menu.



Display options for 2x2

- ☐ Odds ratio
- ☒ Odds ratio (log scale)
- ☐ Relative risk
- ☒ Relative risk (log scale)

Example

We are planning to assign patients to one of two treatment options (new versus current) and to test the null hypothesis that the treatments are equally effective (that is, that the proportion of patients cured will be identical in the two populations).

The cure rate for the current treatment is approximately 60%. We anticipate that the new treatment will prove more effective in curing the illness. However, the new treatment is also expected to have more serious side effects. For this reason, the treatment would be recommended only if it increased the cure rate by 20 percentage points. Accordingly, the effect size selected for the power analysis will be 60% versus 80%.

While it is unlikely that the aggressive treatment will result in a cure rate lower than the standard, this possibility cannot be ruled out entirely, and it would have practical implications for future research. For this reason, the analysis will be two-tailed.

❖ Choose *2x2 for independent samples*.

For the following steps, see Figure 10.1:

❖ Enter the group names—*New treatment* and *Current treatment*. Click *Proportion Positive* and change it to *Proportion Cured*.

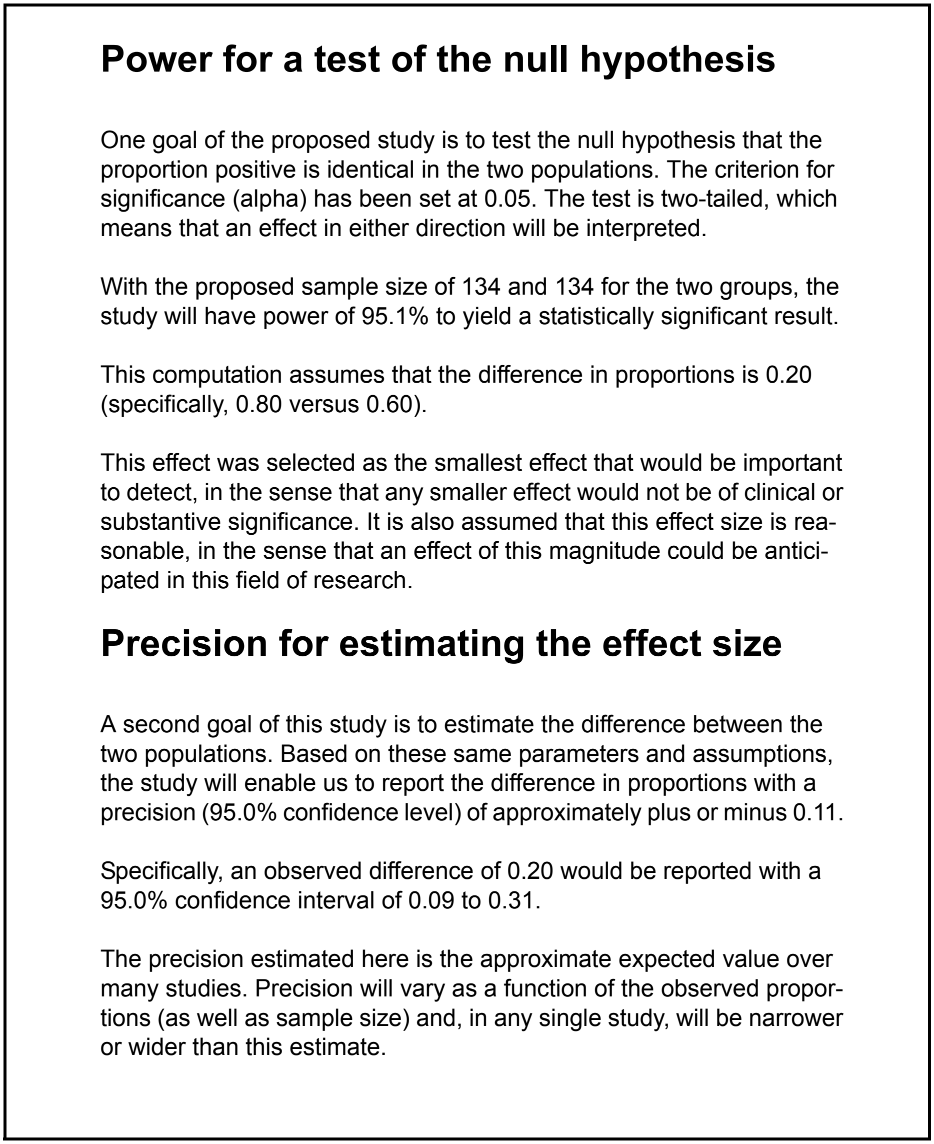
❖ Enter the effect size—proportions of 0.80 and 0.60.

❖ Click *Alpha* and select the following values: *Alpha = 0.05*, *Confidence Level = 95%*, and *Tails = 2*.

❖ Press Ctrl-F (or click the *Find N* icon if the default power has already been set to 95). The program shows that a sample of 134 patients per group will yield power of 95%.

- The program shows that an observed rate difference of 20 points would be reported with a 95% confidence interval of 9 points to 31 points. To narrow the confidence interval, the user could increase the sample size (or modify the confidence level).

- Click the *Report* icon to generate the following report:



Power for a test of the null hypothesis

One goal of the proposed study is to test the null hypothesis that the proportion positive is identical in the two populations. The criterion for significance (alpha) has been set at 0.05. The test is two-tailed, which means that an effect in either direction will be interpreted.

With the proposed sample size of 134 and 134 for the two groups, the study will have power of 95.1% to yield a statistically significant result.

This computation assumes that the difference in proportions is 0.20 (specifically, 0.80 versus 0.60).

This effect was selected as the smallest effect that would be important to detect, in the sense that any smaller effect would not be of clinical or substantive significance. It is also assumed that this effect size is reasonable, in the sense that an effect of this magnitude could be anticipated in this field of research.

Precision for estimating the effect size

A second goal of this study is to estimate the difference between the two populations. Based on these same parameters and assumptions, the study will enable us to report the difference in proportions with a precision (95.0% confidence level) of approximately plus or minus 0.11.

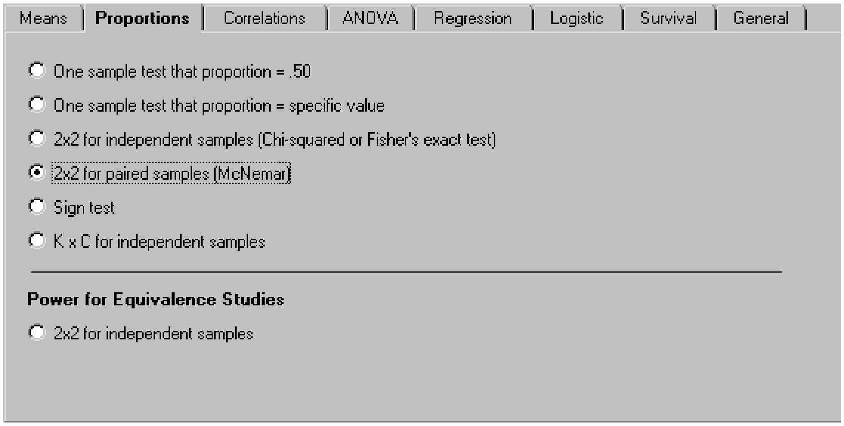
Specifically, an observed difference of 0.20 would be reported with a 95.0% confidence interval of 0.09 to 0.31.

The precision estimated here is the approximate expected value over many studies. Precision will vary as a function of the observed proportions (as well as sample size) and, in any single study, will be narrower or wider than this estimate.

11 Paired Proportions

Selecting the Procedure

To display the available procedures, choose *New Analysis* from the File menu.



○ One sample test that proportion = .50

○ One sample test that proportion = specific value

○ 2x2 for independent samples (Chi-squared or Fisher's exact test)

● 2x2 for paired samples (McNemar)

○ Sign test

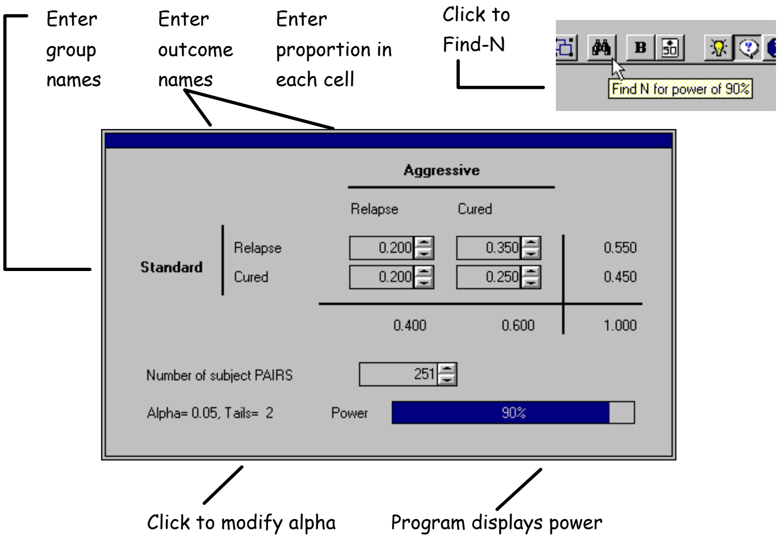
○ K x C for independent samples

Power for Equivalence Studies

○ 2x2 for independent samples

Application

Figure 11.1 2X2 for paired samples (McNemar)



Enter group names

Enter outcome names

Enter proportion in each cell

Click to Find-N

Find N for power of 90%

		Aggressive		
		Relapse	Cured	
Standard	Relapse	0.200	0.350	0.550
	Cured	0.200	0.250	0.450
		0.400	0.600	1.000

Number of subject PAIRS 251

Alpha= 0.05, Tails= 2 Power 90%

Click to modify alpha

Program displays power

The McNemar test of paired proportions is used to compare the proportion of cases in two groups when the cases in the two groups are matched in a way that is relevant to the outcome—for example, when patients are matched on the basis of disease severity and then assigned to one of two treatments, or when siblings are assigned to one of two conditions.

Effect Size

The effect size for the two-sample test of paired proportions is based on the difference between the two proportions. In this sense, the McNemar test is identical to the test of proportions in two independent groups. In the case of the McNemar test, however, we work with pairs of cases rather than individual cases.

Assume, for example, that patients are matched for disease severity and then assigned to either the standard treatment or to an aggressive treatment. The possible outcomes are *sick* or *cured*. When both members of a pair are classified as sick or both are classified as cured, the pair provides no information about the relative utility of the two treatments. Therefore, the test of the hypothesis that the two treatments are equally effective is based entirely on cases where one member of the pair was cured and the other was not.

It follows that for the purpose of computing power, the effective number of pairs is the number falling in the upper right and lower left cells of a 2x2 table (that is, those cells where one member of the pair is cured and the other is not). For the purpose of computing power, the effect size is the difference between the proportion in these two cells.

As is true for any test of proportions, a difference of 10% versus 20% is *not* equivalent to a difference of, say, 40% versus 50%, despite the fact that the difference in either case is 10 percentage points. (Proportions near 50% have larger variance and therefore smaller effect sizes). For this reason, we refer to the effect by citing the rate difference followed by the actual values—for example, a 10-point difference (40% versus 50%).

Alpha and Tails

Click *Alpha* or *Tails* to modify these values.

Sample Size

The program requires the user to enter a value for *Total Number of Pairs* and then specify the proportion of pairs expected to fall into each cell. As above, the number of pairs for computation of power will be determined by the program based on the product of these two values.

For any given number of cases, power will be higher if the proportion of cases falling into the upper right and lower left cells is increased, which increases the effective number of cases. Of course, power will also increase as the disparity between the upper right and lower left cells increases, since this is the basis of the effect size.

The spin control adjacent to the *N of Cases* value can be used to modify the sample size quickly. The size of the increment can be specified by the user (click *N of Cases*).

Click the *Find N* icon to have the program find the number of cases required for the default level of power. The default value for power is 90%, but this can be modified temporarily (Ctrl-F) or permanently (select *Preferences* from the Options menu).

Computational Options for Power

The program allows the user to select the normal (arcsin) approximation or the exact (binomial) test. When the sample size is small, the binomial test is preferable. Otherwise, the normal approximation should be selected (with a large sample size, the binomial option can require several seconds for iterations). To select either option, choose *Computational formulas* from the Options menu.

Example

We are planning to assign patients to match patients on the basis of disease severity and assign the two members of each pair to separate treatment options (standard versus aggressive). We will test the null hypothesis that the treatments are equally effective (in that the proportion of patients cured will be identical in the two populations).

We anticipate that in 20% of pairs, both patients will relapse; in 25%, both will be cured; in 35%, the patient in the aggressive group will have the better outcome; in 20%, the patient in the standard group will have the better outcome.

While it is unlikely that the aggressive treatment will result in a cure rate lower than the standard, this possibility cannot be ruled out entirely, and it would have practical implications for future research. For this reason, the analysis will be two-tailed.

- ❖ Choose *2x2 for paired samples (McNemar).*

For the following steps, see Figure 11.1:

- ❖ Enter group names—*Aggressive* and *Standard.*
- ❖ Enter names for the outcomes—*Relapse* and *Cured.*
- ❖ Enter the proportion falling into each cell (0.20, 0.35, 0.20, and 0.25).
- ❖ Click *Alpha* and select the following values: *Alpha = 0.05*, *Confidence Level = 95%*, *Tails = 2.*
- ❖ Press Ctrl-F (or click the *Find N* icon if the default power has already been set to 90). The program shows that a sample of 251 pairs (502 patients) will yield power of 90%.
- ❖ Click the *Report* icon to generate the following report:

Power for a test of the null hypothesis

One goal of the proposed study is to test the null hypothesis that there is no relationship between the variable used to classify cases (Aggressive versus Standard) and outcome (Relapse versus Cured). This will be tested by the McNemar test for paired proportions. The criterion for significance (alpha) has been set at 0.05 (two-tailed).

The power analysis is based on the following population effect size: in 20.0% of pairs, both cases will be classified as Relapse and in another 25.0%, both cases will be classified as Cured. These cases contribute no information to the hypothesis test.

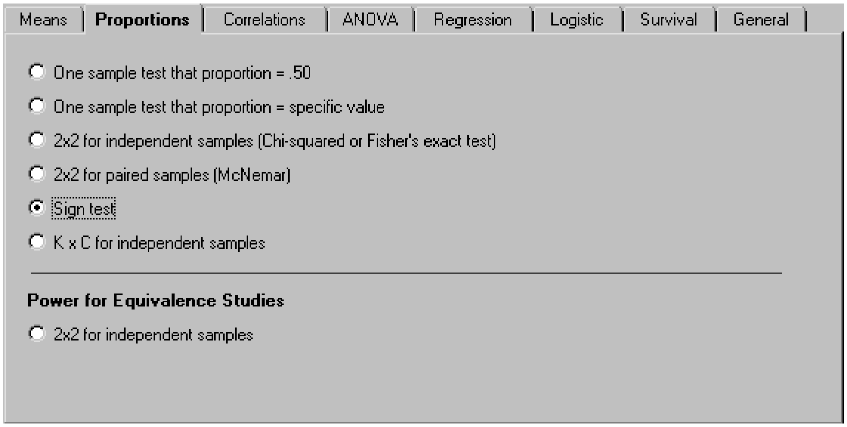
A discrepancy between the two members of a pair is expected in the balance of the population: 35.0% of all pairs will show an outcome of Cured for the Standard case only, while 20.0% of all pairs will show an outcome of Cured for the Aggressive case only. This effect was selected as the smallest effect that would be important to detect, in the sense that any smaller effect would not be of clinical or substantive significance. It is also assumed that this effect size is reasonable, in the sense that an effect of this magnitude could be anticipated in this field of research.

With the proposed sample size of 251, the study will have power of 90.1% to yield a statistically significant result.

12 Sign Test

Selecting the Procedure

To display the available procedures, choose *New analysis* from the File menu.



○ One sample test that proportion = .50

○ One sample test that proportion = specific value

○ 2x2 for independent samples (Chi-squared or Fisher's exact test)

○ 2x2 for paired samples (McNemar)

● Sign test

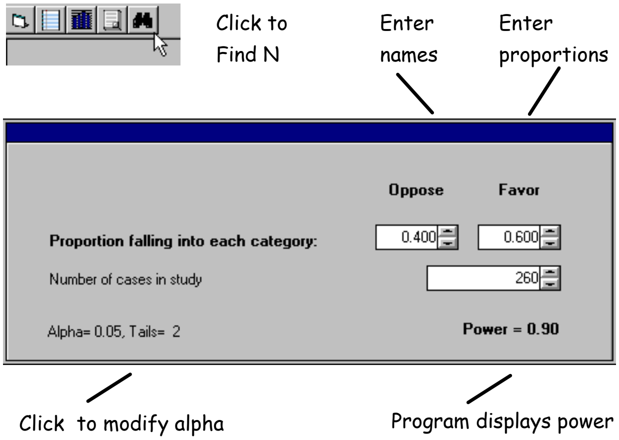
○ K x C for independent samples

Power for Equivalence Studies

○ 2x2 for independent samples

Application

Figure 12.1 Sign test



Click to Find N

Enter names

Enter proportions

Oppose Favor

Proportion falling into each category: 0.400 0.600

Number of cases in study 260

Alpha= 0.05, Tails= 2

Power = 0.90

Click to modify alpha

Program displays power

The sign test is used to compare the proportion of cases in one of two mutually exclusive groups. For example, we may classify people as planning to vote for one of two candidates.

In this test, the proportion of cases falling into either outcome is completely determined by the proportion in the complementary group. Therefore, the user is required to enter the proportion falling into either one of the groups only. A proportion of, say, 40% in one outcome group implies a proportion of 60% in the other.

Effect Size

The effect size for the two-sample test of paired proportions is based on the difference between either proportion and the null hypothesis of 50% (of necessity, the two proportions are equidistant from this null value).

For research in the social sciences, Cohen has suggested the following conventional values—small (0.45 versus 0.55), medium (0.35 versus 0.65), and large (0.25 versus 0.75). Wherever possible, however, the selection of an effect size should be based on the research issues, that is, substantively.

Alpha and Tails

Click *Alpha* or *Tails* to modify these values.

Sample Size

The program requires the user to enter a total number of cases and then specify the proportion falling into either of the two outcomes.

The spin control adjacent to the N of Cases value can be used to modify the sample size quickly. The size of the increment can be specified by the user (click *N of Cases*).

Click the *Find N* icon to have the program find the number of cases required for the default level of power. The default value for power is 90%, but this can be modified temporarily (Ctrl-F) or permanently (choose *Preferences* from the Options menu).

Computational Options for Power

The program allows the user to select the arcsin approximation or the exact (binomial) test. When the sample size is small, the binomial test is preferable. Otherwise, the normal approximation should be selected (with a large sample size, the binomial option can require several seconds for iterations). To select either option, choose *Computational formulas* from the Options menu.

Example

A politician wants to conduct a poll to find out how her constituents feel about an issue on the upcoming ballot. If 60% of the population has a preference, it would be important to reject the null hypothesis of no effect.

❖ Choose *Sign test.*

For the following steps, see Figure 12.1:

❖ Enter the group names—*Oppose* and *Favor.*

❖ Enter the proportion falling into the *Favor* cell (0.60).

❖ Click *Alpha* and select the following values: *Alpha = 0.05* and *Tails = 2*.

❖ Press Ctrl-F (or click the *Find N* icon if the default power has already been set to 90). The program shows that a sample of 260 persons will yield power of 90%.

❖ Click the *Report* icon to generate the following report:

Power for a test of the null hypothesis

One goal of the proposed study is to test the null hypothesis that Oppose and Favor contain equal proportions of cases (that is, that half of the population falls into either classification). The criterion for significance (alpha) has been set at 0.05 (two-tailed).

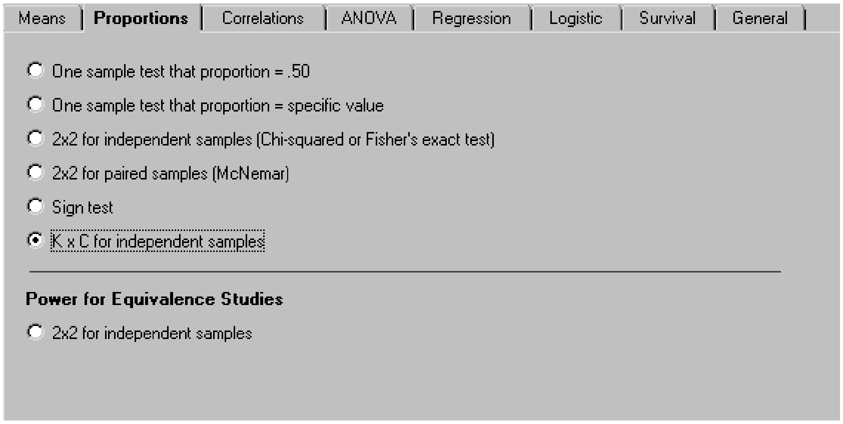
The power analysis is based on a population effect size such that 40% of cases fall into Oppose and 60% fall into Favor, for a discrepancy of 20 percentage points. This effect was selected as the smallest effect that would be important to detect, in the sense that any smaller effect would not be of clinical or substantive significance. It is also assumed that this effect size is reasonable, in the sense that an effect of this magnitude could be anticipated in this field of research.

With the proposed sample size of 260, the study will have power of 90.1% to yield a statistically significant result.

13 K x C Crosstabulation

Selecting the Procedure

To display the available procedures, choose *New analysis* from the File menu.



○ One sample test that proportion = .50

○ One sample test that proportion = specific value

○ 2x2 for independent samples (Chi-squared or Fisher's exact test)

○ 2x2 for paired samples (McNemar)

○ Sign test

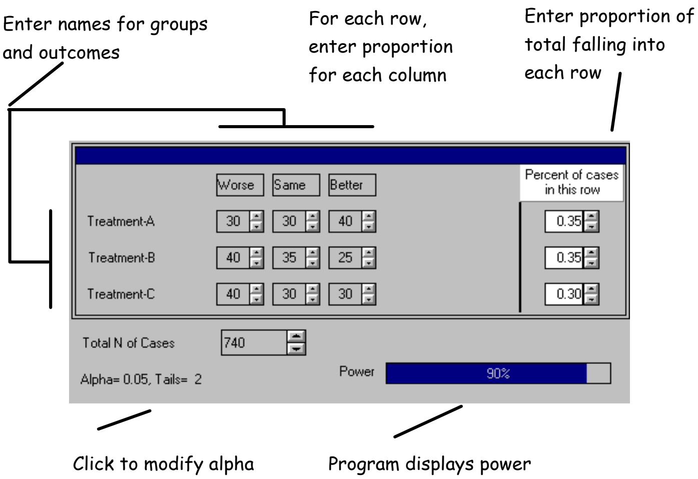
● K x C for independent samples

Power for Equivalence Studies

○ 2x2 for independent samples

Application

Figure 13.1 K x C crosstabulation



	Worse	Same	Better	Percent of cases in this row
Treatment-A	30	30	40	0.35
Treatment-B	40	35	25	0.35
Treatment-C	40	30	30	0.30

Total N of Cases 740

Alpha= 0.05, Tails= 2

Power 90%

Enter names for groups and outcomes

For each row, enter proportion for each column

Enter proportion of total falling into each row

Click to modify alpha

Program displays power

The K × C (K rows by C columns) crosstabulation is used to test the hypothesis that classification on one dimension (for example, assignment to one of three treatments) is related to classification on another dimension (for example, outcome, where patients are assigned to one of three possible outcomes).

Effect Size

This discussion will refer to a running example in which patients are assigned to one of three treatment groups and then classified as having one of three outcomes. However, the test is not limited to situations in which one variable logically precedes the other. Additionally, the test does not require that the matrix of rows by columns be square.

Under the null hypothesis of no effect, outcome will have no relation to treatment. If 10% of patients are classified as improved, then this number should be 10% for each of the three treatment groups. If 40% of patients are classified as the same, then this number should be 40% for each of the treatment groups.

The effect size, w, is based on the disparity between the matrix of proportions that we would expect under the null hypothesis and the matrix that is assumed under the alternative hypothesis.

To enter the effect size for the $K \times C$ crosstabulation:

- Specify the number of rows and columns in the table (choose *Data entry/Study design* from the Options menu).

- Enter the percentage of cases falling into each column for each row in the table. The proportions for each row must sum to 100.

- Enter the proportion of all cases falling into each row of the table. The proportions must sum to 1.00.

This approach is somewhat different from the approach sometimes taken in texts, in which the researcher is required to enter the proportion of the full population falling into each cell (so that *all* cells, rather than the cells in a single row, sum to 1.00). The approach taken by this program ensures that the cells for two rows will appear to be identical under the null, even if more cases are assigned to one row than another. This makes it easier for the researcher to identify the places in which the rows will differ.

The effect size computed in this way (w) is analogous to the chi-square value used to test for significance, but is a pure measure of effect size, whereas chi-square is affected also by the sample size. The relation between the two is given by:

$$\text{chi-square} = w^2 * \text{Ntotal}$$

The effect size, w, can be used to derive two additional indices of effect—the contingency coefficient (C) and Cramér's phi. The program can be set to display these values as well as w.

Alpha

Click *Alpha* to modify its value.

Sample Size

The program assumes initially that cases will be divided evenly across the rows (50% to each of two groups, 33.3% to each of three groups, and so on), but this can be modified by the user. For example, the user can specify that 50% of all cases will be assigned to *Treatment A* and 25% each to *Treatments B* and *C*.

The spin control adjacent to the N of Cases value can be used to modify the sample size quickly. The size of the increment can be specified by the user (click *N of Cases*).

Click the *Find N* icon to have the program find the number of cases required for the default level of power. The default value for power is 90%, but this can be modified temporarily (Ctrl-F) or permanently (choose *Preferences* from the Options menu).

Computational Options for Power

Power is computed with reference to the noncentral chi-square distribution. No options are available.

Example

We are planning to assign patients to one of three treatment options. Subsequent to treatment, each patient will be classified as *Worse*, *Same*, or *Better*. The null hypothesis is that the treatments are equally effective.

❖ Choose *K x C for independent samples*.

For the following steps, see Figure 13.1:

❖ From the Options menu choose *Data entry/Study design* and set 3 rows and 3 columns.

❖ Enter the group names—*Treatments A*, *B*, and *C*.

❖ Enter the names for the outcomes—*Worse*, *Same*, and *Better*.

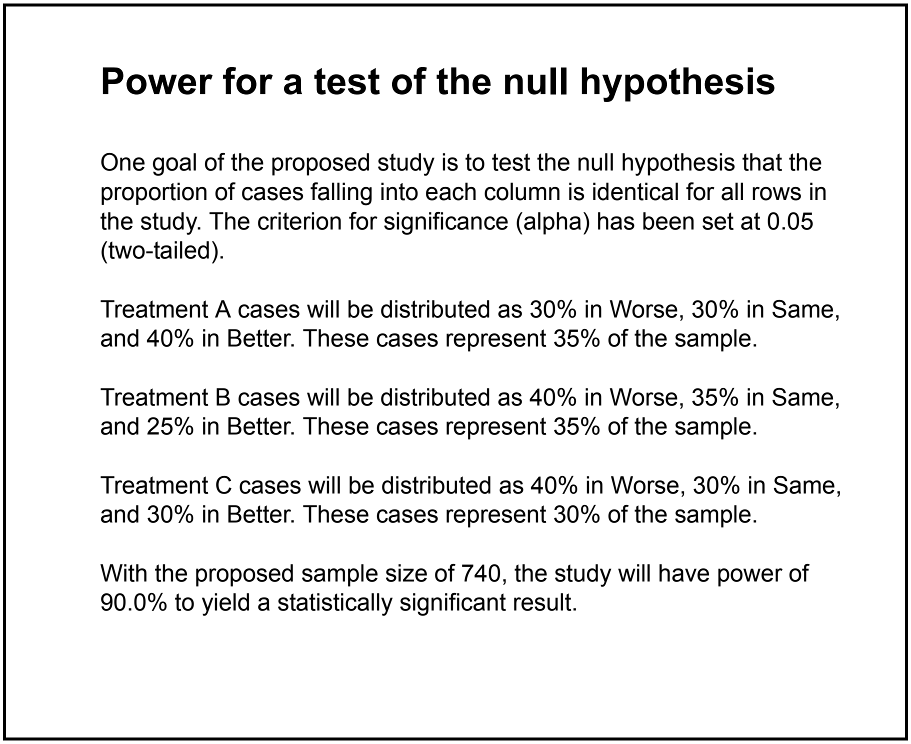
❖ For the first row, enter the percentage falling into each cell (30, 30, and 40). Repeat for the second (40, 35, and 25) and third (40, 30, and 30) rows.

❖ Enter the proportion of the full sample assigned to each of the three treatments—0.35, 0.35, and 0.30.

❖ Click *Alpha* and select the following value: *Alpha = 0.05*.

❖ Press Ctrl-F (or click the *Find N* icon if the default power has already been set to 90). The program shows that a sample of 740 patients will yield power of 90%.

❖ Click the *Report* icon to generate the following report:



Power for a test of the null hypothesis

One goal of the proposed study is to test the null hypothesis that the proportion of cases falling into each column is identical for all rows in the study. The criterion for significance (alpha) has been set at 0.05 (two-tailed).

Treatment A cases will be distributed as 30% in Worse, 30% in Same, and 40% in Better. These cases represent 35% of the sample.

Treatment B cases will be distributed as 40% in Worse, 35% in Same, and 25% in Better. These cases represent 35% of the sample.

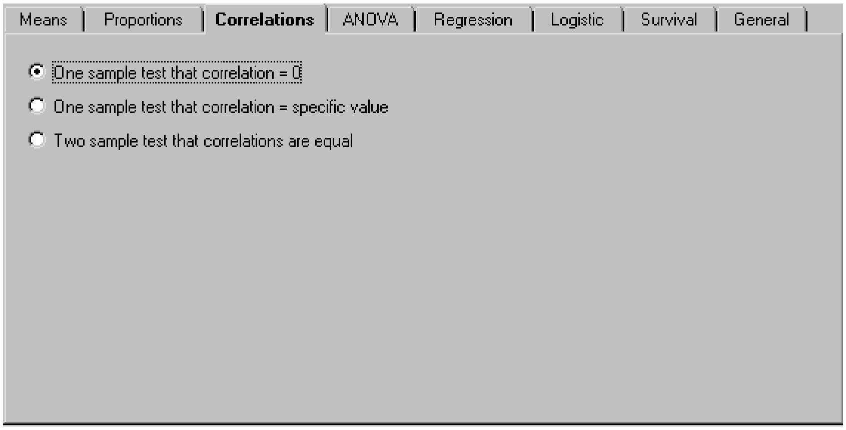
Treatment C cases will be distributed as 40% in Worse, 30% in Same, and 30% in Better. These cases represent 30% of the sample.

With the proposed sample size of 740, the study will have power of 90.0% to yield a statistically significant result.

14 Correlation—One Group

Selecting the Procedure

To display the available procedures, choose *New analysis* from the File menu.



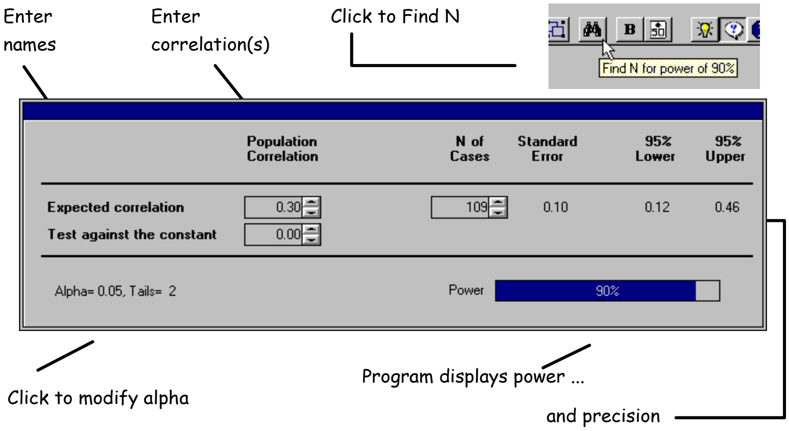
● One sample test that correlation = 0

○ One sample test that correlation = specific value

○ Two sample test that correlations are equal

Application

Figure 14.1 One-sample test that correlation is 0



Enter names

Enter correlation(s)

Click to Find N

	Population Correlation	N of Cases	Standard Error	95% Lower	95% Upper
Expected correlation	0.30	109	0.10	0.12	0.46
Test against the constant	0.00				

Alpha= 0.05, Tails= 2

Power 90%

Click to modify alpha

Program displays power ...

and precision

The one-sample correlation procedure is used to test the hypothesis that a correlation is 0 (or any other specific value).

Effect Size

When a correlation is being tested against the null value of 0, the effect size is simply the correlation coefficient (r). The effect size, r, can range from 0 to an absolute value of 1.00 (the program will accept any value up to 0.999). The program allows r to be either negative or positive, but the effect size is based on the absolute value.

The program will also compute power for a test of the null other than 0. The effect size is based on the difference between the two correlations (the null value and the correlation under the alternate), and the sign of each correlation is important. For correlations, an effect size of 0.30 versus 0.10 is not the same as an effect size of 0.50 versus 0.30, despite the fact that the difference is 20 points in either case.

Alpha, Confidence Level, and Tails

Click *Alpha, Tails*, or the confidence level to modify these values. The value set for tails applies to both the power analysis and the confidence interval.

Sample Size

The spin control adjacent to the *N of Cases* value can be used to modify the sample size quickly. The size of the increment can be specified by the user (click *N of Cases*).

Click the *Find N* icon to have the program find the number of cases required for the default level of power. The default value for power is 90%, but this can be modified temporarily (Ctrl-F) or permanently (choose *Preferences* from the Options menu).

Computational Options for Power

Power for a correlation versus the null of 0 is computed using exact methods (similar to those used for multiple regression).

Power for testing a single correlation versus a constant other than 0 is carried out by means of the Fisher-*Z* transformation.

Example 1

A college is trying out a test that it may use to place students into sections of a mathematics class in the future. At present, all students take the test during the first week of class and then attend the same class for the duration of the semester. The hypothesis is that scores on the placement test will be correlated with scores on the final examination.

The smallest correlation that would be meaningful is a correlation of 0.30, so this is the value used for effect size in computing power. The college decides to run the study with alpha set at 0.05 and power at 90%. While the test is not expected to have a negative correlation with grades, a finding that it does would have important implications, thus the test would be two-tailed.

❖ Choose *One sample test that correlation is zero*.

For the following steps, see Figure 14.1:

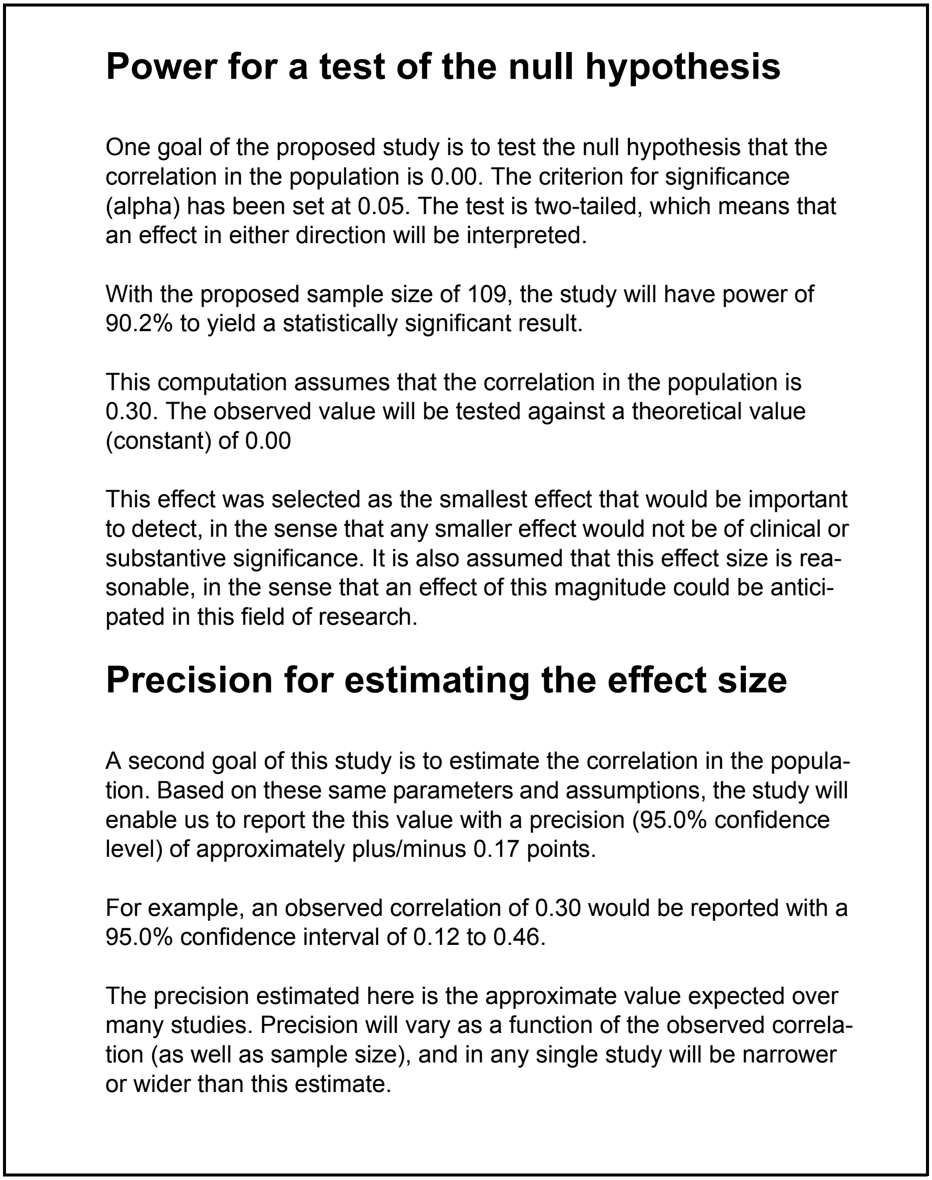
❖ Enter the correlation of 0.30.

❖ Click *Alpha* and select the following values: *Alpha = 0.05* and *Tails =2*.

❖ Press Ctrl-F (or click the *Find N* icon). The program shows that a sample of 109 students will yield power of 90%.

❖ The program shows also that an observed correlation of 0.30 would be reported with a 95% confidence interval of 0.12 to 0.46.

- Click the *Report* icon to generate the following report:



Power for a test of the null hypothesis

One goal of the proposed study is to test the null hypothesis that the correlation in the population is 0.00. The criterion for significance (alpha) has been set at 0.05. The test is two-tailed, which means that an effect in either direction will be interpreted.

With the proposed sample size of 109, the study will have power of 90.2% to yield a statistically significant result.

This computation assumes that the correlation in the population is 0.30. The observed value will be tested against a theoretical value (constant) of 0.00

This effect was selected as the smallest effect that would be important to detect, in the sense that any smaller effect would not be of clinical or substantive significance. It is also assumed that this effect size is reasonable, in the sense that an effect of this magnitude could be anticipated in this field of research.

Precision for estimating the effect size

A second goal of this study is to estimate the correlation in the population. Based on these same parameters and assumptions, the study will enable us to report the this value with a precision (95.0% confidence level) of approximately plus/minus 0.17 points.

For example, an observed correlation of 0.30 would be reported with a 95.0% confidence interval of 0.12 to 0.46.

The precision estimated here is the approximate value expected over many studies. Precision will vary as a function of the observed correlation (as well as sample size), and in any single study will be narrower or wider than this estimate.

Example 2

The situation is identical to Example 1. In this case, however, the administrator decides that it will be of little value to know merely that the correlation exceeds 0. The smallest correlation that would be important as a useful predictor is 0.30, and he wants to test the null hypothesis that the true correlation is 0.30.

If we can assume that the true correlation is 0.50, how many students would be needed to ensure power (90%) to reject this null hypothesis? Other parameters remain as before (alpha = 0.05, test is two-tailed).

❖ Choose *One sample test that correlation is specific value*.

For the following steps, see Figure 14.1:

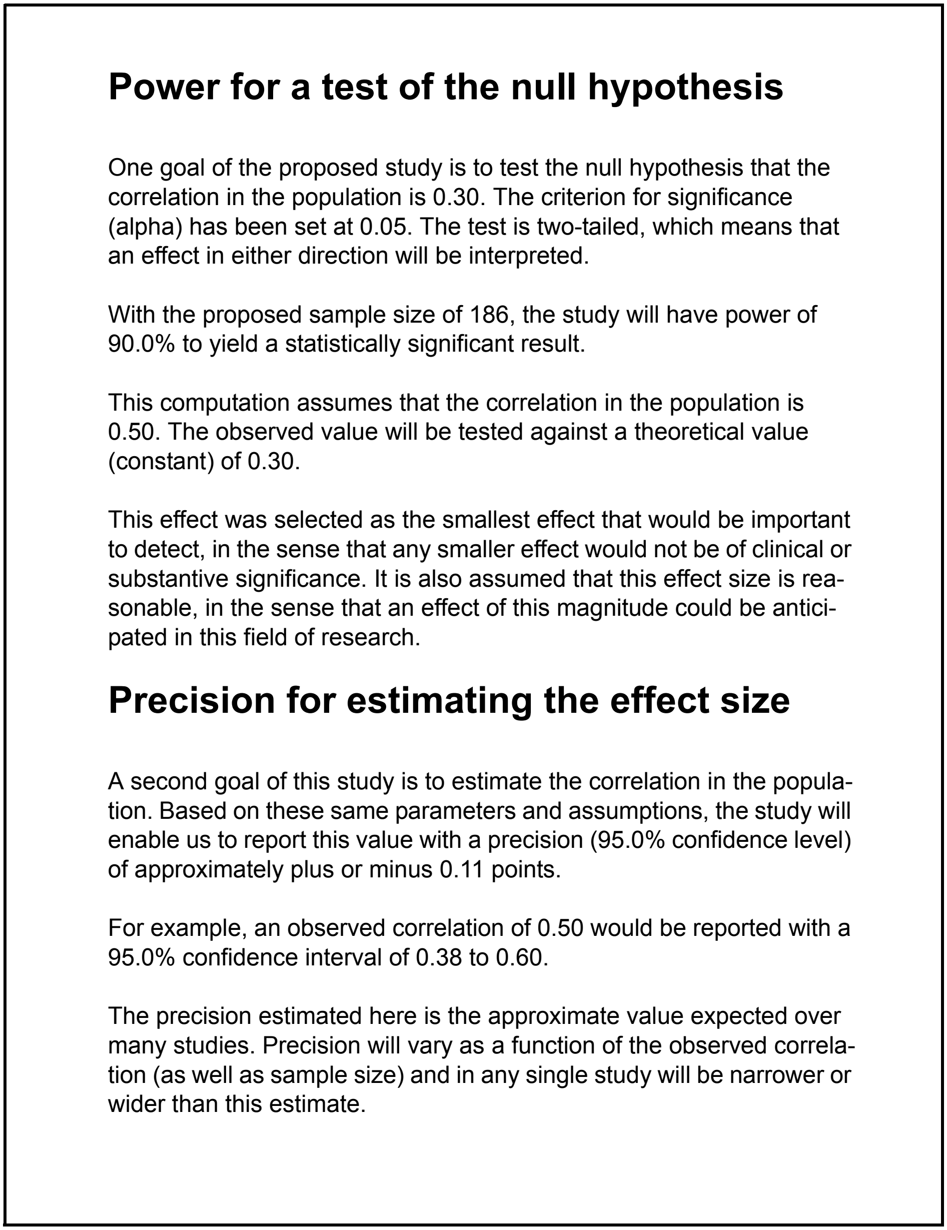
❖ Enter the correlation of 0.50 for the population, to be tested against a constant of 0.30.

❖ Click *Alpha* and select the following values: *Alpha = 0.05* and *Tails =2*.

❖ Press Ctrl-F (or click the *Find N* icon if the default power has already been set to 90). The program shows that a sample of 186 students will yield power of 90%.

❖ The program shows also that an observed correlation of 0.50 would be reported with a 95% confidence interval of 0.38 to 0.60 (note the lack of symmetry).

❖ Click the *Report* icon to generate the following report:



Power for a test of the null hypothesis

One goal of the proposed study is to test the null hypothesis that the correlation in the population is 0.30. The criterion for significance (alpha) has been set at 0.05. The test is two-tailed, which means that an effect in either direction will be interpreted.

With the proposed sample size of 186, the study will have power of 90.0% to yield a statistically significant result.

This computation assumes that the correlation in the population is 0.50. The observed value will be tested against a theoretical value (constant) of 0.30.

This effect was selected as the smallest effect that would be important to detect, in the sense that any smaller effect would not be of clinical or substantive significance. It is also assumed that this effect size is reasonable, in the sense that an effect of this magnitude could be anticipated in this field of research.

Precision for estimating the effect size

A second goal of this study is to estimate the correlation in the population. Based on these same parameters and assumptions, the study will enable us to report this value with a precision (95.0% confidence level) of approximately plus or minus 0.11 points.

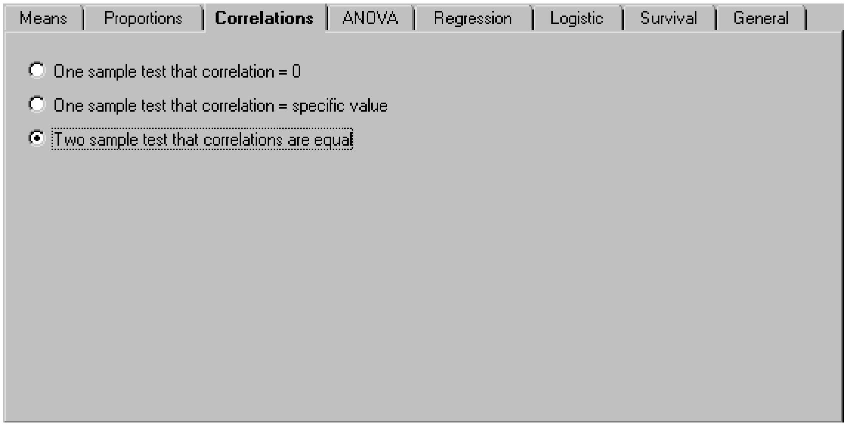
For example, an observed correlation of 0.50 would be reported with a 95.0% confidence interval of 0.38 to 0.60.

The precision estimated here is the approximate value expected over many studies. Precision will vary as a function of the observed correlation (as well as sample size) and in any single study will be narrower or wider than this estimate.

15 Correlation—Two groups

Selecting the Procedure

To display the available procedures, choose *New analysis* from the File menu.



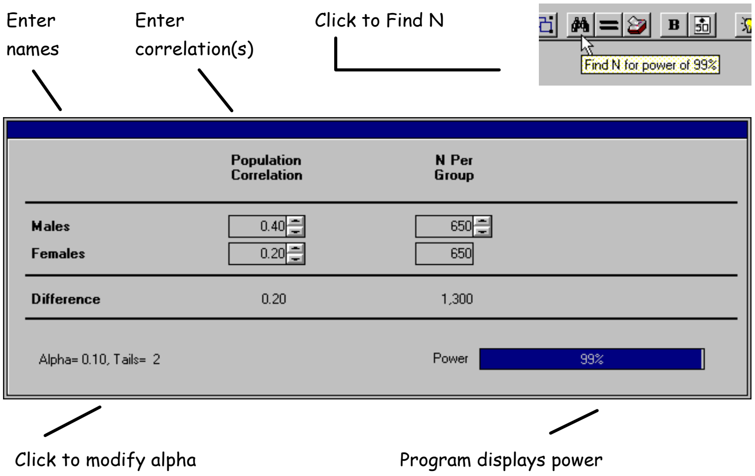
○ One sample test that correlation = 0

○ One sample test that correlation = specific value

● Two sample test that correlations are equal

Application

Figure 15.1 Two sample test that correlations are equal



Enter names

Enter correlation(s)

Click to Find N

	Population Correlation	N Per Group
Males	0.40	650
Females	0.20	650
Difference	0.20	1,300

Alpha= 0.10, Tails= 2

Power 99%

Click to modify alpha

Program displays power

The two-sample correlation procedure is used to test the hypothesis that the correlation between X and Y is identical in populations A and B—for example, to test that the predictive utility of a screening test (indexed by the correlation between the test and a criterion) is identical for males and females.

Effect Size

The effect size used here is based on the difference between the two correlations coefficients. However, a difference of 0.20 versus 0.40 is *not* as detectable as a difference of 0.40 versus 0.60, despite the fact that the difference is 20 points in both cases. Therefore, the difference will be expressed by presenting the absolute difference followed by the two correlations.

The effect size, r, can range from the case where the two correlations are identical (indicating no effect) to the case where one correlation is –0.999 and the other is +0.999. When we are comparing two correlations, the sign of each value is, of course, important.

Alpha, Confidence Level, and Tails

Click *Alpha*, *Tails*, or the confidence level to modify these values.

Sample Size

The spin control adjacent to the *N of Cases* value can be used to modify the sample size quickly. The size of the increment can be specified by the user (click *N of Cases*).

Click the *Find N* icon to have the program find the number of cases required for the default level of power. The default value for power is 90%, but this can be modified temporarily (Ctrl-F) or permanently (select *Preferences* from the Options menu).

Computational Options for Power

Power is computed using the Fisher-*Z* transformation.

Example

A college is trying out a test that it may use to place students into sections of a mathematics class in the future. At present, all students take the test during the first week of class and then attend the same class for the duration of the semester. The hypothesis is that scores on the placement test will correlate with scores on the final examination.

It is anticipated that the correlation between the placement test and score on the final grade will be about 0.30. There is concern, however, that the ability of this test to function as a predictor of outcome will be substantially stronger for males, who as a group have had more formal training in mathematics, than for the females, who may have had less formal training.

If indeed the test functions as a strong predictor for males but not for females, it would be unfair to use this as a placement test for the whole group. The administration decides that the test will be considered to have this problem if the correlation between the placement test and the final grade is 20 points higher for one gender than for the other (0.40 for males versus 0.20 for females). The study's power should be 99%—if there really is a gender gap of 20 points, the study should have power of 99% to yield a significant effect.

❖ Choose *Two sample test that correlations are equal.*

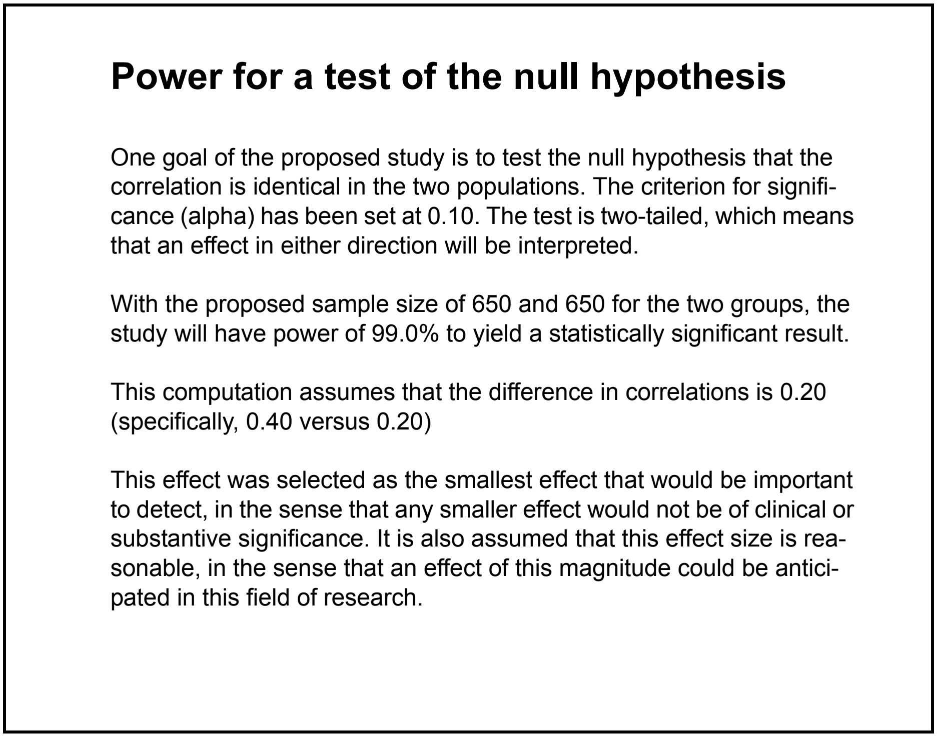
For the following steps, see Figure 15.1.

❖ Enter the correlations of 0.40 and 0.20.

❖ Click *Alpha* and select the following values: *Alpha = 0.10, Tails =2*.

❖ Press Ctrl-F (or click the *Find N* icon if the default power has already been set to 99). The program shows that a sample of 650 students per group will yield power of 99%.

❖ Click the *Report* icon to generate the following report:



Power for a test of the null hypothesis

One goal of the proposed study is to test the null hypothesis that the correlation is identical in the two populations. The criterion for significance (alpha) has been set at 0.10. The test is two-tailed, which means that an effect in either direction will be interpreted.

With the proposed sample size of 650 and 650 for the two groups, the study will have power of 99.0% to yield a statistically significant result.

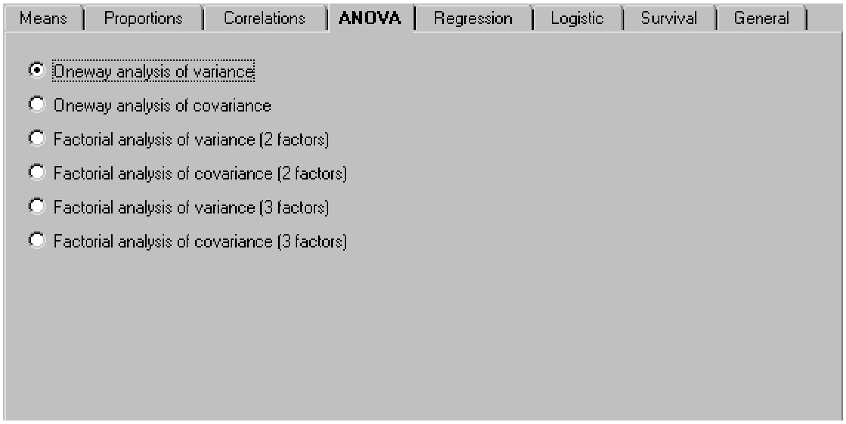
This computation assumes that the difference in correlations is 0.20 (specifically, 0.40 versus 0.20)

This effect was selected as the smallest effect that would be important to detect, in the sense that any smaller effect would not be of clinical or substantive significance. It is also assumed that this effect size is reasonable, in the sense that an effect of this magnitude could be anticipated in this field of research.

16 Analysis of Variance/Covariance (Oneway)

Selecting the Procedure

To display the available procedures, choose *New analysis* from the File menu.



● Oneway analysis of variance

○ Oneway analysis of covariance

○ Factorial analysis of variance (2 factors)

○ Factorial analysis of covariance (2 factors)

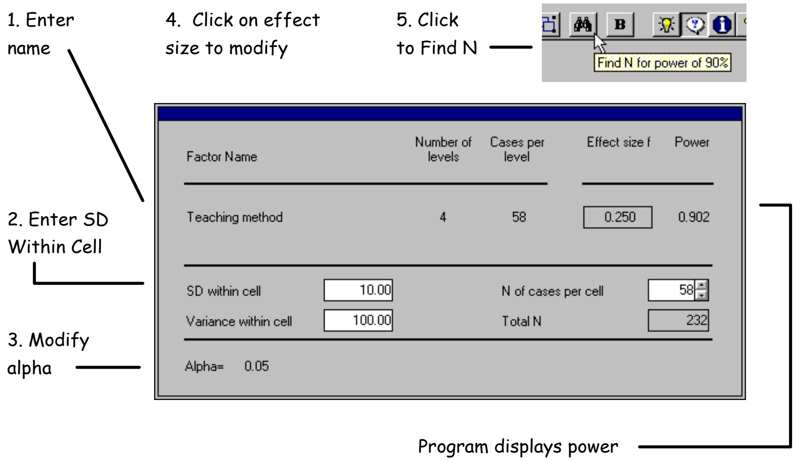
○ Factorial analysis of variance (3 factors)

○ Factorial analysis of covariance (3 factors)

Application

Oneway analysis of variance (ANOVA) is used to compare the means in more than two groups. Assume, for example, that students are assigned at random to one of four classroom methods, taught for a period of time, and then assessed on some continuous measure. Oneway ANOVA might be used to test the null hypothesis of no difference in effect among the four groups.

Figure 16.1 Oneway analysis of variance



1. Enter name

2. Enter SD Within Cell

3. Modify alpha

4. Click on effect size to modify

5. Click to Find N

Find N for power of 90%

Factor Name	Number of levels	Cases per level	Effect size f	Power
Teaching method	4	58	0.250	0.902

SD within cell 10.00

Variance within cell 100.00

N of cases per cell 58

Total N 232

Alpha= 0.05

Program displays power

Effect Size

The effect size (f) used in analysis of variance is an extension of the effect size (d) used for a t-test. Recall that d is the mean difference between groups divided by the dispersion within groups. Similarly, f is based on the dispersion (standard deviation) between groups divided by the dispersion (standard deviation) within groups. (The effect size, f, is a pure measure of effect size and should not be confused with the *F* statistic, which takes into account sample size as well as effect size.)

The program allows the user to enter f directly or by using Cohen's conventions for research in the social sciences (small = 0.10, medium = 0.25, and large = 0.40).

Alternatively, the user is allowed to provide data that the program uses to compute f. In this case, the program requires the user to provide the within-cells standard deviation on the main screen. Then, enter data for the between-groups standard deviation in one of three formats:

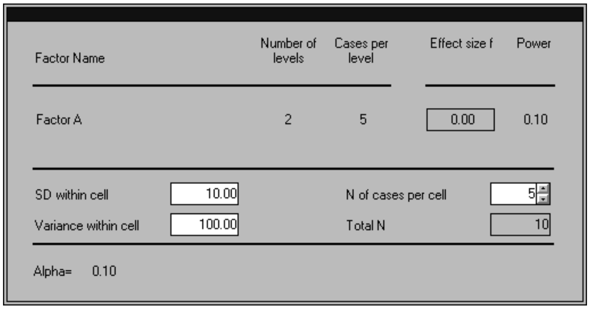
- Enter the range of means for the factor.
- Enter the mean for each group.
- Enter the between-groups standard deviation or variance.

The program will use the data provided in any of these formats to compute the between-groups standard deviation and then proceed to compute f.

Entering the Effect Size (f) for Oneway ANOVA

- On the main screen, enter the *SD within cell* value (optional if the user will enter f directly).

Figure 16.2 Oneway ANOVA main screen



Factor Name	Number of levels	Cases per level	Effect size f	Power
Factor A	2	5	0.00	0.10

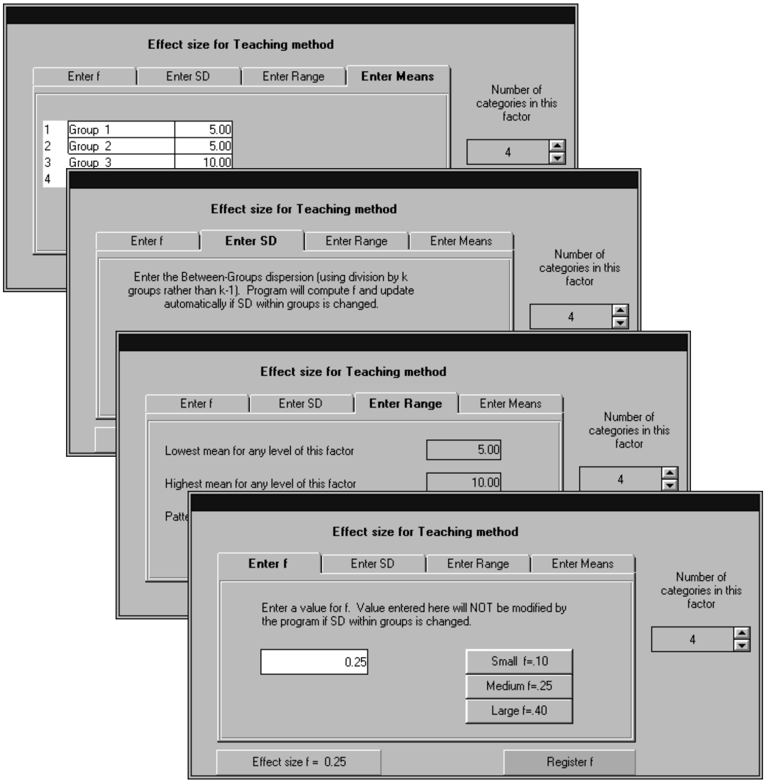
SD within cell 10.00 N of cases per cell 5

Variance within cell 100.00 Total N 10

Alpha= 0.10

- Click on the effect size shown (initially, this is 0.00). The program will immediately transfer control to one of four panels (see Figure 16.3).

Figure 16.3 Effect size panels



Effect size for Teaching method

Number of categories in this factor 4

1	Group 1	5.00
2	Group 2	5.00
3	Group 3	10.00
4		

Effect size for Teaching method

Enter the Between-Groups dispersion (using division by k groups rather than k-1). Program will compute f and update automatically if SD within groups is changed.

Number of categories in this factor 4

Effect size for Teaching method

Lowest mean for any level of this factor 5.00

Highest mean for any level of this factor 10.00

Patt

Number of categories in this factor 4

Effect size for Teaching method

Enter a value for f. Value entered here will NOT be modified by the program if SD within groups is changed.

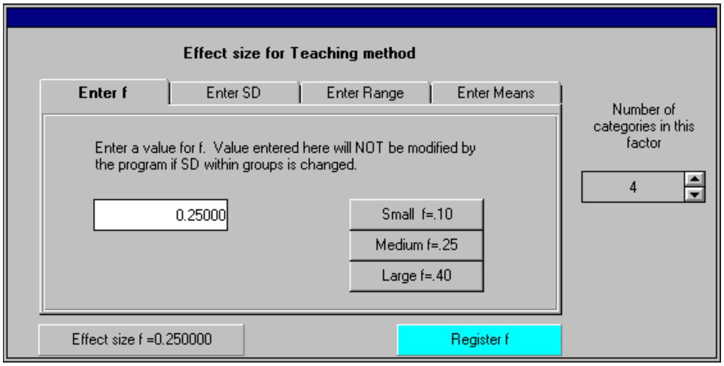
0.25

Number of categories in this factor 4

Effect size f = 0.25

- On the panel, enter the number of levels and the effect size.

Figure 16.4 Panel to enter f directly or using conventional values



Effect size for Teaching method

Enter a value for f. Value entered here will NOT be modified by the program if SD within groups is changed.

0.25000

Number of categories in this factor: 4

Effect size f =0.250000

This panel is intended for users who are familiar with the effect size, f, and feel comfortable specifying the f value directly. It is also appropriate for users who have little basis for estimating the effect size and therefore prefer to work with Cohen's conventions for small, medium, or large effects.

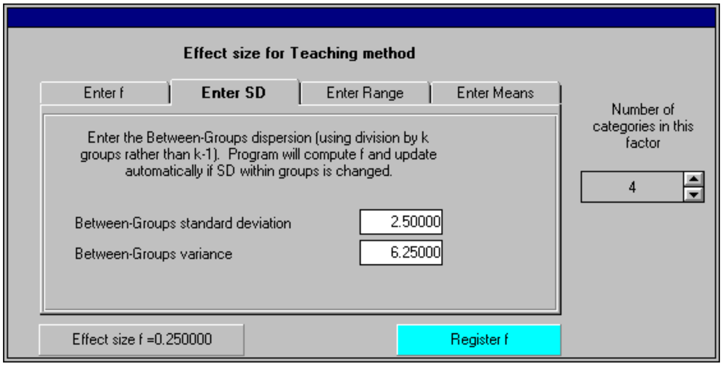
- Enter the number of groups.

- Click on one of the conventional values for effect size or enter a value directly. For example, a value of 0.25 would fall in the medium range according to Cohen's conventions.

- Click *Compute f* to compute the corresponding f.

- Click *Register f* to transfer the value to the main screen.

Figure 16.5 Panel to enter between-groups standard deviation



Effect size for Teaching method

Enter the Between-Groups dispersion (using division by k groups rather than k-1). Program will compute f and update automatically if SD within groups is changed.

Between-Groups standard deviation: 2.50000

Between-Groups variance: 6.25000

Number of categories in this factor: 4

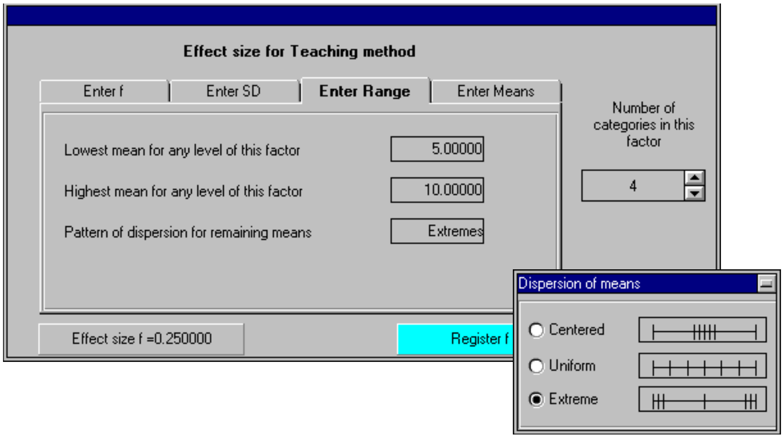
Effect size f =0.250000

This panel is appropriate for researchers who are able to provide an estimate of the between-groups dispersion.

- Enter the number of groups.
- Enter the between-groups dispersion (either the standard deviation or the variance).
- Click *Compute f* to compute the corresponding f.
- Click *Register f* to transfer the value to the main screen.

In this example, the user entered the between-groups standard deviation (2.5) and the between-groups variance (6.25). Using this information and the within-cells standard deviation (10, entered on the main screen), the program computes the effect size, f (in this case, $2.5/10 = 0.25$).

Figure 16.6 Panel to enter the range of group means



Lowest mean for any level of this factor 5.00000

Highest mean for any level of this factor 10.00000

Pattern of dispersion for remaining means Extremes

Number of categories in this factor 4

Effect size f =0.250000

○ Centered
○ Uniform
● Extreme

This panel is appropriate for researchers who are able to estimate the range of means but not the mean for each group.

- Enter the number of groups.
- Enter the single lowest and highest means.
- Once the two extreme groups have been specified, the remaining groups can fall into one of three patterns: all remaining groups fall at the center of the range (*Centered*), which will yield the smallest effect; the remaining groups are distributed evenly across the range

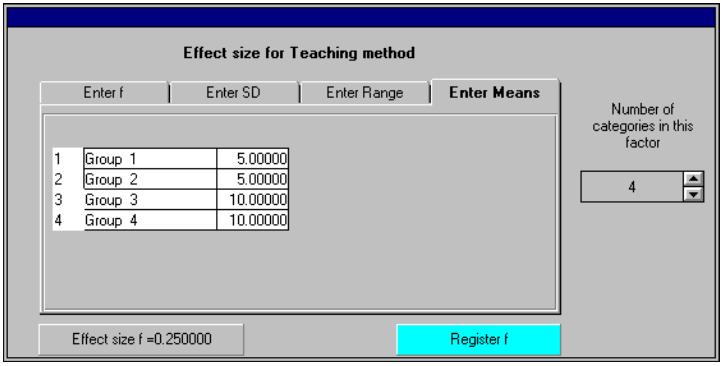
(*Uniform*); or the remaining groups fall at either extreme (*Extreme*). Click on the value shown to activate the pop-up box and make a selection.

- Click *Compute f* to compute the corresponding f.
- Click *Register f* to transfer the value to the main screen.

Note that when the study has only two groups, the three patterns are identical and, thus, will have no effect on f.

On the main screen, the user specified that the within-cells standard deviation is 10. In Figure 16.6, the user specified that the four groups will have means that range from 5 to 10 and that the remaining two groups will have means at either extreme of this range (that is, at 5 and 10). The program has computed the corresponding f as 0.25.

Figure 16.7 Panel to enter the mean for each group



Effect size for Teaching method

1	Group 1	5.00000
2	Group 2	5.00000
3	Group 3	10.00000
4	Group 4	10.00000

Number of categories in this factor: 4

Effect size f =0.250000

This panel is appropriate for researchers who are able to provide an estimate for every one of the group means.

- Enter the number of groups. The program will display an entry box for each group.
- Enter a name (optional) and mean (required) for each group.
- Click *Compute f* to compute the corresponding f.
- Click *Register f* to transfer the value to the main screen.

In this example, the user specified mean values of 5, 5, 10, and 10. These values provide the between-groups standard deviation (2.5), and the user has already provided the within-cells standard deviation (10). The program computes the effect size, f, as 0.25.

Correspondence between the Four Approaches

The four methods provided for computing f are mathematically equivalent to each other. In the example, the user is working with means of 5, 5, 10, and 10. The dispersion can be described by specifying these four distinct values. It can also be described by entering the range (5 to 10) and the pattern (extreme) of the two remaining means. Finally, it can be described by entering the standard deviation for the four means (2.5).

Each of these approaches yields the identical value for the between-groups standard deviation (2.5). Given the within-cells standard deviation (10) from the main screen, all yield the identical value for f (0.25).

Of course, the different methods yield the same value only when provided with equivalent information. Often, researchers will want to estimate f using more than one method as a check that the estimate is accurate. The program will retain the data entered in any panel, but only the registered data will be transferred to the main screen.

Effect Size Updated Automatically

When you initially enter the between-groups standard deviation (either by entering the between-groups standard deviation, the range of means, or the four individual means), the program will compute f using this value and the within-cells standard deviation (which is entered on the main screen). If you later modify the within-cells standard deviation, the effect size will be adjusted automatically to reflect this change (just as the effect size, d, is modified in a t-test when you update the within-groups standard deviation).

However, if you have decided to enter the effect size by specifying f directly, the effect size is not updated. The program assumes that a user who has specified a medium effect (0.25) may not know whether the between-groups/within-groups standard deviation is 0.25/1 or 2.5/10, but merely that a medium effect would be important to detect. In this case, the value entered for the within-cells standard deviation has no impact on f.

Alpha

Click *Alpha* to modify this value. ANOVA is sensitive to an effect in any direction and as such is nondirectional.

Sample Size

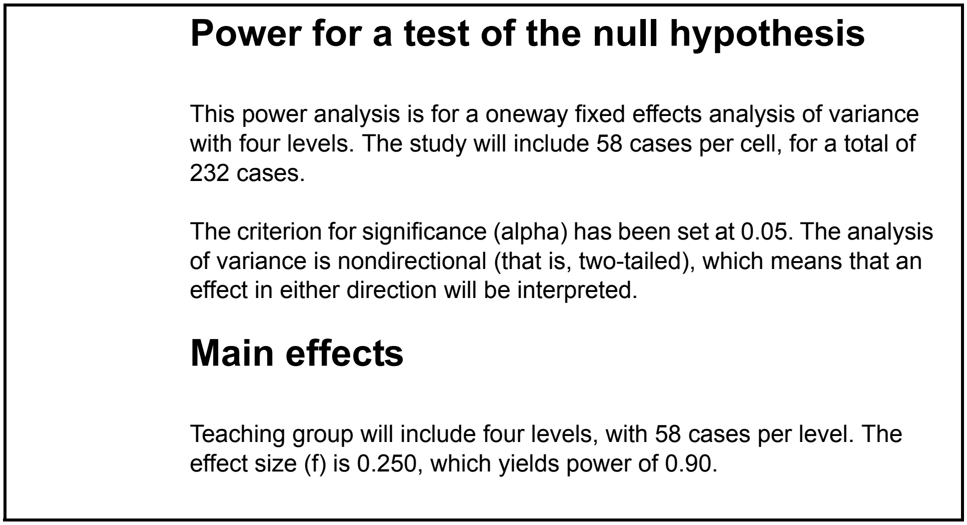
To modify the sample size, use the spin control. By default, this will increase the number of cases by 5, but this can be modified to any number (choose *N-Cases* from the Options menu).

Click the *Find N* icon to have the program find the number of cases required for the default level of power. The default value for power is 90%, but this can be modified temporarily (Ctrl-F) or permanently (choose *Preferences* from the Options menu).

Example 1

Students are assigned at random to one of four teaching groups, taught for a period of time, and then assessed on a continuous measure. A substantive effect would be a medium effect of f = 0.25.

- Choose *Oneway analysis of variance*.
- Enter the name *Teaching group.*
- Click on the value shown for effect size (0.00).
- Activate the *Enter f* tab. Set the number of groups (4) and effect size (medium).
- (Alternatively, enter the within-cells standard deviation on main screen and the between-groups standard deviation using the alternative panel.)
- Click *Compute f* and then click *Register f* to return to the main screen.
- Click *Alpha* and select the following value: *Alpha = 0.05*.
- Click the *Find N* icon (or press Ctrl-F). The program shows that 58 cases per cell will yield power of 90%.
- Click the *Report* icon to generate the following report:



Power for a test of the null hypothesis

This power analysis is for a oneway fixed effects analysis of variance with four levels. The study will include 58 cases per cell, for a total of 232 cases.

The criterion for significance (alpha) has been set at 0.05. The analysis of variance is nondirectional (that is, two-tailed), which means that an effect in either direction will be interpreted.

Main effects

Teaching group will include four levels, with 58 cases per level. The effect size (f) is 0.250, which yields power of 0.90.

Oneway Analysis of Covariance

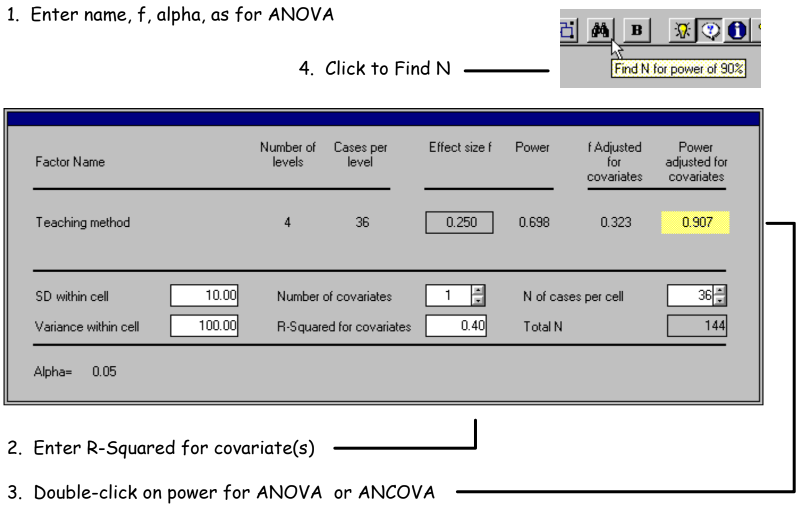
Analysis of covariance is identical to analysis of variance except for the presence of the covariate. The covariates are able to explain some of the variance in the outcome measure, and by taking this into account, we reduce the error term, which yields a more powerful test.

We will illustrate the use of this procedure by extending the previous example. As in Example 1, students are assigned at random to one of three training groups, taught for a period of time, and then assessed on a continuous measure. In the ANOVA example, we simply assessed the students at the end point. For the ANCOVA example, we will assume that the students are assessed at the baseline as well and that the baseline ratings serve as a covariate.

The specification of effect size by the user proceeds exactly as that for ANOVA. Specifically, the within-groups standard deviation and the between-groups standard deviation are entered as though no covariate is present. If the user chooses to enter f directly, enter f as though no covariate is present. The covariate serves to reduce the error term (which boosts the effect size, f), but this is taken into account by the program, which displays both the unadjusted f and the adjusted f.

Figure 16.8 shows the oneway analysis of covariance. We assume that the baseline scores account for 40% of the scores at the end point.

Figure 16.8 Oneway analysis of covariance



1. Enter name, f, alpha, as for ANOVA

4. Click to Find N

Factor Name	Number of levels	Cases per level	Effect size f	Power	f Adjusted for covariates	Power adjusted for covariates
Teaching method	4	36	0.250	0.698	0.323	0.907

SD within cell	10.00	Number of covariates	1	N of cases per cell	36
Variance within cell	100.00	R-Squared for covariates	0.40	Total N	144

Alpha= 0.05

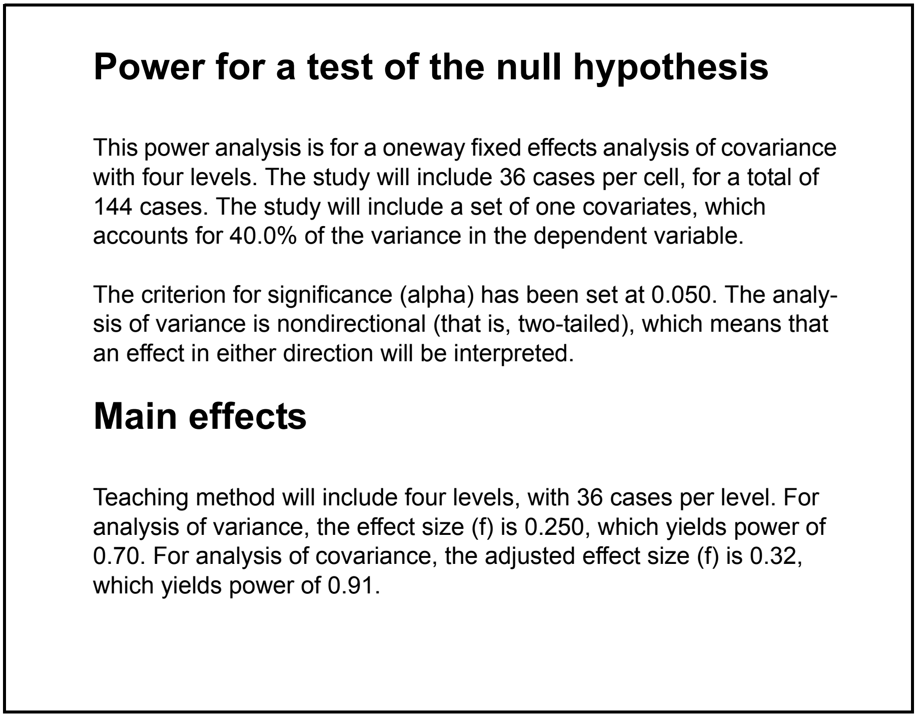
2. Enter R-Squared for covariate(s)

3. Double-click on power for ANOVA or ANCOVA

Example 2

Students are assigned at random to one of four teaching groups, taught for a period of time, and then assessed on a continuous measure. A substantive effect would be a medium effect of f = 0.25. The pre-score will serve as a covariate and is expected to account for 40% of the variance in post-scores.

- Choose *Oneway analysis of covariance*.
- Enter the name *Teaching method*.
- Click on the value shown for effect size (0.00).
- Activate the *Enter f* tab. Set the number of groups (4) and effect size (medium).
- (Alternatively, enter the within-cells standard deviation on main screen and the between-groups standard deviation using the alternative panel.)
- Click *Compute f* and then click *Register f* to return to the main screen.
- Click *Alpha* and select the following value: *Alpha = 0.05*.
- Double-click the *Power adjusted for covariates* value. This tells the program to find, in the next step, the required number of cases for ANCOVA (alternatively, double-click on the power shown for ANOVA to find the required number of cases for ANOVA).
- Press Ctrl-F (or click the *Find N* icon if the default power has already been set to 90). The program shows that 36 cases per cell (as compared with 58 for ANOVA) will yield power of 90%.
- Click the *Report* icon to generate the following report:



Power for a test of the null hypothesis

This power analysis is for a oneway fixed effects analysis of covariance with four levels. The study will include 36 cases per cell, for a total of 144 cases. The study will include a set of one covariates, which accounts for 40.0% of the variance in the dependent variable.

The criterion for significance (alpha) has been set at 0.050. The analysis of variance is nondirectional (that is, two-tailed), which means that an effect in either direction will be interpreted.

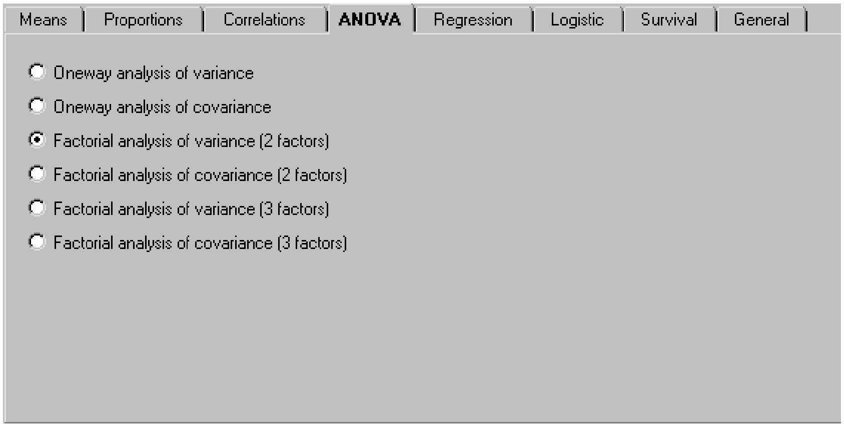
Main effects

Teaching method will include four levels, with 36 cases per level. For analysis of variance, the effect size (f) is 0.250, which yields power of 0.70. For analysis of covariance, the adjusted effect size (f) is 0.32, which yields power of 0.91.

17 Analysis of Variance/Covariance (Factorial)

Selecting the Procedure

To display the available procedures, choose *New analysis* from the File menu.



○ Oneway analysis of variance

○ Oneway analysis of covariance

● Factorial analysis of variance (2 factors)

○ Factorial analysis of covariance (2 factors)

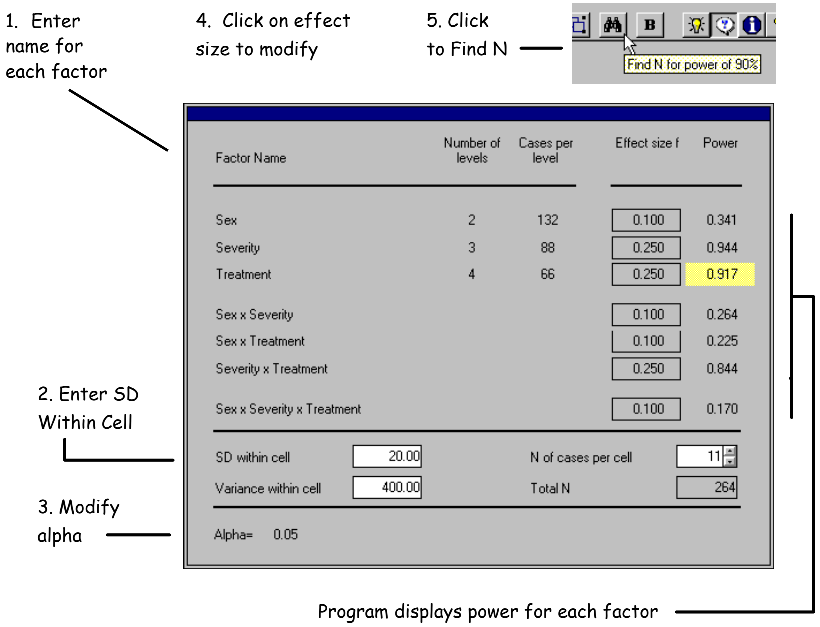
○ Factorial analysis of variance (3 factors)

○ Factorial analysis of covariance (3 factors)

Application

The program can compute power for a fixed-effect balanced factorial analysis of variance with two or three factors, including interactions.

Figure 17.1 Factorial analysis of variance (three factors)



1. Enter name for each factor

4. Click on effect size to modify

5. Click to Find N

Factor Name	Number of levels	Cases per level	Effect size f	Power
Sex	2	132	0.100	0.341
Severity	3	88	0.250	0.944
Treatment	4	66	0.250	0.917
Sex x Severity			0.100	0.264
Sex x Treatment			0.100	0.225
Severity x Treatment			0.250	0.844
Sex x Severity x Treatment			0.100	0.170

SD within cell 20.00 N of cases per cell 11

Variance within cell 400.00 Total N 264

Alpha= 0.05

2. Enter SD Within Cell

3. Modify alpha

Program displays power for each factor

Effect Size

The effect size (f) used in analysis of variance is an extension of the effect size (d) used for a t-test. Recall that d is the mean difference between groups divided by the dispersion within groups. Similarly, f is based on the dispersion (standard deviation) between groups divided by the dispersion (standard deviation) within groups. (The effect size, f, is a pure measure of effect size and should not be confused with the *F* statistic, which takes into account sample size as well as effect size.)

In factorial ANOVA, the within-cells standard deviation is affected by the inclusion of multiple factors. For example, if the analysis includes *Gender* and *Treatment* as factors, the within-cells standard deviation is the within-gender, within-treatment standard deviation. The user should take this into account when estimating the within-cells standard deviation, which will have the effect of reducing within-cell variance and increasing the effect size.

The program allows the user to enter f for each factor directly or by using Cohen's conventions for research in the social sciences—small = 0.10, medium = 0.25, and large = 0.40—for main effects or interactions.

Alternatively, the user is allowed to provide data that the program uses to compute f. In this case, the program requires the user to provide the within-cells standard deviation on the main screen. Then, enter data for the between-groups standard deviation in one of three formats:

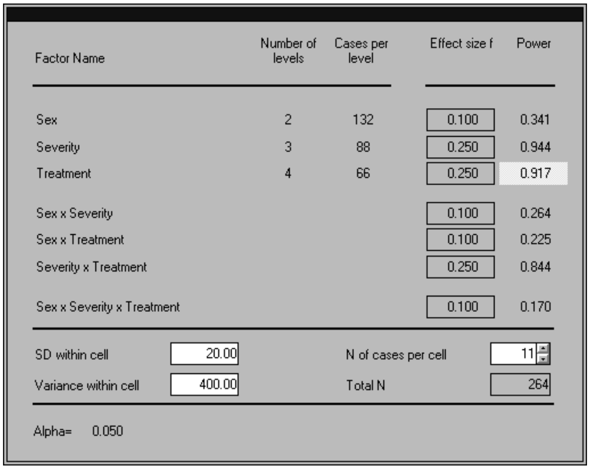
- Enter the between-groups standard deviation or variance (for main effects or interactions).
- Enter the range of means for the factor (main effects only).
- Enter the mean for each group (main effects only).

The program will use the data provided in any of these formats to compute the between-groups standard deviation and then proceed to compute f.

Entering the Effect Size (f) for Factorial ANOVA

❖ On the main screen, enter the *SD within cell* value (optional if the user will enter f directly).

Figure 17.2 Factorial ANOVA (three factors) main screen



Factor Name	Number of levels	Cases per level	Effect size f	Power
Sex	2	132	0.100	0.341
Severity	3	88	0.250	0.944
Treatment	4	66	0.250	0.917
Sex x Severity			0.100	0.264
Sex x Treatment			0.100	0.225
Severity x Treatment			0.250	0.844
Sex x Severity x Treatment			0.100	0.170

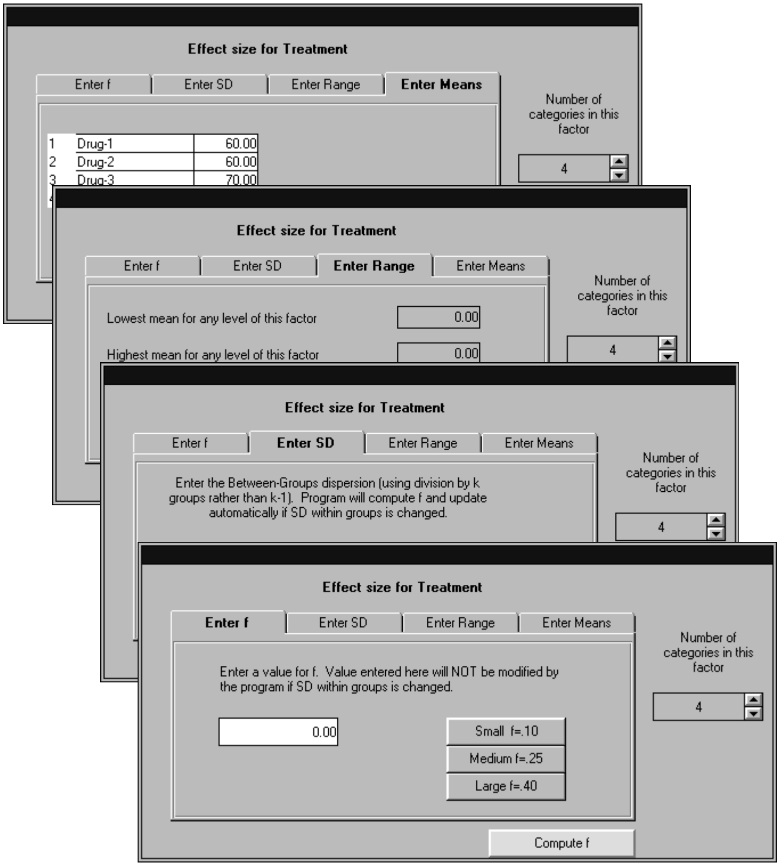
SD within cell 20.00 N of cases per cell 11

Variance within cell 400.00 Total N 264

Alpha= 0.050

❖ Click on the effect size shown for any of the main effects or interactions (initially, this is 0.000). The program will immediately transfer control to one of four panels (see Figure 17.3). Choose one panel and enter the value(s) for f, standard deviation, range, or means.

Figure 17.3 Effect size panels



Effect size for Treatment

1	Drug-1	60.00
2	Drug-2	60.00
3	Drug-3	70.00

Effect size for Treatment

Lowest mean for any level of this factor 0.00

Highest mean for any level of this factor 0.00

Effect size for Treatment

Enter the Between-Groups dispersion (using division by k groups rather than k-1). Program will compute f and update automatically if SD within groups is changed.

Effect size for Treatment

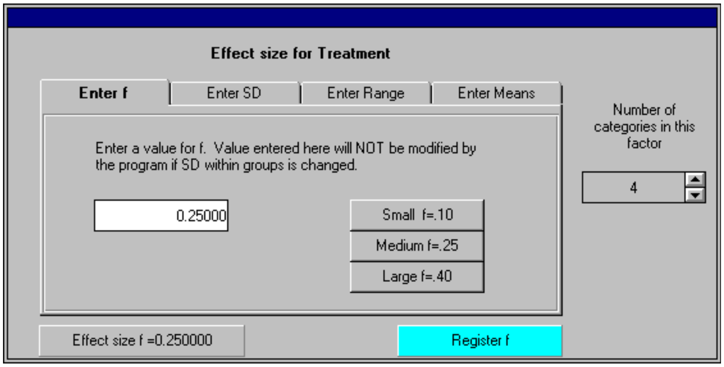
Enter a value for f. Value entered here will NOT be modified by the program if SD within groups is changed.

0.00

Number of categories in this factor 4

❖ Repeat this process for each factor. The method used to enter the effect size is set on a factor-by-factor basis.

Figure 17.4 Panel to enter f directly or using conventional values



Enter a value for f. Value entered here will NOT be modified by the program if SD within groups is changed.

0.25000

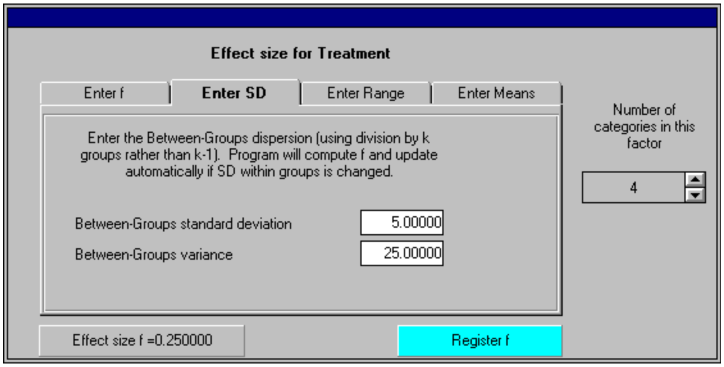
Number of categories in this factor: 4

Effect size f =0.250000

This panel is intended for users who are familiar with the effect size, f, and feel comfortable specifying the f value directly. It is also appropriate for users who have little basis for estimating the effect size and therefore prefer to work with Cohen's conventions for small, medium, or large effects.

- ❖ Enter the number of groups.
- ❖ Click on one of the conventional values for effect size or enter a value directly. For example, a value of 0.25 would fall in the medium range according to Cohen's conventions.
- ❖ Click *Compute f* to compute the corresponding f.
- ❖ Click *Register f* to transfer the value to the main screen.

Figure 17.5 Panel to enter the between-groups standard deviation


Effect size for Treatment
Enter f
Enter SD
Enter Range
Enter Means
Number of categories in this factor
Register f

Enter the Between-Groups dispersion (using division by k groups rather than k-1). Program will compute f and update automatically if SD within groups is changed.

Between-Groups standard deviation 5.00000

Between-Groups variance 25.00000

Number of categories in this factor: 4

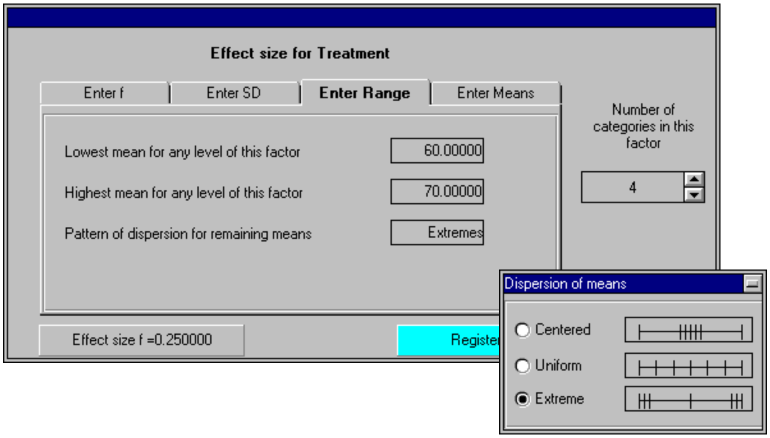
Effect size f =0.250000

This panel is appropriate for researchers who are able to provide an estimate of the beween-groups dispersion.

- Enter the number of groups.
- Enter the between-groups dispersion (either the standard deviation or the variance).
- Click *Compute f* to compute the corresponding f.
- Click *Register f* to transfer the value to the main screen.

In this example, the user entered the between-groups standard deviation (5) and the between-groups variance (25). Using this information and the within-cells standard deviation (20, entered on the main screen), the program computes the effect size, f (in this case, $5/20 = 0.25$).

Figure 17.6 Panel to enter the range of group means



Lowest mean for any level of this factor 60.00000

Highest mean for any level of this factor 70.00000

Pattern of dispersion for remaining means Extremes

Number of categories in this factor 4

Effect size f =0.250000

○ Centered

○ Uniform

● Extreme

This panel is appropriate for researchers who are able to estimate the range of means but not the mean for each group.

- Enter the number of groups.
- Enter the single lowest and highest means.
- Once the two extreme groups have been specified, the remaining groups can fall into one of three patterns: all remaining groups fall at the center of the range (*Centered*), which will yield the smallest effect; the remaining groups are distributed evenly across the

range (*Uniform*); or the remaining groups fall at either extreme (*Extreme*). Click on the value shown to activate the pop-up box and make a selection.

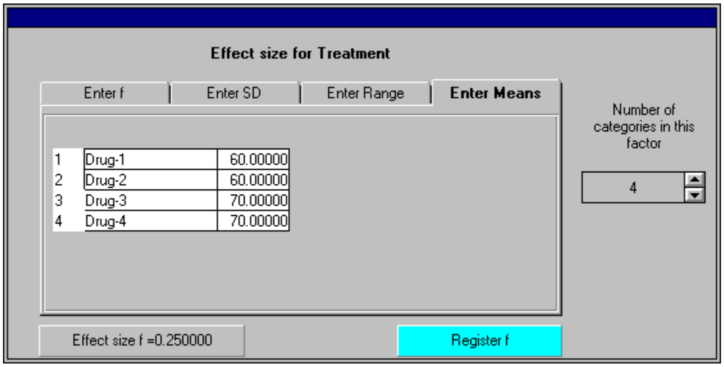
❖ Click *Compute f* to compute the corresponding f.

❖ Click *Register f* to transfer the value to the main screen.

Note that when the study has only two groups, the three patterns are identical and, thus, will have no effect on f.

On the main screen, the user specified that the within-groups standard deviation is 20. In Figure 17.6, the user specified that the four groups will have means that range from 60 to 70 and that the remaining two groups will have means at either extreme of this range (that is, at 60 and 70). The program has computed the corresponding f as 0.25.

Figure 17.7 Panel to enter the mean for each group



Effect size for Treatment

1	Drug-1	60.00000
2	Drug-2	60.00000
3	Drug-3	70.00000
4	Drug-4	70.00000

Number of categories in this factor: 4

Effect size f =0.250000

This panel is appropriate for researchers who are able to provide an estimate for every one of the group means.

❖ Enter the number of groups. The program will display an entry box for each group.

❖ Enter a name (optional) and mean (required) for each group.

❖ Click *Compute f* to compute the corresponding f.

❖ Click *Register f* to transfer the value to the main screen.

In this example, the user specified mean values of 60, 60, 70, and 70. These values provide the between-groups standard deviation (5), and the user has already provided the within-cells standard deviation (20). The program computes the effect size, f, as 0.25.

Correspondence between the Four Approaches

The four methods provided for computing f are mathematically equivalent to each other. In the example, the user is working with means of 60, 60, 70, and 70. The dispersion can be described by specifying these four distinct values. It can also be described by entering the range (60 to 70) and the pattern (extreme) of the two remaining means. Finally, it can be described by entering the standard deviation for the four means (5).

Each of these approaches yields the identical value for the between-groups standard deviation (5). Given the within-cells standard deviation (20) from the main screen, all yield the identical value for f (0.25).

Of course, the different methods yield the same value only when provided with equivalent information. Often, researchers will want to estimate f using more than one method as a check that the estimate is accurate. The program will retain the data entered in any panel, but only the registered data will be transferred to the main screen.

Effect Size Updated Automatically

When you initially enter the between-groups standard deviation (either by entering the between-groups standard deviation, the range of means, or the four individual means), the program will compute f using this value and the within-cells standard deviation (which is entered on the main screen). If you later modify the within-cells standard deviation, the effect size will be adjusted automatically to reflect this change (just as the effect size, d, is modified in a t-test as you update the within-groups standard deviation).

However, if you have decided to enter the effect size by specifying f directly, the effect size is not updated. The program assumes that a user who has specified a medium effect (0.25) may not know whether the between-groups/within-groups standard deviation is 0.25/1 or 2.5/10, but merely that a medium effect would be important to detect. In this case, the value entered for the within-cells standard deviation has no impact on f.

This is handled on a factor-by-factor basis. If the effect size for *Sex* is entered directly as 0.10 while the effect size for *Treatment* is entered by providing the between-groups standard deviation, changes to the within-groups standard deviation will not affect the former but will affect the latter. All effect sizes are displayed on the main screen, and the user should check these values after making changes to the within-groups standard deviation.

Alpha

Click *Alpha* to modify this value. ANOVA is sensitive to an effect in any direction and as such is nondirectional.

Sample Size

To modify the sample size, use the spin control. By default, this will increase the number of cases by 5, but this can be modified to any number (choose *N-Cases* from the Options menu).

Click the *Find N* icon to have the program find the number of cases required for the default level of power. The default value for power is 90%, but this can be modified temporarily (Ctrl-F) or permanently (choose *Preferences* from the Options menu).

Note. When you click the *Find N* icon, the program assumes that you want to work with the first factor (that is, to find the number of cases required to yield adequate power for Factor A). To select another factor, double-click on the power shown for that factor (the selected value will be highlighted) and click the *Find N* icon again.

Example 1

A researcher plans to assign patients to one of four treatments regimens. Following a period of treatment, she will measure the patients' level of antibodies. The analysis will take into account patient gender (two levels), disease severity at baseline (three levels), and treatment (four levels). The researcher is primarily interested in the effect of treatment and in the treatment-by-severity interaction. A clinically important effect for either would be a medium effect of f = 0.25.

❖ Choose *Factorial analysis of variance (3 factors)*.

❖ Enter a name for each factor—*Sex*, *Severity*, and *Treatment*. Names for the interactions are assigned automatically.

❖ Click on the effect size for the first factor (*Sex*). Specify two levels with f = 0.10.

❖ Click on the effect size for the second factor (*Severity*). Specify three levels with f = 0.25.

❖ Click on the effect size for the third factor (*Treatment*). Specify four levels with f = 0.25.

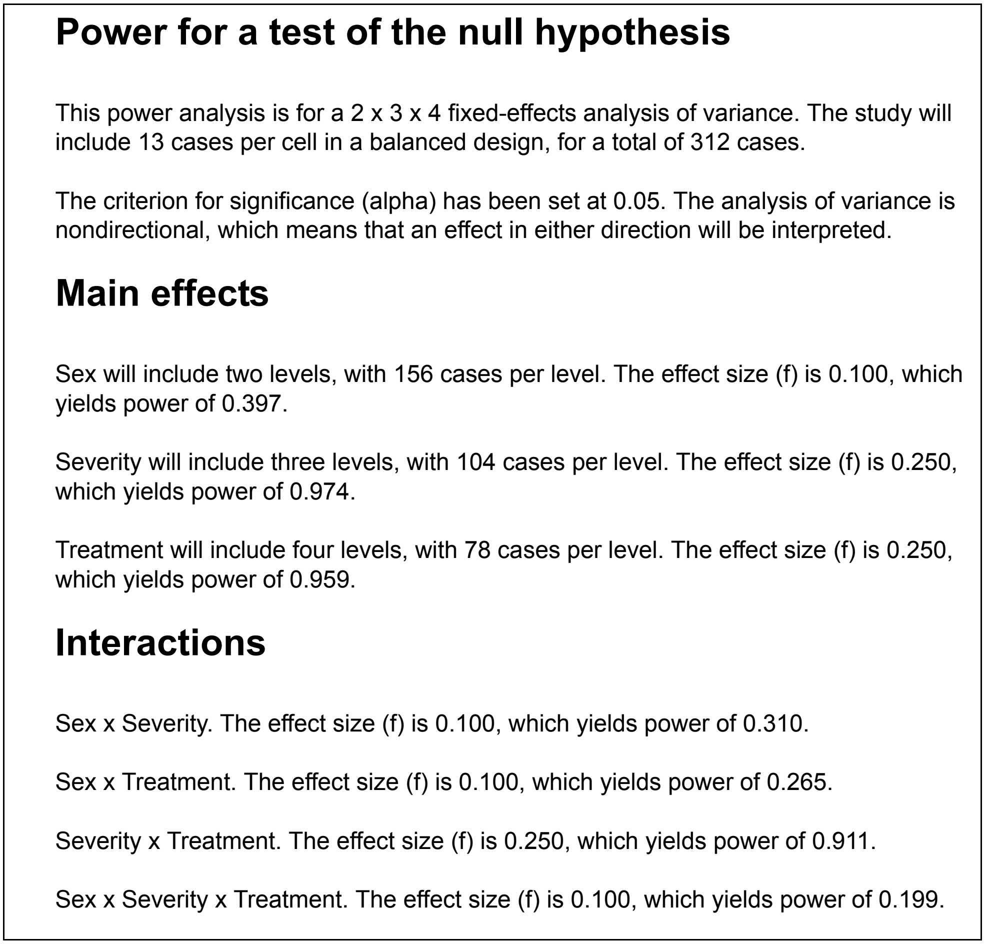
❖ Click on the effect size for each of the interactions. In this example, the effect size is set at 0.25 for the interaction of *Severity* by *Treatment* and at 0.10 for all other interactions.

❖ Click *Alpha* and select the following value: *Alpha = 0.05*.

- Double-click on the power shown for *Treatment*. Then press Ctrl-F (or click the *Find N* icon if the default power has already been set to 90). The program shows that 11 cases per cell (the total number of cases is 264) will yield power of 92% for *Treatment*, which is the designated factor. The power to find an effect for the interaction of *Treatment* by *Severity* is 84%.

- Double-click on the power shown for the interaction of *Severity* by *Treatment*. Then click the *Find N* icon. The program shows that 13 cases per cell (the total number of cases is 312) will yield power of 91% for this interaction. With this sample size, the power to find an effect for *Treatment* is 96%.

- Click the *Report* icon to generate the following report:



Power for a test of the null hypothesis

This power analysis is for a 2 x 3 x 4 fixed-effects analysis of variance. The study will include 13 cases per cell in a balanced design, for a total of 312 cases.

The criterion for significance (alpha) has been set at 0.05. The analysis of variance is nondirectional, which means that an effect in either direction will be interpreted.

Main effects

Sex will include two levels, with 156 cases per level. The effect size (f) is 0.100, which yields power of 0.397.

Severity will include three levels, with 104 cases per level. The effect size (f) is 0.250, which yields power of 0.974.

Treatment will include four levels, with 78 cases per level. The effect size (f) is 0.250, which yields power of 0.959.

Interactions

Sex x Severity. The effect size (f) is 0.100, which yields power of 0.310.

Sex x Treatment. The effect size (f) is 0.100, which yields power of 0.265.

Severity x Treatment. The effect size (f) is 0.250, which yields power of 0.911.

Sex x Severity x Treatment. The effect size (f) is 0.100, which yields power of 0.199.

Factorial Analysis of Covariance

Analysis of covariance is identical to analysis of variance except for the presence of the covariate. The covariates are able to explain some of the variance in the outcome measure. This serves to reduce the error term, which yields a more powerful test.

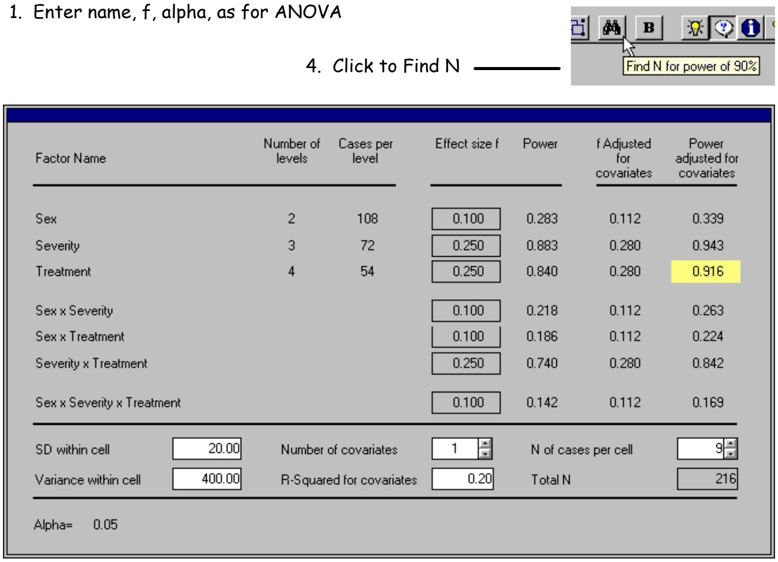
We will illustrate the use of this procedure by extending the previous example. As in Example 1, patients are assigned at random to one of four treatment groups, treated for a period of time, and then assessed on a continuous measure. In the ANOVA example, we simply assessed the level of antibodies at the end point. For the ANCOVA example, we will assume that the patients are assessed at the baseline as well and that the baseline ratings serve as a covariate.

The specification of effect size by the user proceeds exactly as that for ANOVA. Specifically, the within-groups standard deviation and the between-groups standard deviation are entered as though no covariate is present. If the user chooses to enter f directly, enter f as though no covariate is present. The covariate serves to reduce the error term (which boosts the effect size, f), but this is taken into account by the program, which displays both the unadjusted f and the adjusted f.

Figure 17.8 shows the factorial analysis of covariance. We assume that the baseline scores account for 40% of the scores at the end point.

Figure 17.8 Factorial analysis of covariance (three factors)

1. Enter name, f, alpha, as for ANOVA



4. Click to Find N

Factor Name	Number of levels	Cases per level	Effect size f	Power	f Adjusted for covariates	Power adjusted for covariates
Sex	2	108	0.100	0.283	0.112	0.339
Severity	3	72	0.250	0.883	0.280	0.943
Treatment	4	54	0.250	0.840	0.280	0.916
Sex x Severity			0.100	0.218	0.112	0.263
Sex x Treatment			0.100	0.186	0.112	0.224
Severity x Treatment			0.250	0.740	0.280	0.842
Sex x Severity x Treatment			0.100	0.142	0.112	0.169

SD within cell	20.00	Number of covariates	1	N of cases per cell	9
Variance within cell	400.00	R-Squared for covariates	0.20	Total N	216

Alpha= 0.05

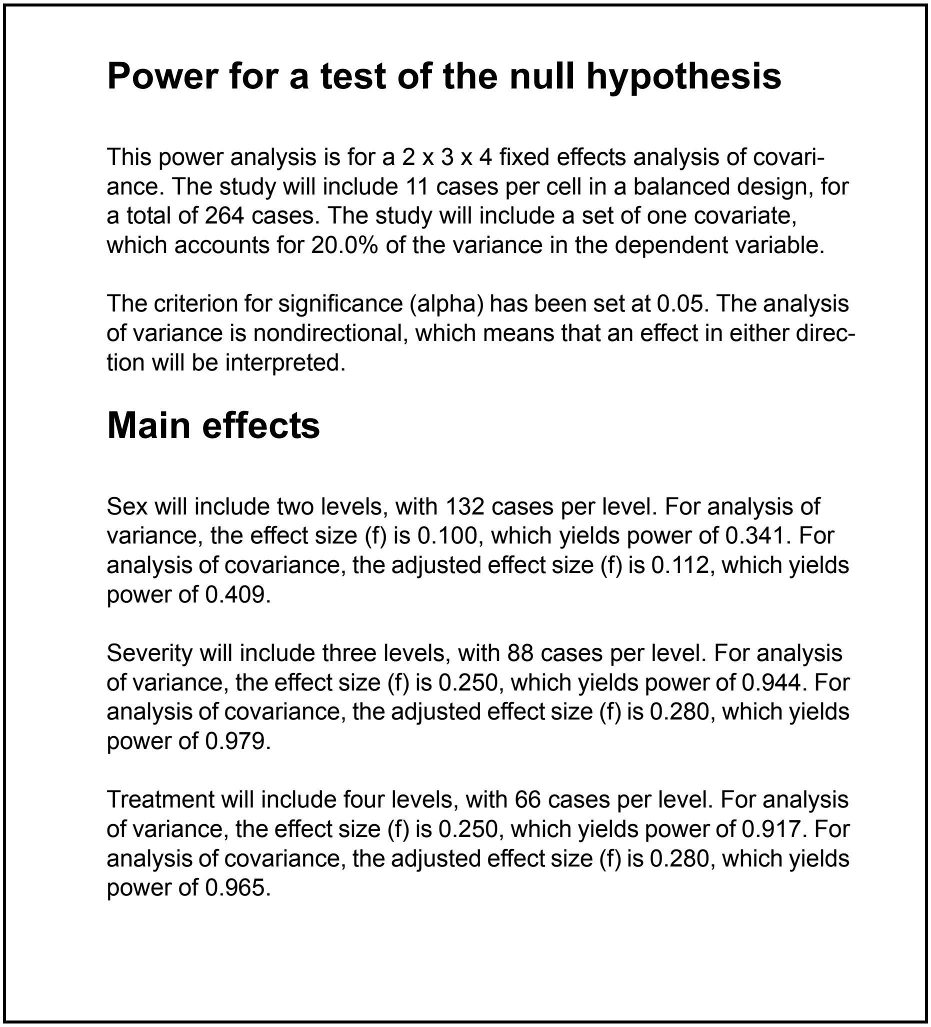
2. Enter R-Squared for covariate(s)

3. Double-click on power for ANOVA or ANCOVA

Example 2

A researcher plans to assign patients to one of four treatments regimens. Following a period of treatment, she will measure the patients' level of antibodies. The analysis will take into account patient gender (two levels), disease severity at baseline (three levels), and treatment (four levels). The researcher is primarily interested in the effect of treatment and in the treatment-by-severity interaction. A clinically important effect for either would be a medium effect of f = 0.25. The pre-score will serve as a covariate and is expected to account for 20% of the variance in post-scores.

- Choose *Factorial analysis of covariance (3 factors)*.
- Enter a name for each factor—*Sex*, *Severity*, and *Treatment*. Names for the interactions are assigned automatically.
- Click on the effect size for the first factor (*Sex*). Specify two levels with f = 0.10.
- Click on the effect size for the second factor (*Severity*). Specify three levels with f = 0.25.
- Click on the effect size for the third factor (*Treatment*). Specify four levels with f = 0.25.
- Click on the effect size for each of the interactions. In this example, the effect size is set at 0.25 for the interaction of *Severity* by *Treatment* and at 0.10 for all other interactions.
- Click *Alpha* and select the following value: *Alpha = 0.05*.
- On the main screen, enter the number of covariates (1) and the *R*-squared for covariates (0.20).
- Double-click the *Power adjusted for covariates* value in the *Treatment* row. Press Ctrl-F (or click the *Find N* icon if the default power has already been set to 90). The program shows that 9 cases per cell (the total number of cases is 216) will yield a power of 92% for *Treatment*, which is the designated factor. The power to find an effect for the interaction of *Treatment* by *Severity* is 84%.
- Double-click on the power shown for the interaction of *Severity* by *Treatment*. Then click the *Find N* icon. The program shows that 11 cases per cell (the total number of cases is 264) will yield a power of 92% for this interaction. With this sample size, power to find an effect for *Treatment* is 97%.
- Click the *Report* icon to generate the following report:



Power for a test of the null hypothesis

This power analysis is for a 2 x 3 x 4 fixed effects analysis of covariance. The study will include 11 cases per cell in a balanced design, for a total of 264 cases. The study will include a set of one covariate, which accounts for 20.0% of the variance in the dependent variable.

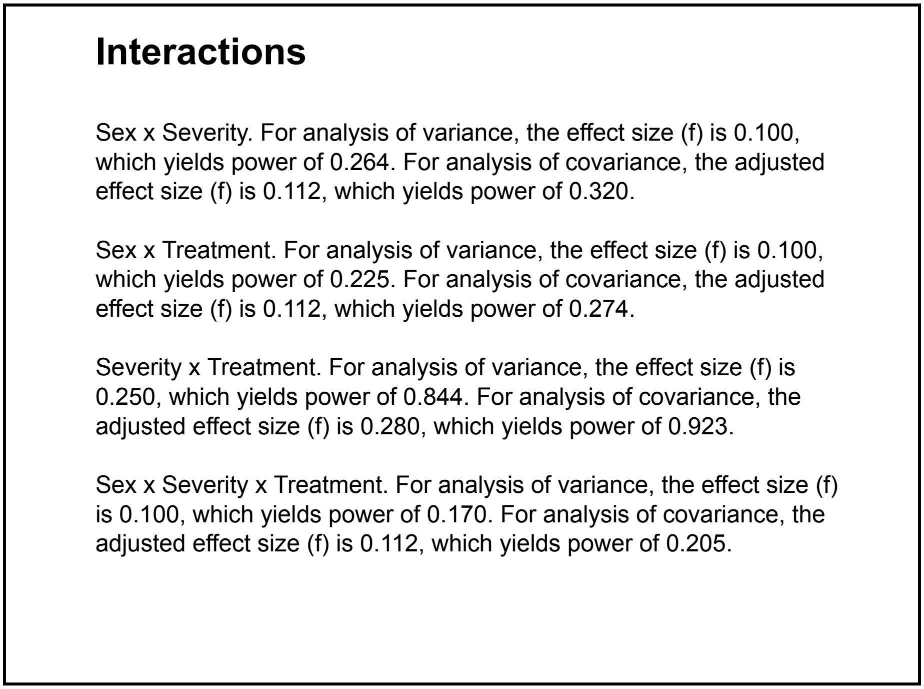
The criterion for significance (alpha) has been set at 0.05. The analysis of variance is nondirectional, which means that an effect in either direction will be interpreted.

Main effects

Sex will include two levels, with 132 cases per level. For analysis of variance, the effect size (f) is 0.100, which yields power of 0.341. For analysis of covariance, the adjusted effect size (f) is 0.112, which yields power of 0.409.

Severity will include three levels, with 88 cases per level. For analysis of variance, the effect size (f) is 0.250, which yields power of 0.944. For analysis of covariance, the adjusted effect size (f) is 0.280, which yields power of 0.979.

Treatment will include four levels, with 66 cases per level. For analysis of variance, the effect size (f) is 0.250, which yields power of 0.917. For analysis of covariance, the adjusted effect size (f) is 0.280, which yields power of 0.965.



Interactions

Sex x Severity. For analysis of variance, the effect size (f) is 0.100, which yields power of 0.264. For analysis of covariance, the adjusted effect size (f) is 0.112, which yields power of 0.320.

Sex x Treatment. For analysis of variance, the effect size (f) is 0.100, which yields power of 0.225. For analysis of covariance, the adjusted effect size (f) is 0.112, which yields power of 0.274.

Severity x Treatment. For analysis of variance, the effect size (f) is 0.250, which yields power of 0.844. For analysis of covariance, the adjusted effect size (f) is 0.280, which yields power of 0.923.

Sex x Severity x Treatment. For analysis of variance, the effect size (f) is 0.100, which yields power of 0.170. For analysis of covariance, the adjusted effect size (f) is 0.112, which yields power of 0.205.

Generate Table

To generate a table, click the *Table* icon. For a two-factor ANOVA, the table will show power for Factor A, Factor B, and the interaction as a function of sample size. For a three-factor ANOVA, the table will show power for Factor A, Factor B, and Factor C as a function of effect size. The table will not include power for interactions. In either case, the program can include alpha as a second factor as well.

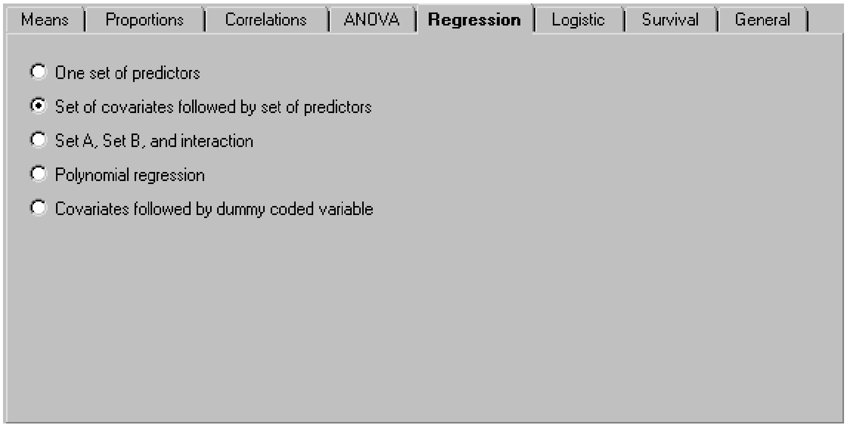
Generate Graph

To generate a graph, click the *Graph* icon. For a two-factor ANOVA, the program will graph power for Factor A, Factor B, and the interaction as a function of sample size. For a three-factor ANOVA, the program will graph power for Factor A, Factor B, and Factor C as a function of sample size. In either case, the program will graph power as a function of alpha for Factor A only.

18 Multiple Regression

Selecting the Procedure

To display the available procedures, choose *New analysis* from the File menu.



○ One set of predictors

● Set of covariates followed by set of predictors

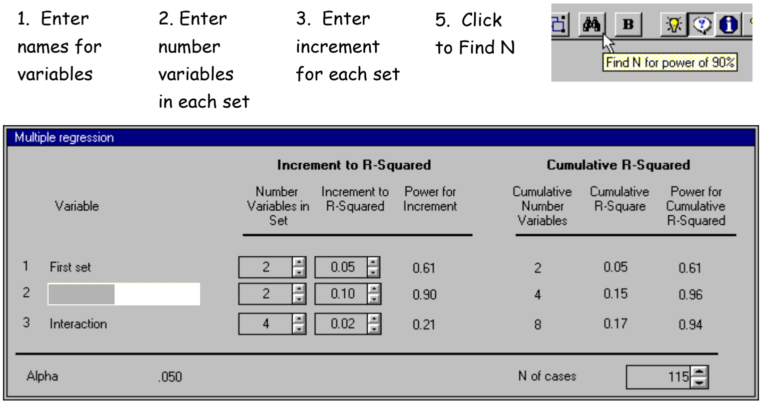
○ Set A, Set B, and interaction

○ Polynomial regression

○ Covariates followed by dummy coded variable

Application

Figure 18.1 Set A, set B, and interaction



1. Enter names for variables

2. Enter number variables in each set

3. Enter increment for each set

5. Click to Find N

Multiple regression

	Variable	Number Variables in Set	Increment to R-Squared	Power for Increment	Cumulative Number Variables	Cumulative R-Square	Power for Cumulative R-Squared
		Increment to R-Squared			Cumulative R-Squared		
1	First set	2	0.05	0.61	2	0.05	0.61
2		2	0.10	0.90	4	0.15	0.96
3	Interaction	4	0.02	0.21	8	0.17	0.94

Alpha .050

N of cases 115

4. Click to modify alpha

Program shows power for each increment at each step

Power shows power for cumulative R-SQ at each step

Multiple regression is used to study the relationship between sets of independent (predictor) variables and a single dependent variable. The “set” of variables is defined broadly and may consist of a single variable, such as age, or a set of variables that together serve as an index of some construct (for example, disease severity and duration together serve as an indicator of prognosis).

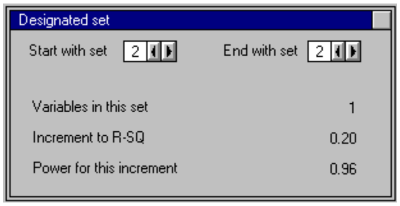
The variables may be continuous, in which case the actual rating (or some transformation thereof) is used as the predictor. The variables may also be categorical, in which case dummy coding (or one of its variants) is used to create a set of independent variables (sex is coded as 0 or 1).

Typically, sets of variables are entered into the multiple regression in a predetermined sequence. At each point in the analysis, the researcher may test the significance of the increment (the increase in R^2 for the new set over and above previous sets) or the significance of all variables in the equation. The program is able to compute power for either of these tests.

- At left, each set is identified by name
- The panel of columns labeled *Increment to R-Squared* shows data for the increment due to each set—the user provides the number of variables and the increment, and the program computes power.
- The panel of columns to the right shows data for the cumulative R^2. All data here (the cumulative number of variables, the cumulative R^2, and power) are computed automatically.

The Designated Set

The program allows the user to designate one set as the set of primary interest (choose *Customize screen* from the Options menu, and then select *Display superset*.)



Designated set

Start with set 2 End with set 2

Variables in this set 1

Increment to R-SQ 0.20

Power for this increment 0.96

- This set may be identical with a set in the main panel, or it may include two or more contiguous sets.
- The program will display power for the increment attributed to this set.
- The program's tools (Find N, Report, Table, and Graph) will provide data for this set.

Effect Size

Effect size for multiple regression is given by f^2, defined as explained variance/error variance. This is similar to the index (f) used for ANOVA, except that f is based on the standard deviations, while f^2 is based on the variances, as is common in multiple regression. Cohen (1988) notes that f^2 can be thought of as a kind of signal-to-noise ratio.

When used for a single set of variables, f^2 is equal to $R^2/(1-R^2)$. When used for more than one set of variables, R^2 in the numerator is the increment to R^2 for the current set, while R^2 in the denominator is the cumulative R^2 for all sets in the regression (but see "Error Model" on p. 148).

Cohen provides the following conventions for research in the social sciences: small ($f^2 = 0.02$), medium ($f^2 = 0.15$) and large ($f^2 = 0.35$). Assuming a single set of variables, these would correspond to R^2 values of about 0.02, 0.13, and 0.26.

Alpha and Tails

Click *Alpha* to modify it. Multiple regression is sensitive to effects in any direction and thus is naturally two-tailed.

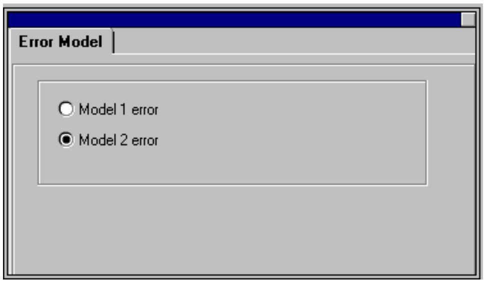
Sample Size

To modify the sample size, click the spin control for *N of cases*. To find the number of cases required to yield the required power, press Ctrl-F or click the *Find N* icon. The program will find the number of cases required to yield required power for the set defined in the Designated set box.

Computational Options for Power

Error Model

Multiple regression can be run with the model 1 error or the model 2 error (not to be confused with the type 1 error versus the type 2 error, or with one-tail versus two-tail). Choose *Computational formulas* from the Options menu.



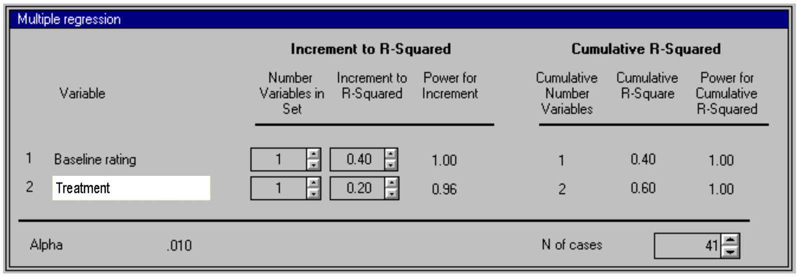
Error Model

○ Model 1 error

● Model 2 error

The two models differ in the definition of the error term used to compute the effect for each element in the model. In either model, F^2 is defined as R^2/Error. In model 1, error is defined as $1 - R^2$ through the current set. In model 2, error is defined as $1 - R^2$ for all variables in the model.

Assume, for example, that we will enter a set of covariates (*Baseline rating*), followed by a main set. We want to compute power for the covariates. The data are shown here.



Multiple regression

	Variable	Increment to R-Squared: Number Variables in Set	Increment to R-Squared: Increment to R-Squared	Increment to R-Squared: Power for Increment	Cumulative R-Squared: Cumulative Number Variables	Cumulative R-Squared: Cumulative R-Square	Cumulative R-Squared: Power for Cumulative R-Squared
1	Baseline rating	1	0.40	1.00	1	0.40	1.00
2	Treatment	1	0.20	0.96	2	0.60	1.00

Alpha .010

N of cases 41

With only the covariates in the equation, R^2 is 0.40, meaning that $(1 - 0.40)$, or 60% of the variance, is classified as error.

With both sets of variables in the model, the cumulative R^2 is 0.60, which means that the unexplained variance is $(1 - 0.60)$, or 40%.

The effect size for the covariates would be computed as $0.40/0.60 = 0.66$ under model 1, or as $0.40/0.40 = 1.00$ under model 2.

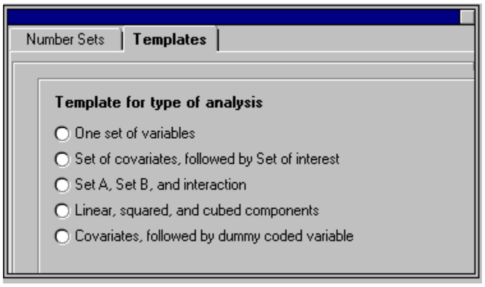
The effect size for set B is $0.20/0.40 = 0.50$ under either model, since set B is entered at the last step and all other variables are already in the equation. If there were additional sets to be entered subsequently to set B, then the selection of an error model would affect set B as well.

The default in this program (as in most statistical analysis programs) is the model 2 error. This yields higher power in most cases, but if the number of cases is low (so that small changes in the degrees of freedom are important) and/or subsequent sets incorporate many variables (thus consuming degrees of freedom), while yielding a small increment, it is possible that model 1 could yield a more powerful test.

Options for Study Design

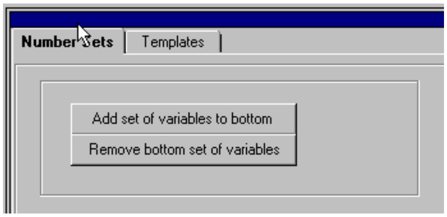
Templates

The program comes with templates corresponding to some of the more common types of analyses (choose *Data entry/Study design* from the Options menu).


Number Sets
Templates
Template for type of analysis
○ One set of variables
○ Set of covariates, followed by Set of interest
○ Set A, Set B, and interaction
○ Linear, squared, and cubed components
○ Covariates, followed by dummy coded variable

Customizing the Study Design

The templates cover some typical situations but are intended only as a starting point. The program allows the user to enter any number of sets (1 – 10), and each set can include as many as 99 variables (choose *Data entry/Study design* from the Options menu).


Number Sets
Templates
Add set of variables to bottom
Remove bottom set of variables

Generating a Table

Click the *Table* icon to generate a table of power by sample size. The table in multiple regression will show power for only one effect size (the designated effect). The table can be modified to show power as a function of alpha as well as sample size.

Generating a Graph

Click the *Graph* icon to generate a graph of power by sample size.

The graph in multiple regression will display only one effect size (the designated effect) but can be modified to show power as a function of sample size and alpha.

Example 1

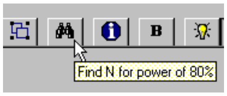
One Set of Variables

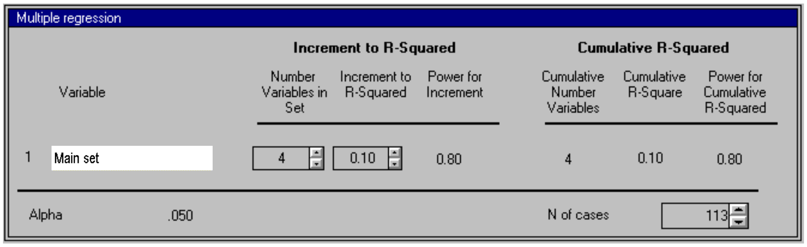
A researcher believes that aptitude, which she has operationally defined as scores on four tests, will predict students' grade point average (GPA). A meaningful effect is defined as one that explains 10% of the variance in the GPA. The study will be run with alpha at 0.05.

- ❖ On the Regression panel, select *One set of predictors*. (Or, within the module, choose *Data entry/Study design* from the Options menu, click the *Templates* tab, and select the template *One set of variables*.)
- ❖ Enter the number of variables—4.
- ❖ Enter the *Increment to R-Squared*—0.10.
- ❖ Click *Alpha* and select the value 0.05.

- Press Ctrl-F (or click the *Find N* icon if the default power has already been set to 80). The program shows that a sample of 113 cases will yield power of 80% for the given parameters.

Find N required for power of 80%



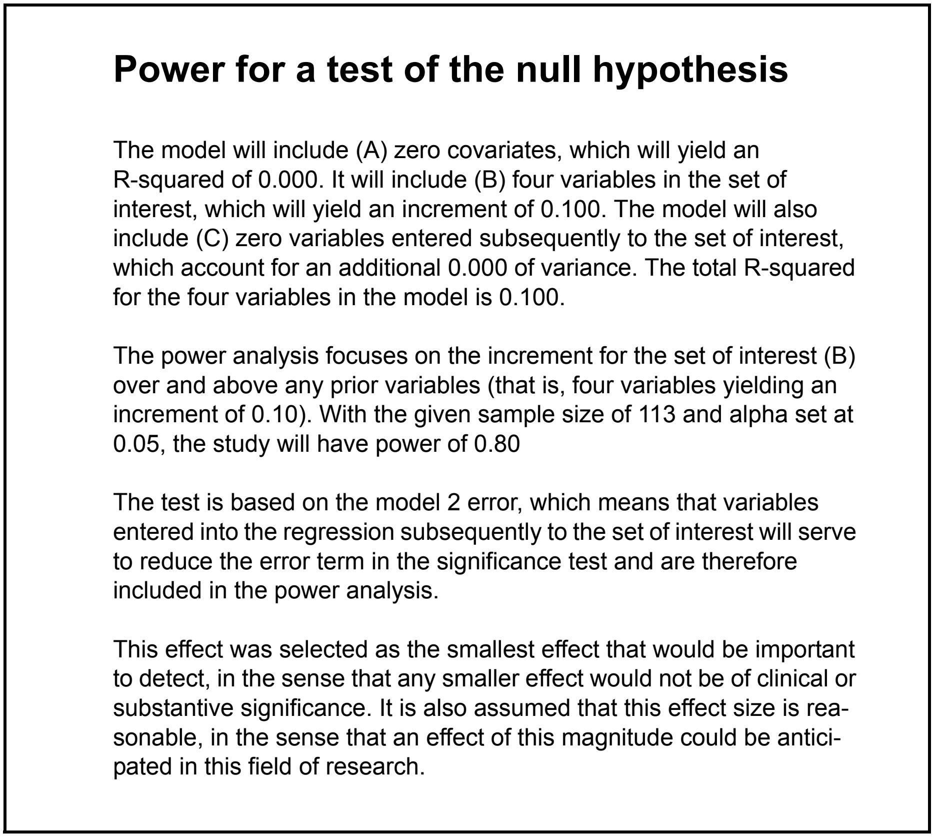

Multiple regression

	Variable	Increment to R-Squared			Cumulative R-Squared		
		Number Variables in Set	Increment to R-Squared	Power for Increment	Cumulative Number Variables	Cumulative R-Square	Power for Cumulative R-Squared
1	Main set	4	0.10	0.80	4	0.10	0.80

Alpha .050

N of cases 113

- Click the *Report* icon to generate the following report:



Power for a test of the null hypothesis

The model will include (A) zero covariates, which will yield an R-squared of 0.000. It will include (B) four variables in the set of interest, which will yield an increment of 0.100. The model will also include (C) zero variables entered subsequently to the set of interest, which account for an additional 0.000 of variance. The total R-squared for the four variables in the model is 0.100.

The power analysis focuses on the increment for the set of interest (B) over and above any prior variables (that is, four variables yielding an increment of 0.10). With the given sample size of 113 and alpha set at 0.05, the study will have power of 0.80

The test is based on the model 2 error, which means that variables entered into the regression subsequently to the set of interest will serve to reduce the error term in the significance test and are therefore included in the power analysis.

This effect was selected as the smallest effect that would be important to detect, in the sense that any smaller effect would not be of clinical or substantive significance. It is also assumed that this effect size is reasonable, in the sense that an effect of this magnitude could be anticipated in this field of research.

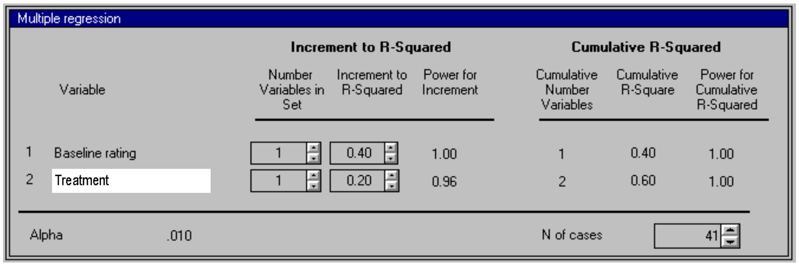
Example 2

Set of Covariates Followed by Set of Interest

Patients are being recruited for a clinical trial that will assess the impact of a drug on asthma symptoms. Patients' symptoms at baseline will serve as a covariate. Patients' drug assignments (dummy coded as 0 for placebo and 1 for drug) will serve as the variable of interest. Symptom level following treatment will serve as the dependent variable.

It is expected that the baseline rating will explain 40% of the variance in outcome. The treatment would be clinically useful if it could explain an additional 20% of the variance. This will serve as the effect size in the power analysis.

The study will be run with alpha set at 0.01, and the researcher wants to compute the sample size required for power of 95%.



Multiple regression

		Increment to R-Squared			Cumulative R-Squared		
	Variable	Number Variables in Set	Increment to R-Squared	Power for Increment	Cumulative Number Variables	Cumulative R-Square	Power for Cumulative R-Squared
1	Baseline rating	1	0.40	1.00	1	0.40	1.00
2	Treatment	1	0.20	0.96	2	0.60	1.00

Alpha .010

N of cases 41

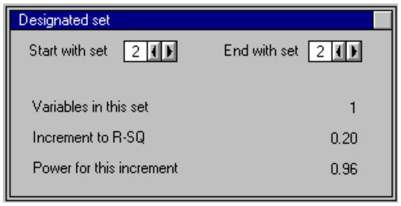
- On the Regression panel, select *Set of covariates followed by set of predictors*. (Or, within the module, choose *Data entry/Study design* from the Options menu, click the *Templates* tab, and select the template *Set of covariates followed by set of interest*.)

- On line 1, enter *Baseline rating*, with one variable and 40% of the variance explained.

- On line 2, enter *Treatment*, with one variable and 20% of the variance explained.

- Click *Alpha* and select the value 0.01.

The next step is to find the number of cases required for power of 99%. However, the program is now showing data for several analyses—power for the covariates, power for the increment due to treatment, and power for the covariates and treatment combined. Therefore, we need to identify the test for which we want 99% power.



Start with set 2   End with set 2

Variables in this set 1

Increment to R-SQ 0.20

Power for this increment 0.96

❖ The program displays a box for this designated set, as shown above. (If the Designated set box is not displayed, choose *Customize screen* from the Options menu and select *Display superset*.) Using the spin controls, define the set as including set 2 only.

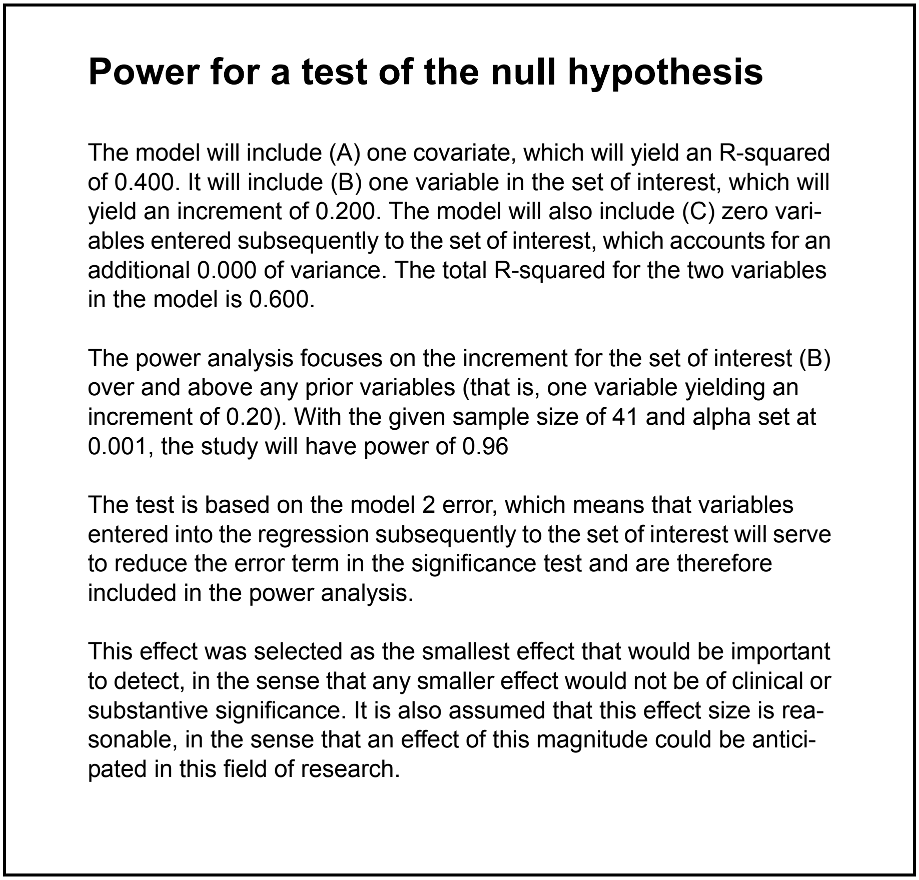
- On the main panel, set 2 (*Treatment*) is now highlighted.
- In line 2, the number of variables is shown as 1.
- In line 2, the increment is shown as 20%.

The program will focus on this designated set when it finds the number of cases required for power.

❖ Press Ctrl-F and click 0.95. The program shows that a sample of 41 cases will yield power of 96% for the designated set (that is, the increment of treatment over the covariates).

Note: The *N of cases* shown is 41 (power = 0.96). If the number of cases is set manually to 40, power will display as 0.95, which suggests that this would have met the criterion. However, if additional digits are displaced (choose *Decimals* from the Options menu), the program displays power for N = 40 as 0.949 rather than 0.95.

❖ Click the *Report* icon to generate the following report:



Power for a test of the null hypothesis

The model will include (A) one covariate, which will yield an R-squared of 0.400. It will include (B) one variable in the set of interest, which will yield an increment of 0.200. The model will also include (C) zero variables entered subsequently to the set of interest, which accounts for an additional 0.000 of variance. The total R-squared for the two variables in the model is 0.600.

The power analysis focuses on the increment for the set of interest (B) over and above any prior variables (that is, one variable yielding an increment of 0.20). With the given sample size of 41 and alpha set at 0.001, the study will have power of 0.96

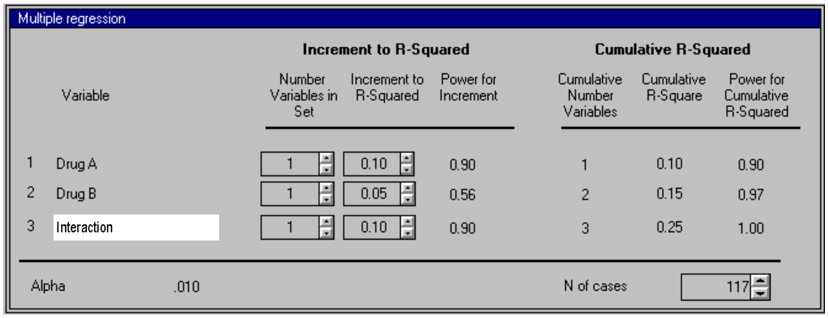
The test is based on the model 2 error, which means that variables entered into the regression subsequently to the set of interest will serve to reduce the error term in the significance test and are therefore included in the power analysis.

This effect was selected as the smallest effect that would be important to detect, in the sense that any smaller effect would not be of clinical or substantive significance. It is also assumed that this effect size is reasonable, in the sense that an effect of this magnitude could be anticipated in this field of research.

Example 3

Two Sets of Variables and Their Interaction

Assume that we want to study the impact of two drugs on the white blood cell count (WBC). We anticipate that the use of either drug will reduce the WBC by a modest amount, but the use of both drugs will have a synergistic effect, and we want to test the impact of this interaction. The smallest effect that would be important to detect would be an increment of 10%. The study will be run with alpha of 0.01, and we want to find the number of cases required for power of 90%



Multiple regression

		Increment to R-Squared			Cumulative R-Squared		
	Variable	Number Variables in Set	Increment to R-Squared	Power for Increment	Cumulative Number Variables	Cumulative R-Square	Power for Cumulative R-Squared
1	Drug A	1	0.10	0.90	1	0.10	0.90
2	Drug B	1	0.05	0.56	2	0.15	0.97
3	Interaction	1	0.10	0.90	3	0.25	1.00

Alpha .010

N of cases 117

- ❖ On the Regression panel, select *Set A, Set B, and interaction*. (Or, within the module, choose *Data entry/Study design* from the Options menu, click the *Templates* tab, and select that template.

- ❖ For line 1, enter *Drug A*, with one variable and 10% of the variance explained.

- ❖ For line 2, enter *Drug B*, with one variable and 5% of the variance explained.

- ❖ For line 3 (*Interaction*), enter one variable and 10% of the variance explained.

- ❖ Click *Alpha* and select the value 0.01.

The program will display power for each test on the screen, but we need to identify one test that will be addressed by the report, tables, graphs, and the Find N procedure.

- ❖ If the Designated set box is not displayed, choose *Customize screen* from the Options menu, and select *Display superset*. Using the spin controls, define the set as including set 3 only. The box shows the designated set as beginning and ending with Set 3 (the interaction), which incorporates one variable and an increment of 10%.

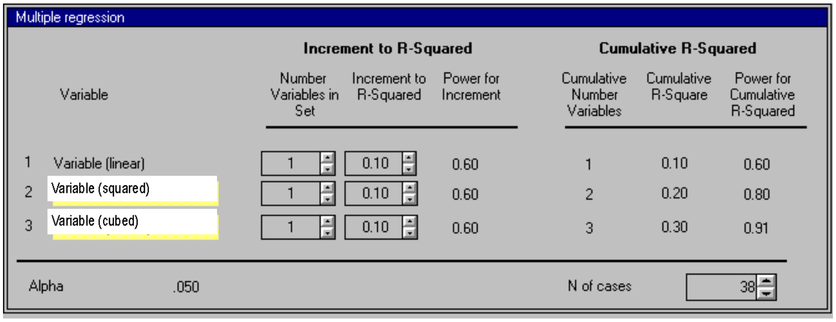
- ❖ Press Ctrl-F (or click the *Find N* icon if the default power has already been set to 90). The program shows that a sample of 117 cases will yield power of 90% for the test just cited.

Optionally, click the *Report*, *Table*, or *Graph* icon to generate additional data, Any of these features will work with the set designated in the box.

Example 4

Polynomial Regression

Patients being treated with neuroleptics will be followed for three months and then assessed for signs of emerging abnormal movements (rated on a five-point scale). The study will include patients in the age range of 20–80. It is anticipated that the risk of these movements increases with age, but the risk is not linear. We anticipate that the risk rises for patients over the age of 50 and rises again, sharply, for patients in their 70's. We plan to enter as variables *age*, *age-squared*, and *age-cubed*.



Multiple regression

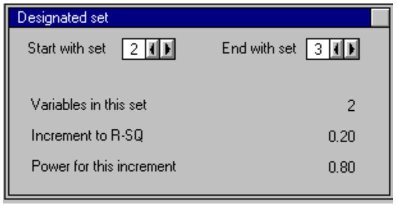
		Increment to R-Squared			Cumulative R-Squared		
	Variable	Number Variables in Set	Increment to R-Squared	Power for Increment	Cumulative Number Variables	Cumulative R-Square	Power for Cumulative R-Squared
1	Variable (linear)	1	0.10	0.60	1	0.10	0.60
2	Variable (squared)	1	0.10	0.60	2	0.20	0.80
3	Variable (cubed)	1	0.10	0.60	3	0.30	0.91

Alpha .050

N of cases 38

- ❖ On the Regression panel, select *Polynomial regression*. (Or, within the module, choose *Data entry/Study design* from the Options menu, click the *Templates* tab, and select the template *Linear, squared, and cubed components*.
- ❖ For line 1, enter *Age*, with one variable and 10% of the variance explained.
- ❖ For line 2, enter *SQ*, with one variable and 10% of the variance explained.
- ❖ For line 3, enter *CU*, with two variables and 10% of the variance explained.
- ❖ Click *Alpha* and select the value 0.05.

To test the hypothesis that the relationship between age and outcome is nonlinear, the researcher wants to treat *age-squared* and *age-cubed* as a set and test the increment of this set over (linear) *age*.


Designated set

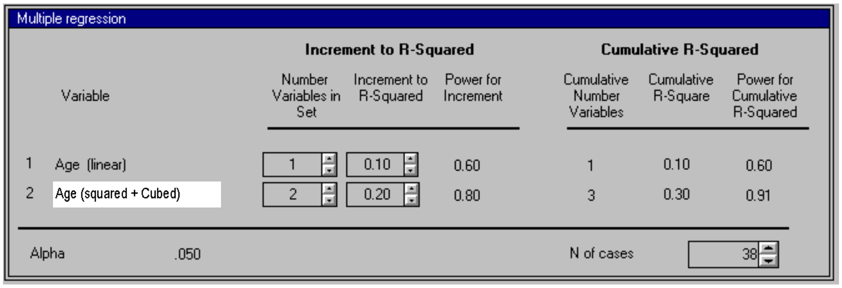
Start with set 2 End with set 3

Variables in this set	2
Increment to R-SQ	0.20
Power for this increment	0.80

- ❖ If the Designated set box is not displayed, choose *Customize screen* from the Options menu and select *Display superset*. Using the spin controls, define the set as including lines 2 and 3. In the main panel, both lines are highlighted, and the set is defined as having two variables and an increment of 20%.

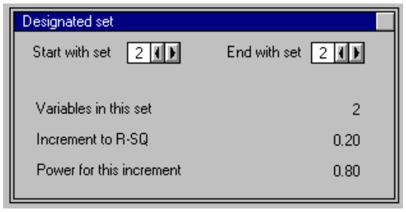
- ❖ Press Ctrl-F (or click the *Find N* icon if the default power has already been set to 80). The program shows that a sample of 38 cases will yield power of 80% for the two-variable set.

Alternatively, you could set the main panel to have only two lines, with *age-squared* and *age-cubed* combined on line 2. The number of variables for line 2 would be 2, with 20% of the variance explained.


Multiple regression

		Increment to R-Squared			Cumulative R-Squared		
	Variable	Number Variables in Set	Increment to R-Squared	Power for Increment	Cumulative Number Variables	Cumulative R-Square	Power for Cumulative R-Squared
1	Age (linear)	1	0.10	0.60	1	0.10	0.60
2	Age (squared + Cubed)	2	0.20	0.80	3	0.30	0.91

Alpha .050 N of cases 38


Designated set

Start with set 2 End with set 2

Variables in this set	2
Increment to R-SQ	0.20
Power for this increment	0.80

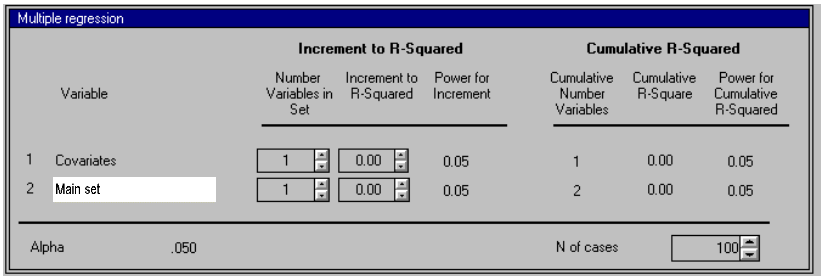
As shown here, either approach will yield the same result. Putting the three variables on three lines, however, allows the researcher to quickly look at power for any combination of sets (the increment for *age-squared*, the increment for *age-cubed*, the increment for *age-squared* combined with *age-cubed*, and so on) without the need to reenter data.

Example 5

One Set of Covariates, Two Sets of Variables, and Interactions

A hospital is planning a study to assess treatment strategies for schizophrenia. The dependent variable is the caregiver's evaluation of the patient's overall functioning at the end of a year. A similar index at study entry will serve as a covariate. The researcher wants to test the impact of the drug assignment, the incremental impact of the support, and the hypothesis that the entire treatment program is having an effect.

❖ On the Regression panel, select *Set of covariates followed by set of predictors*. (Or, within the module, choose *Data entry/Study design* from the Options menu, click the *Templates* tab, and select the template *Set of covariates followed by set of interest*.)

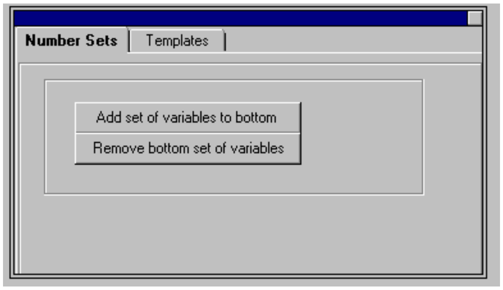


Multiple regression

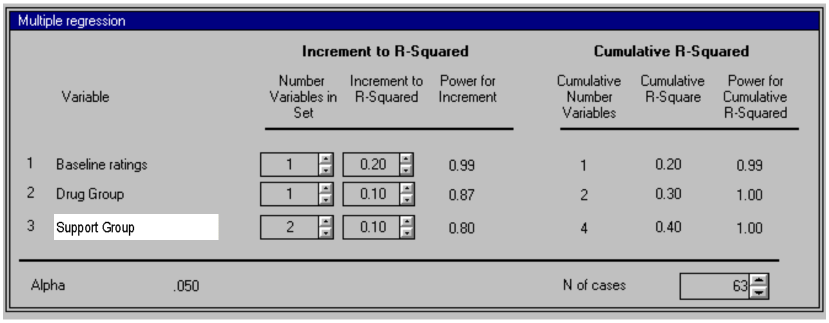
		Increment to R-Squared			Cumulative R-Squared		
	Variable	Number Variables in Set	Increment to R-Squared	Power for Increment	Cumulative Number Variables	Cumulative R-Square	Power for Cumulative R-Squared
1	Covariates	1	0.00	0.05	1	0.00	0.05
2	Main set	1	0.00	0.05	2	0.00	0.05

Alpha .050 N of cases 100

❖ Add an additional row (choose *Data Entry/Study design* from the Options menu).


Number Sets
Templates
Add set of variables to bottom
Remove bottom set of variables

- For line 1, enter *Baseline ratings*, with one variable and 20% of the variance explained.
- For line 2, enter *Drug Group*, with one variable and 10% of the variance explained.
- For line 3, enter *Support Group*, with two variables and 10% of the variance explained.



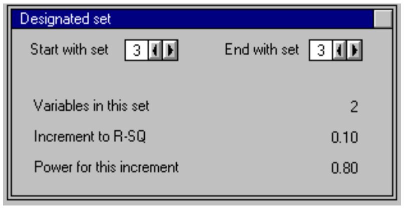
		Increment to R-Squared			Cumulative R-Squared		
	Variable	Number Variables in Set	Increment to R-Squared	Power for Increment	Cumulative Number Variables	Cumulative R-Square	Power for Cumulative R-Squared
1	Baseline ratings	1	0.20	0.99	1	0.20	0.99
2	Drug Group	1	0.10	0.87	2	0.30	1.00
3	Support Group	2	0.10	0.80	4	0.40	1.00

Alpha .050

N of cases 63

The program shows immediately that power is lower for the support group than for the drug group (for each, we are using 10% as the proportion of variance explained, but the support group requires two variables in order to yield this level of prediction). To ensure that power will be adequate for both factors, we must ensure adequate power for the support group, knowing that power for the drug effect will be higher.

- In the main panel, identify *Support Group* as the designated set. If the Designated set box is not displayed, choose *Customize screen* from the Options menu and select *Display superset*. Using the spin controls, define the set as beginning and ending with set 3.



Start with set 3 End with set 3

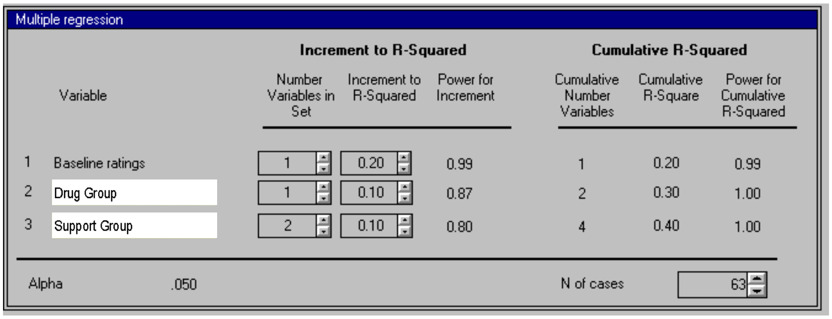
Variables in this set	2
Increment to R-SQ	0.10
Power for this increment	0.80

- Press Ctrl-F (or click the *Find N* icon if the default power has already been set to 80). The program shows that 63 cases will yield power of 80% for the test of the support group (the power is displayed in the Designated set box and also on the main panel, in line 3).

  The main panel also shows that power for the drug group effect is 0.87, and that power for the cumulative set of baseline, drug group and support group approaches 1.00.

Note that power for the entire treatment plan (drug and support, as an increment over baseline) is not displayed on the main panel. However, it can be displayed in the Designated set box.

❖ In the Designated set box, use the spin controls to define the set as including sets 2 and 3. These sets are highlighted on the main panel. The Designated set box shows that power for these combined sets is 0.97.

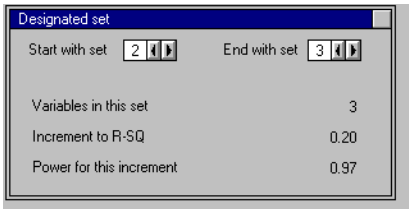


Multiple regression

		Increment to R-Squared			Cumulative R-Squared		
	Variable	Number Variables in Set	Increment to R-Squared	Power for Increment	Cumulative Number Variables	Cumulative R-Square	Power for Cumulative R-Squared
1	Baseline ratings	1	0.20	0.99	1	0.20	0.99
2	Drug Group	1	0.10	0.87	2	0.30	1.00
3	Support Group	2	0.10	0.80	4	0.40	1.00

Alpha .050

N of cases 63



Designated set

Start with set 2

End with set 3

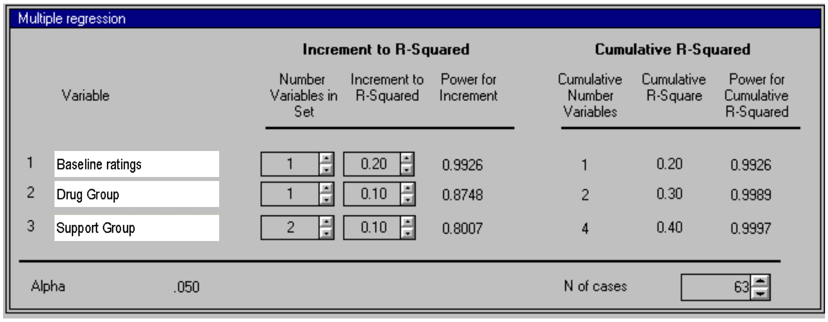
Variables in this set 3

Increment to R-SQ 0.20

Power for this increment 0.97

Power for the Combination of All Three Sets

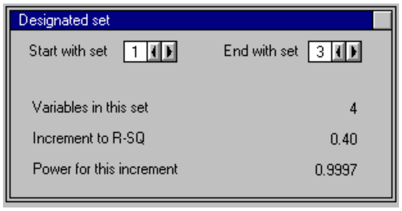
You can also define the designated set as including all three sets as shown below. Power for the designated set is shown as 0.9997 (with four decimals displayed for the purpose of this example), which is the same as the value shown in the main panel for the cumulative power of the three sets.



	Variable	Increment to R-Squared Number Variables in Set	Increment to R-Squared Increment to R-Squared	Increment to R-Squared Power for Increment	Cumulative R-Squared Cumulative Number Variables	Cumulative R-Squared Cumulative R-Square	Cumulative R-Squared Power for Cumulative R-Squared
1	Baseline ratings	1	0.20	0.9926	1	0.20	0.9926
2	Drug Group	1	0.10	0.8748	2	0.30	0.9989
3	Support Group	2	0.10	0.8007	4	0.40	0.9997

Alpha .050

N of cases 63



Start with set 1 End with set 3

Variables in this set	4
Increment to R-SQ	0.40
Power for this increment	0.9997

19 Logistic Regression (Continuous Predictors)

Application

This procedure is used when the dependent variable is **dichotomous** (for example, the patient responded or failed to respond). The program will work with any of the following types of predictors:

- One continuous predictor (the program will compute power for this predictor alone).
- Two continuous predictors (the program will compute power for the two predictors as a set, or for either predictor with the other partialled.

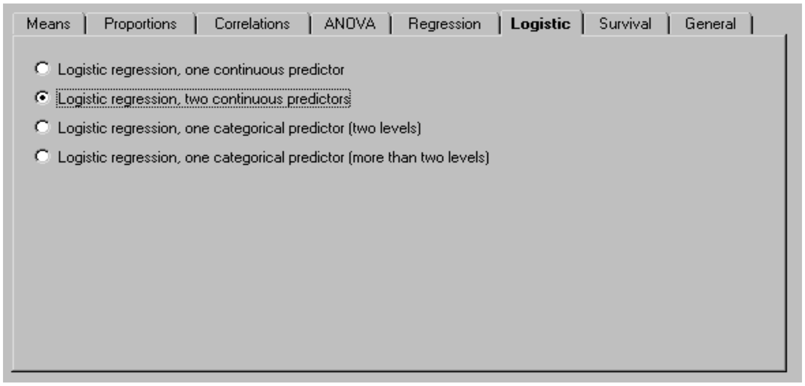
Note: For information on categorical predictors, see Chapter 20.

Selecting the Procedure

❖ To display the available procedures, choose *New analysis* from the File menu.

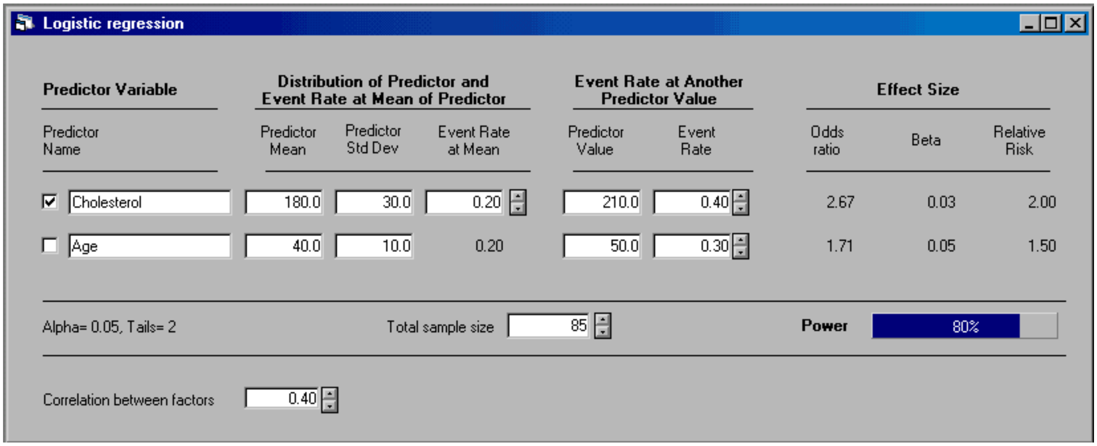
❖ Click the *Logistic* tab.

❖ Select one of the options for continuous predictors, and click *OK* to proceed to the module.



○ Logistic regression, one continuous predictor
● Logistic regression, two continuous predictors
○ Logistic regression, one categorical predictor (two levels)
○ Logistic regression, one categorical predictor (more than two levels)

The program displays the following interactive screen. The screen shown is for two continuous variables, or predictors.



Logistic regression

Predictor Variable		Distribution of Predictor and Event Rate at Mean of Predictor			Event Rate at Another Predictor Value		Effect Size		
	Predictor Name	Predictor Mean	Predictor Std Dev	Event Rate at Mean	Predictor Value	Event Rate	Odds ratio	Beta	Relative Risk
☑	Cholesterol	180.0	30.0	0.20	210.0	0.40	2.67	0.03	2.00
□	Age	40.0	10.0	0.20	50.0	0.30	1.71	0.05	1.50

Alpha= 0.05, Tails= 2

Total sample size 85

Power 80%

Correlation between factors 0.40

Predictor 1

Distribution of predictor and event rate at mean. In this example, we are using the cholesterol level to predict the presence of a specific chemical in patients' blood.

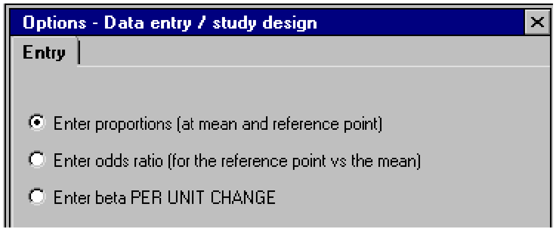
❖ Enter a name for the predictor, the mean and standard deviation expected in the sample, and the event rate at the mean.

In this example, the predictor is cholesterol, the mean cholesterol level is 180, with a standard deviation of 30. The event rate is shown as 0.20. This means that for patients with a cholesterol level of 180, there is a 20% likelihood of this chemical being present.

Event rate at another predictor value. To specify the logistic curve, you need to provide the event rate at some other value of cholesterol. You can provide the event rate at *any* other value. For example, you may enter:

- Cholesterol level at 190, with an event rate of 0.26.
- Cholesterol level of 200, with an event rate of 0.32.
- Cholesterol level of 210 (as shown), which happens to be one standard deviation above the mean, with an event rate of 0.40.

Odds ratio, beta, and relative risk. The effect size in logistic regression is affected by the slope of the curve. The program allows you to specify this curve in any of three ways. To select among these options click *Effect Size* or select *Data entry / study design* from the Options menu.



Options - Data entry / study design

Entry

● Enter proportions (at mean and reference point)

○ Enter odds ratio (for the reference point vs the mean)

○ Enter beta PER UNIT CHANGE

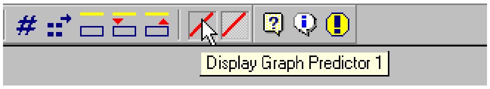
- Provide the event rate at the mean, specify any other value, and provide the event rate at that value.

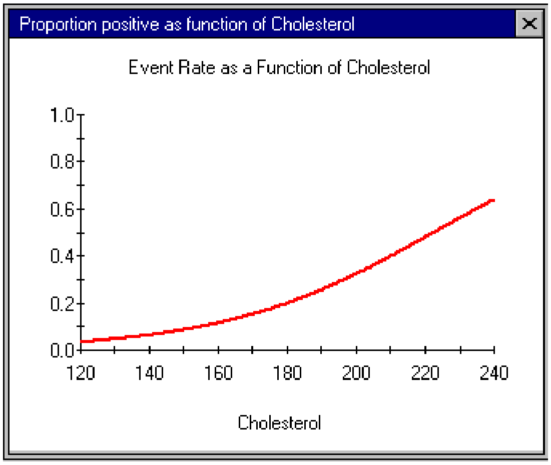
- Provide the event rate at the mean, specify any other value, and provide the odds ratio for the events rate corresponding to this value (versus the mean).

- Provide the event rate at the mean, specify any other value, and specify beta per unit change.

  In this case, the second value provided will not affect the effect size but will allow the program to display the corresponding event rate.

Graph of event rate. Click the *Display Graph Predictor 1* icon to display a graph showing the relationship between the predictor and the event rate.





Proportion positive as function of Cholesterol

The line on this graph should serve as a reality check to ensure that the effect size is both appropriate and realistic. Notice that the graph extends for two standard deviations on either side of the predictor mean. Since the vast majority of the subjects will have cholesterol values in this range, it is the slope of the curve in this range that is critical in determining power. If the slope in this range is shallow, power will be adversely affected.

Predictor 2

Distribution of predictor and event rate at mean. In this example, we want to also take account of *age*, which is the second predictor.

❖ Enter a name for the predictor and the mean and standard deviation expected in the sample.

In this example, the mean value for age is 40 and the standard deviation is 10. The event rate is shown as 0.20, which is the value entered for cholesterol. (The event rate at the mean of the sample will be the same for any predictor variable.)

Event rate at another predictor value. To specify the logistic curve, we need to provide the event rate at some other value of age. The procedure here is identical to that for the first predictor. Again, the program displays the odds ratio, beta, and relative risk, as well as a graph.

Correlation between the two predictors. When you are working with two predictors, you need to provide the correlation between them. Note the following:

- An increase in the correlation will sometimes increase the power and will sometimes decrease the power. This depends on the hypothesis being tested (whether the second predictor is included in the set or partialled).
- The direction of the correlation is important. While this may seem counter-intuitive, consider the current example. We have specified that an increased cholesterol level or a higher age will be associated with a higher event rate. If we want to use the two together to predict the event rate and they have a positive correlation with each other, the predictive power of the set will tend to be high. In contrast, if they tend to run in opposite directions, the predictive power of the set will be adversely affected. In other words, if we switch the sign of the correlation, we also need to switch the sign of beta, or power will be affected.
- If you are not able to specify the correlation, you may want to work with a range of values and see how much power is affected. For example, if the correlation is bound to fall in the range of 0.20 to 0.50, you may find that power tends to remain relatively stable for correlations in this range (or it may not). You can use the tables and graphs and include correlation as a factor, to quickly assess its importance.

Hypothesis to Be Tested

When there is a single predictor variable, the program will compute power for this predictor. When there are two predictor variables, the program will compute power for:

- The two variables as a set.
- The first variable when the second variable is partialled.
- The second variable when the first variable is partialled.

To select among these options, use the check boxes that appear to the left of the variables. When both are selected, they are both included in the set. When only one is selected, the other will be partialled.

Alpha, Confidence Level, and Tails

The program shows Alpha and Tails. To modify, click *Alpha*. When the test is one-tailed, the program is testing for an increase in the event rate.

Sample Size

Enter the sample size or use the Find N icon to locate the sample size required for a given level of power.

Power

The program shows the level of power. For the combination of accrual, sample size, hazard rates, attrition, and alpha and tails, the power level is 87%. If you modify any values on the screen, power changes accordingly.

Example

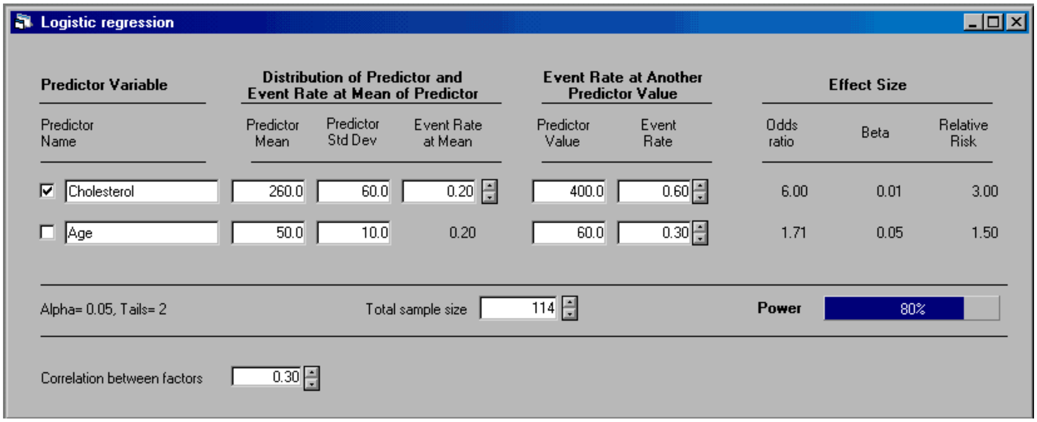
Synopsis

We want to look at the relationship between cholesterol levels and the presence of a marker for cardiac events. The likelihood of this marker is known to increase with age, so *age* will be taken into account as well as *cholesterol*.

❖ Choose *New analysis* from the File menu.

- Click the *Logistic* tab.
- Select *Logistic regression, two continuous predictors*, and click *OK*.

The program displays the following interactive screen:



Logistic regression

Predictor Variable	Distribution of Predictor and Event Rate at Mean of Predictor			Event Rate at Another Predictor Value		Effect Size		
Predictor Name	Predictor Mean	Predictor Std Dev	Event Rate at Mean	Predictor Value	Event Rate	Odds ratio	Beta	Relative Risk
☑ Cholesterol	260.0	60.0	0.20	400.0	0.60	6.00	0.01	3.00
□ Age	50.0	10.0	0.20	60.0	0.30	1.71	0.05	1.50

Alpha= 0.05, Tails= 2

Total sample size 114

Power 80%

Correlation between factors 0.30

- We expect that the mean cholesterol value will be on the order of 260, with a standard deviation of 60. For patients falling at the mean cholesterol level (260) or the mean age (50), the likelihood of the marker being present is about 20%. These values are entered under Distribution of Predictor and Event Rate at Mean of Predictor.
- The likelihood of the marker being present is expected to fall on the order of 60% for patients with cholesterol levels of 400. For Event Rate at Another Predictor Value, the researcher enters 400.0 as the predictor value and 0.60 as the event rate.
- To ensure that this effect size has been described accurately, the program displays it in several formats. The odds ratio for the marker being present versus absent at cholesterol levels of 260 versus 400 is shown as 6.00. This value is high, but considering the fact that it represents the difference between patients at the mean and those with a cholesterol value of 400, it seems appropriate. This is confirmed in the graph of cholesterol values by the presence of the marker, which is consistent with the expectations based on prior research. The effect is also shown to represent a beta of 0.01 and a relative risk of 3.0. Critically, this represents the smallest effect that would be important to detect. That is, we want to have adequate power to obtain a significant result if the effect is this large or larger.
- For *age*, we enter a mean value of 50 with a standard deviation of 10. The event rate at the mean must be the same as the event rate at the cholesterol mean, which has already been entered as 0.20. We provide a second value for age (60) and the event rate at this

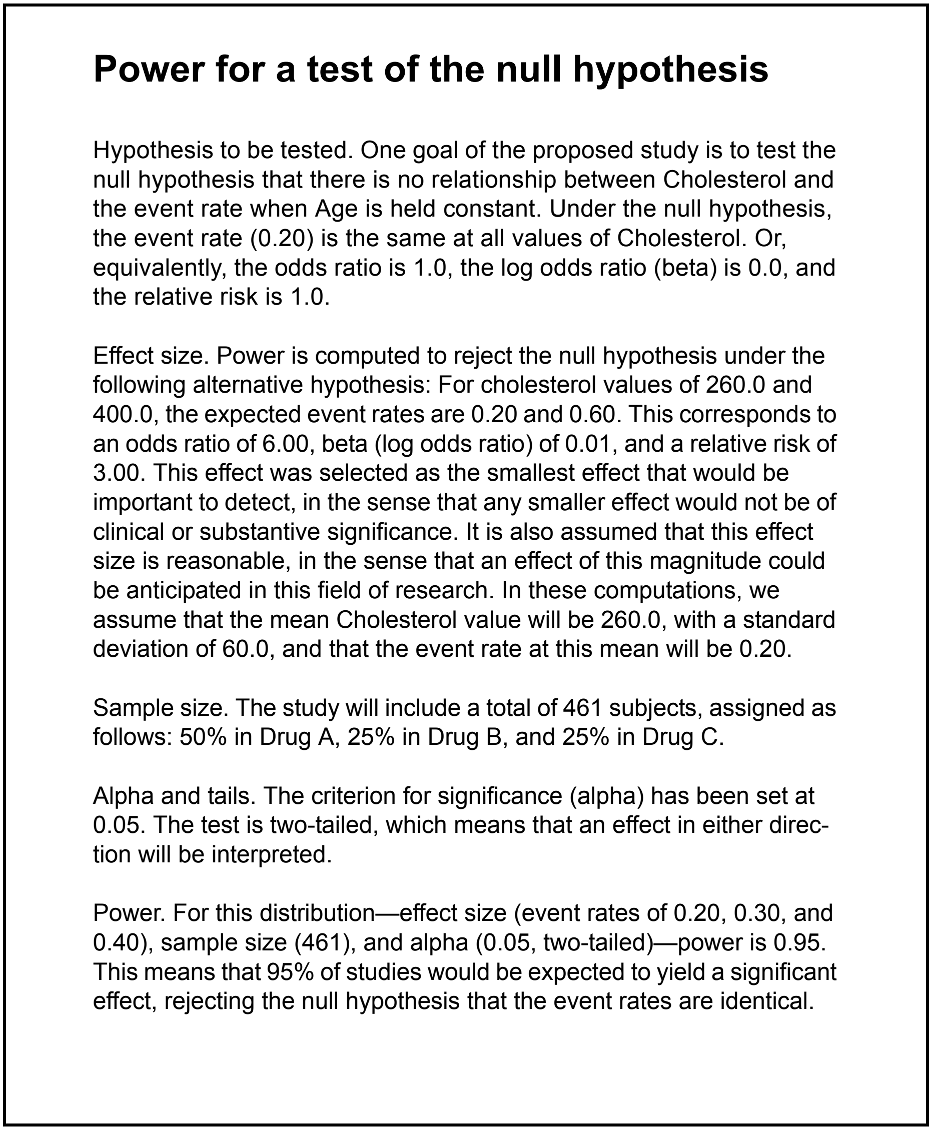
age (0.30). This corresponds to an odds ratio of 1.71, a beta of 0.05, and a relative risk of 1.5. Again, the graph of event rate by age is consistent with our clinical experience.

- The correlation between cholesterol and age is expected to fall on the order of 0.30, so this value is entered for the correlation.
- Alpha is shown as 0.05, two-tailed. To modify alpha, click *Alpha* and select an alternative value.
- Our primary concern is whether we will be able to establish the relationship between cholesterol level and the marker when age is partialled. Therefore, we place a check mark adjacent to *cholesterol* only.
- Click the *Find N* icon and select a power level of 80%.

  The program shows that we would need 114 patients to have a power of 80% to obtain a significant effect for cholesterol with age partialled.

To create a report:

- Click the *Report* icon to generate the following report:



Power for a test of the null hypothesis

Hypothesis to be tested. One goal of the proposed study is to test the null hypothesis that there is no relationship between Cholesterol and the event rate when Age is held constant. Under the null hypothesis, the event rate (0.20) is the same at all values of Cholesterol. Or, equivalently, the odds ratio is 1.0, the log odds ratio (beta) is 0.0, and the relative risk is 1.0.

Effect size. Power is computed to reject the null hypothesis under the following alternative hypothesis: For cholesterol values of 260.0 and 400.0, the expected event rates are 0.20 and 0.60. This corresponds to an odds ratio of 6.00, beta (log odds ratio) of 0.01, and a relative risk of 3.00. This effect was selected as the smallest effect that would be important to detect, in the sense that any smaller effect would not be of clinical or substantive significance. It is also assumed that this effect size is reasonable, in the sense that an effect of this magnitude could be anticipated in this field of research. In these computations, we assume that the mean Cholesterol value will be 260.0, with a standard deviation of 60.0, and that the event rate at this mean will be 0.20.

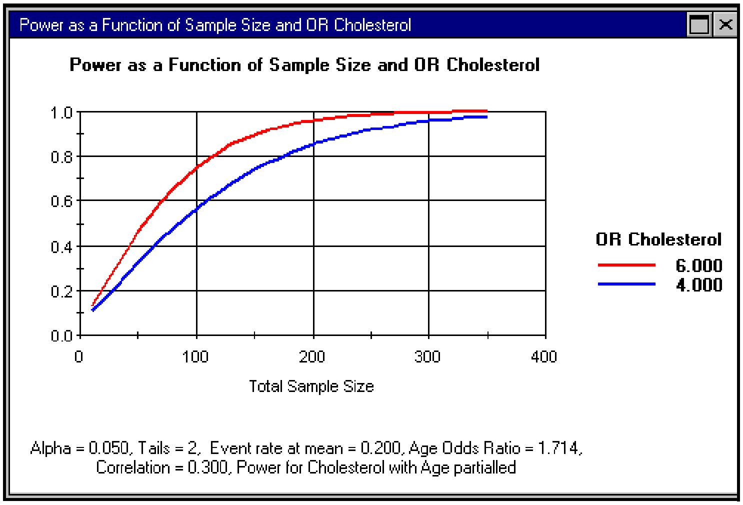
Sample size. The study will include a total of 461 subjects, assigned as follows: 50% in Drug A, 25% in Drug B, and 25% in Drug C.

Alpha and tails. The criterion for significance (alpha) has been set at 0.05. The test is two-tailed, which means that an effect in either direction will be interpreted.

Power. For this distribution—effect size (event rates of 0.20, 0.30, and 0.40), sample size (461), and alpha (0.05, two-tailed)—power is 0.95. This means that 95% of studies would be expected to yield a significant effect, rejecting the null hypothesis that the event rates are identical.

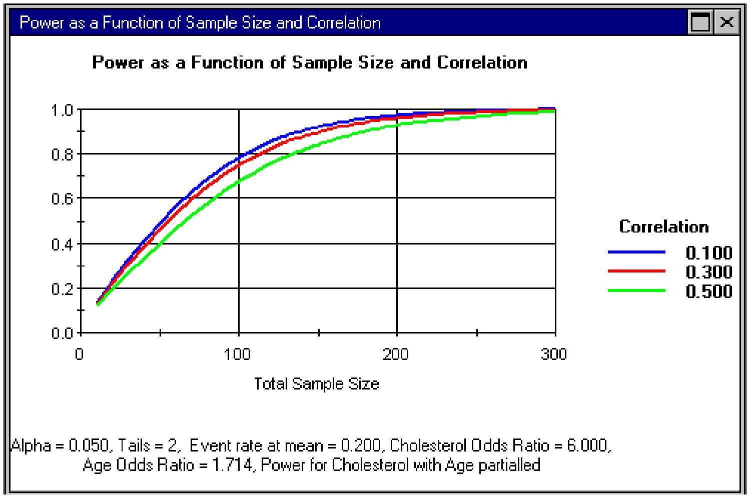
To create table and graphs:

At this point, we have established that power will be 80% for the given set of assumptions. However, we are concerned that some of our assumptions may be in error. For example, what if the odds ratio is not 6.0 but only 4.0? Click the *Tables and graphs* icon to create a table and graph for two values of the odds ratio—6.0 (as before) and 4.0 (as the lower limit). The graph shows that if the odds ratio is actually 4.0 rather than 6.0, power would drop by some 20 points. In other words, to ensure power of 80%, we need a sample of around 180, rather than 120.


Power as a Function of Sample Size and OR Cholesterol
Power as a Function of Sample Size and OR Cholesterol
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
Total Sample Size
OR Cholesterol
6.000
4.000

Alpha = 0.050, Tails = 2, Event rate at mean = 0.200, Age Odds Ratio = 1.714,
Correlation = 0.300, Power for Cholesterol with Age partialled

At this point, the facts are relatively clear. We may decide that the extra resources required to detect a smaller effect cannot be justified (that is, we need to commit more patients but are trying to detect an effect that is less clinically important) and work with the sample of 120. Or we may decide than an odds ratio of 4.0 is clinically important and then work with a sample size of 180.



Alpha = 0.050, Tails = 2, Event rate at mean = 0.200, Cholesterol Odds Ratio = 6.000,
Age Odds Ratio = 1.714, Power for Cholesterol with Age partialled

20 Logistic Regression (Categorical Predictors)

Application

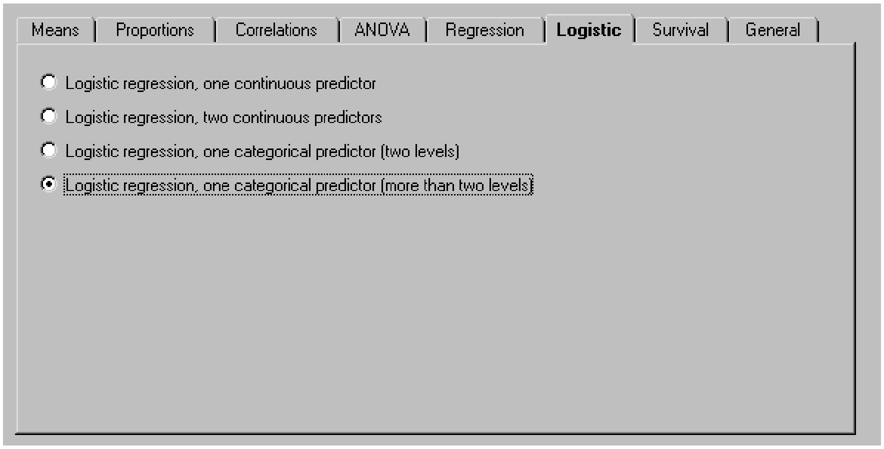
This procedure is used when the dependent variable is **dichotomous** (for example, whether the patient responded or failed to respond) and the predictor is **categorical** (if the predictor is continuous, see Chapter 19). The program allows the user to specify one categorical predictor with two to five groups and will compute power for an omnibus test of all groups, or it will allow the user to select the specific groups to be compared.

Selecting the Procedure

- To display the available procedures, choose *New analysis* from the File menu.

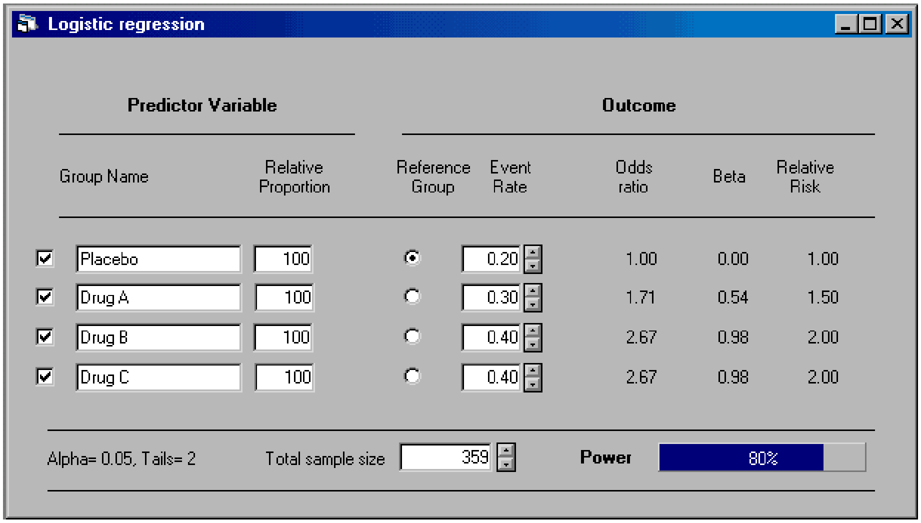
- Click the *Logistic* tab.

- Select one of the options for categorical predictors, and click *OK* to proceed to the module.



○ Logistic regression, one continuous predictor
○ Logistic regression, two continuous predictors
○ Logistic regression, one categorical predictor (two levels)
● Logistic regression, one categorical predictor (more than two levels)

The program displays the following interactive screen. The screen shown is for one continuous variable with four categories.



Logistic regression

	Predictor Variable		Outcome				
	Group Name	Relative Proportion	Reference Group	Event Rate	Odds ratio	Beta	Relative Risk
☑	Placebo	100	●	0.20	1.00	0.00	1.00
☑	Drug A	100	○	0.30	1.71	0.54	1.50
☑	Drug B	100	○	0.40	2.67	0.98	2.00
☑	Drug C	100	○	0.40	2.67	0.98	2.00

Alpha= 0.05, Tails= 2   Total sample size 359   Power 80%

In this example, patients are being assigned to one of four treatment groups—*Placebo*, *Drug A*, *Drug B*, or *Drug C*—and we will compare the proportion responding in these groups.

Labels

For each group, enter a name, which will be used in the reports.

Relative Proportion

For each group, enter the relative proportion assigned to that group. For example, if subjects will be assigned to all groups in equal numbers, type 1.0 for each group. If twice as many cases will be assigned to the first group than to any other group, type 2.0 for the first group and 1.0 for each of the other groups.

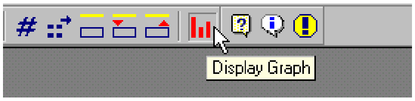
Event Rate

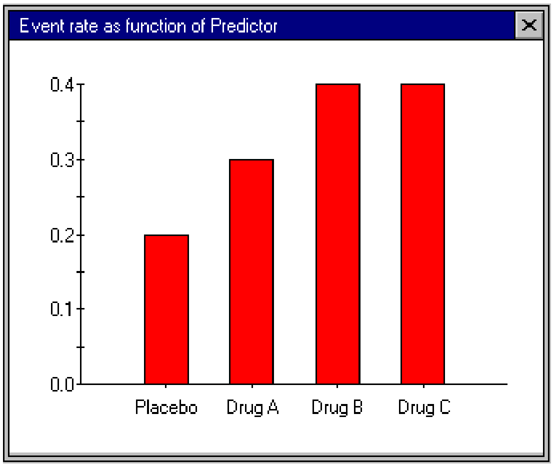
Enter the expected event rate for each group. For example, this might be the proportion of subjects expected to respond in each group. As is true for all procedures, power will be computed for the effect size (the difference between groups) entered here and will be lower for any smaller effect size. Therefore, the values used here should be considered carefully.

Odds Ratio, Beta, and Relative Risk

Select any one of the groups to serve as the reference group. For the reference group, the odds ratio and relative risk are shown as 1.0 and beta is shown as 0.0. For each of the other groups, these values are shown in comparison to the reference group. This feature is helpful for quantifying the magnitude of the effect size, but the selection of a group as the reference group has no impact on power.

Click the *Display Graph* icon to display a graph showing the relationship between the predictor and the event rate. This graph should serve as a reality check to ensure that the effect size is both appropriate and realistic.





Hypothesis to Be Tested

The program displays a check box to the left of each group. If the check box is selected, the group will be included in the test. In this example, if all check boxes are selected, power is computed for an omnibus test of all four groups. If only the first two check boxes are selected, power is computed for a test of *Drug A* versus *Placebo*.

Important. All groups, including those not included in the test, still have an impact on power because the subjects are being assigned to all groups in whatever ratio was specified. For example, if there are four groups with an equal allocation of patients, a sample of 400 implies that 100 are assigned to each, for a total of 100×4 , or 400; if only two check boxes are selected, the effective sample size becomes 100×2 , or 200.

Alpha, Confidence Level, and Tails

The program shows values for alpha and tails. To modify the values, click *Alpha*. When the test is one-tailed, the program is testing for an increase in the event rate.

Sample Size

Enter the sample size, or use the Find N icon to locate the sample size required for a given level of power.

Power

The program shows the level of power. For this distribution, effect size (event rates of 0.20, 0.30, 0.40, and 0.40), sample size (359), and alpha (0.05, two-tailed), power is 0.80. This means that 80% of studies are expected to yield a significant effect, rejecting the null hypothesis that the event rates are identical.

Example

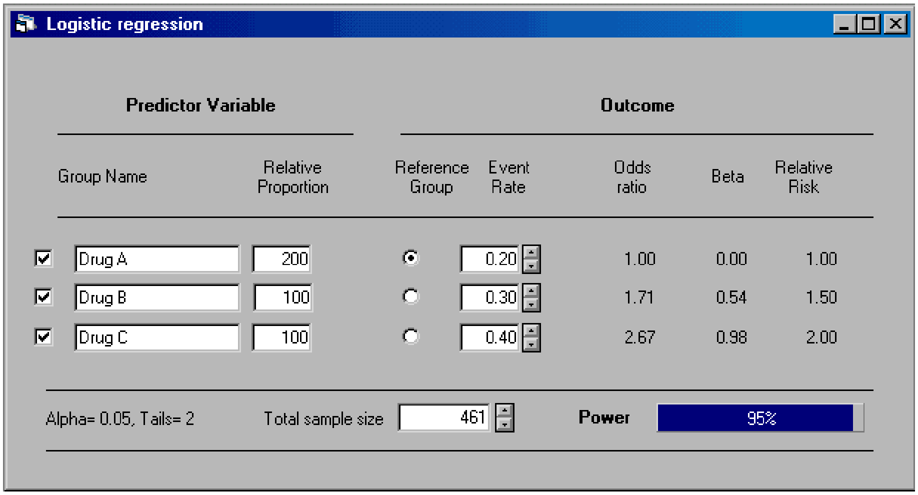
Synopsis

Patients are assigned to one of three drugs (*Drug A*, *Drug B*, or *Drug C*) and then classified as either responding or not responding to the treatment. We want to test for any differences in response rates across the three drugs.

Interactive Computation

- Choose *New analysis* from the File menu.
- Click the *Logistic* tab.
- Select *Logistic regression, one categorical predictor (more than two levels)*, and click *OK*.

  The program displays the following interactive screen:



	Predictor Variable		Outcome				
	Group Name	Relative Proportion	Reference Group	Event Rate	Odds ratio	Beta	Relative Risk
☑	Drug A	200	●	0.20	1.00	0.00	1.00
☑	Drug B	100	○	0.30	1.71	0.54	1.50
☑	Drug C	100	○	0.40	2.67	0.98	2.00

Alpha= 0.05, Tails= 2 Total sample size 461 **Power** 95%

If necessary, use the Tools menu to add or remove groups.

- *Drug A* is the current standard treatment and yields a response rate of 0.20. We want to compute power on the assumption that *Drug B* will yield a response rate of 0.30 and that *Drug C* will yield a response rate of 0.40. These values are entered for the three drugs. *Drug A* is set as the reference group (odds ratio of 1.0). The odds ratios for the other groups relative to this group are displayed.
- We plan to enter 50% of the patients into the standard treatment (*Drug A*) and 25% into each of the experimental treatments (*Drug B* and *Drug C*). We enter the relative proportions as 200, 100, and 100 (other numbers, such as 50, 25, and 25, would work as well).
- Alpha is shown as 0.05, two-tailed. To modify alpha, click *Alpha* and select an alternative value.

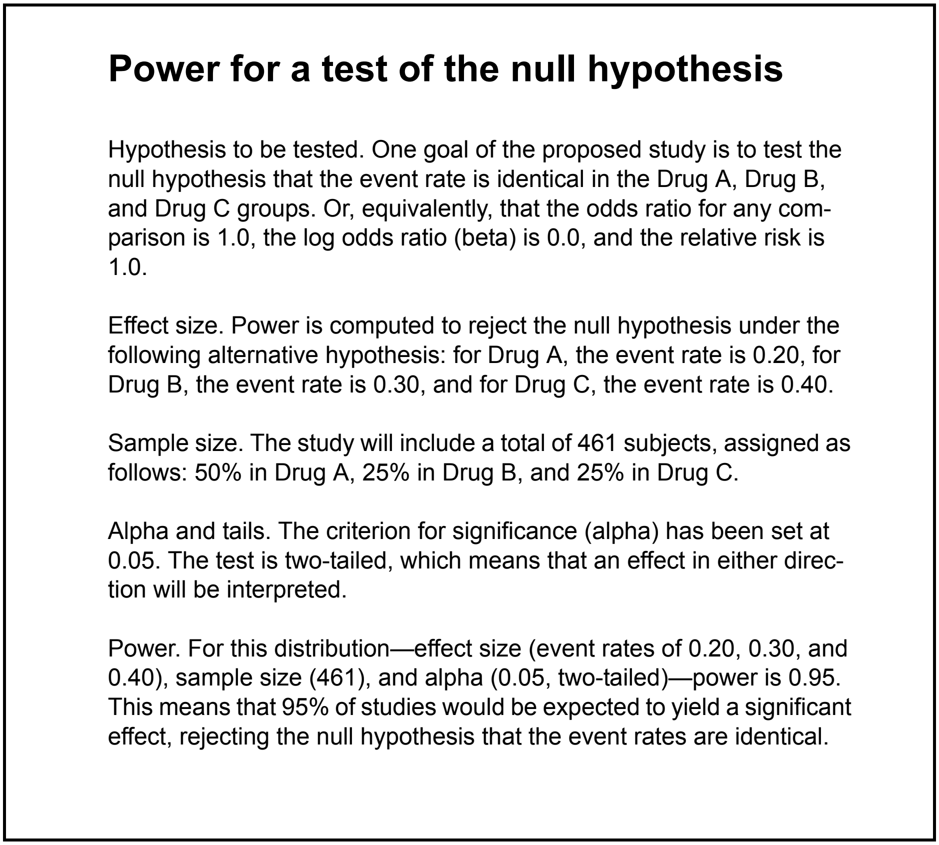
- Our plan is to run an overall test of significance incorporating all three groups. Select the check box adjacent to each group. This indicates that each group will be included in the test.

- Click the *Find N* icon.

  The program shows that we would need a total of 461 patients to have power of 95%. These patients would be assigned to the three groups in the ratio of 2:1:1 (otherwise, the power would be different).

  To create a report:

- Click the *Report* icon to generate the following report:



Power for a test of the null hypothesis

Hypothesis to be tested. One goal of the proposed study is to test the null hypothesis that the event rate is identical in the Drug A, Drug B, and Drug C groups. Or, equivalently, that the odds ratio for any comparison is 1.0, the log odds ratio (beta) is 0.0, and the relative risk is 1.0.

Effect size. Power is computed to reject the null hypothesis under the following alternative hypothesis: for Drug A, the event rate is 0.20, for Drug B, the event rate is 0.30, and for Drug C, the event rate is 0.40.

Sample size. The study will include a total of 461 subjects, assigned as follows: 50% in Drug A, 25% in Drug B, and 25% in Drug C.

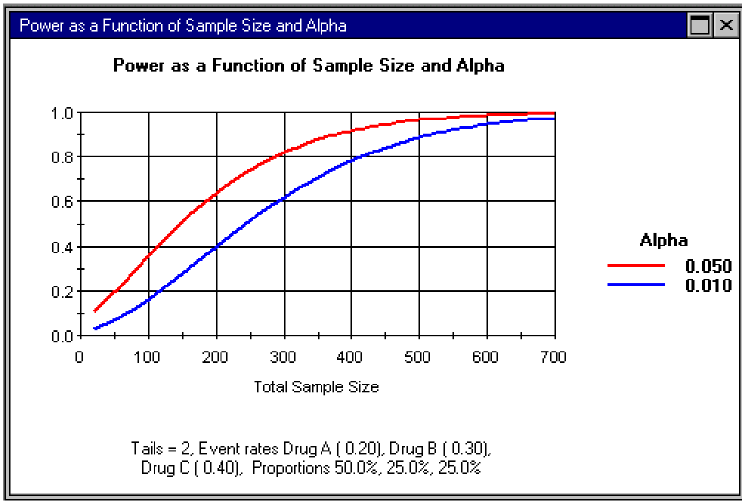
Alpha and tails. The criterion for significance (alpha) has been set at 0.05. The test is two-tailed, which means that an effect in either direction will be interpreted.

Power. For this distribution—effect size (event rates of 0.20, 0.30, and 0.40), sample size (461), and alpha (0.05, two-tailed)—power is 0.95. This means that 95% of studies would be expected to yield a significant effect, rejecting the null hypothesis that the event rates are identical.

To create tables and graphs:

At this point, we have established that power will be 95% for the given set of parameters. However, we want to explore other options. For example, what if we used a somewhat smaller or larger sample or set alpha at 0.01?

Click the *Tables and graphs* icon to create a table and graph for two values of alpha—0.05 and 0.01. The graph below shows that the slope of the power curve is fairly shallow in the range of 95% power. That is, with alpha set at 0.05, as we increased the sample size from 300 to 380 to 450, power went from 80% to approximately 90% to approximately 95%. We may decide to lower the sample size to 380, keeping power at 90, and a substantial savings in the sample size. We also see that in moving from alpha = 0.05 to alpha = 0.01 , we need to increase the sample size by about 100 patients, and this is true for any level of power in the range of concern.



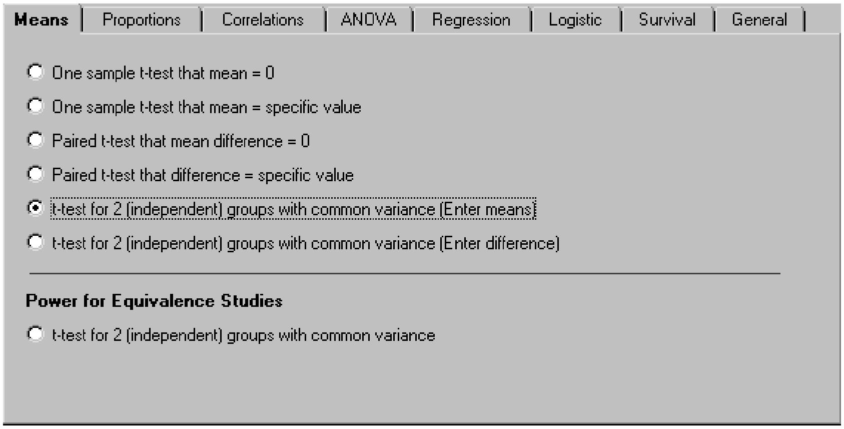
Tails = 2, Event rates Drug A (0.20), Drug B (0.30), Drug C (0.40), Proportions 50.0%, 25.0%, 25.0%

21 Survival Analysis

Survival analysis is used when subjects will be followed over a period of time, and you want to compare the proportion of subjects surviving at each point in time. Survival analysis is used most often in clinical trials, and the term *Survival* comes from these studies where the outcome is, literally, survival. The method, however, can be used to track the time to any event (such as time to success) and is used in many fields. The program will compute power for studies that compare survival rates in two groups.

Selecting the Procedure

To display the available procedures, choose *New analysis* from the File menu.



○ One sample t-test that mean = 0

○ One sample t-test that mean = specific value

○ Paired t-test that mean difference = 0

○ Paired t-test that difference = specific value

● t-test for 2 (independent) groups with common variance (Enter means)

○ t-test for 2 (independent) groups with common variance (Enter difference)

Power for Equivalence Studies

○ t-test for 2 (independent) groups with common variance

Tip

As a rule, select the Advanced option in each category only when the advanced features are needed. The other options allow for quicker data entry and more flexible creation of tables and graphs.

Options for Accrual

Subjects entered prior to the first study interval. All subjects are accrued and waiting before the study begins, so all subjects enter the study on the first day. If the follow-up period is 24 months, this means that each subject will be followed for the full 24 months unless that subject dies, for example, or drops out of the study.

Subjects entered during study, at one rate. Subjects enter the study over a period of time, at a constant rate. For example, you may plan to enter subjects over a period of 6 months and then follow them for an additional 18 months after the last one enters. The first subject could be followed for as long as 24 months while the last one could be followed for no more than 18 months.

Advanced: accrual varies. The accrual rate varies. For example, you may plan to enter a small number of patients each month for three months as recruiting efforts are being refined, and then increase the entry rate after that. Or, you may plan to enter some patients before the study begins and additional patients after the study is underway. If this option is selected, you will not be able to include accrual time as a variable in tables and graphs.

Options for Hazard Rates

Hazard rate is constant. The hazard rate is constant over time. For example, the hazard rate is 5% for one group and 10% for the other group for all time periods in the study.

Advanced: Hazard rate varies. The hazard rate varies from one time period to the next. For example, the hazard rates in the two groups are initially 2% versus 4%, but later rise to 5% versus 10%. The rates in the two groups may be proportional, but this is not required.

Options for Attrition

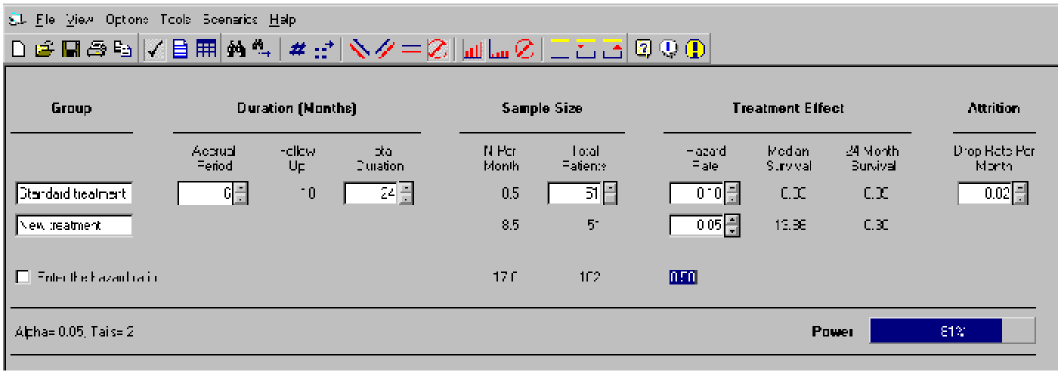
No attrition. No patients will be lost to attrition. This might be the case, for example, if the study will last for only a brief period of time and the intervention takes place at the start, as in the case of surgery versus no surgery, rather than on a continuous basis.

Attrition is constant. Select this option if the attrition rate will be constant throughout the study and will be the same for both groups. This might be appropriate, for example, if the only reason for attrition is likely to be patients moving away from the geographic area. If this option is selected, attrition can be included as a variable in the tables and graphs.

Advanced: Attrition varies. This option allows you to provide an attrition rate separately for each time interval and for each group. For example, you might expect that the attrition rate is high initially as some patients reconsider their commitment to the study, but then levels off. Or, you might expect the attrition to be higher in one treatment group than another. If this option is selected, attrition cannot be included as a variable in the tables and graphs.

Interactive Screen

The screen shown here is one of 18 possible main screens for the survival module. The actual display is determined by the design options selected under procedures. For example, if the hazard rate is constant, the program displays a text box for hazard rate; if the hazard rate varies from one time interval to the next, the program displays a grid instead. A similar approach is applied to the displays for accrual and attrition. The screen is divided into sections labeled *Group, Duration, Sample Size, Treatment Effect,* and *Attrition.* The use of these sections is described here.



Group	Duration (Months)			Sample Size		Treatment Effect			Attrition
	Accrual Period	Follow Up	Study Duration	N Per Month	Total Patients	Hazard Rate	Median Survival	24 Month Survival	Drop Rate Per Month
Standard treatment	0	10	24	0.5	51	0.10	[illegible]	[illegible]	0.02
New treatment				8.5	51	0.05	13.86	0.30	
□ Enter the hazard ratio				17.0	102	0.50			

Alpha= 0.05, Tails= 2

Power 81%

Groups

The first section on this screen is labeled Group. Provide a name for each of the two treatment groups. This name will be used in the reports and graphs.

Tips

By default, the program refers to time intervals as Intervals. You may set this to any other time period such as *Months* or *Years* by choosing *Customize screen* from the Options menu. This has no impact on the calculations but is useful for working with the interactive screen and generating reports.

By default, the program refers to subjects as *Subjects*. You may set this to any other name such as *Patients* by choosing *Customize screen* from the Options menu.

Duration

The total duration of a study includes an accrual period and a follow-up period. For example, a study may run for 24 months, with patients being accrued for the initial 6 months and then followed for an additional 18 months.

You may enter the accrual period and the follow-up period, in which case the program will display the total duration. Or, you may enter the accrual period and the total duration, in which case the program will display the follow-up period. To select either of these options, click *Duration* or use the Options menu.

Note. If you selected *No Accrual* on the Procedures screen, then this screen will display a column for follow-up but not for accrual or for total duration.

How Study Duration and Accrual Affect Power

As a rule, the longer the study duration, the higher the power, because a longer period of observation allows you to observe more events.

For a given number of patients and a given study duration, power is higher to the extent that patients are entered earlier (and followed longer). Specifically,

- Power is higher if patients are entered prior to the study, rather than during the study.
- Power is higher if the accrual period is short, rather than long.
- For a given accrual period, power is highest if the accrual rate is initially high. It is lower if the accrual rate is constant, and it is lowest if the accrual rate increased over time.

These rules follow from the fact that the earlier a patient is entered, the longer that patient can be followed before the study ends. The impact of accrual rates on power will vary from one study to the next. For example, if patients will be followed for 10 years, it doesn't matter if they are accrued over 2 months or over 6 months. By contrast, if patients will be followed for only a year, then this may be an important distinction.

Treatment Effect

The hazard rate is the instantaneous risk of an event (such as dying) at any point in time. When the intervals are reasonably small, this value is close to the risk of dying within the interval.

Treatment Effect When the Hazard Rate Is Constant

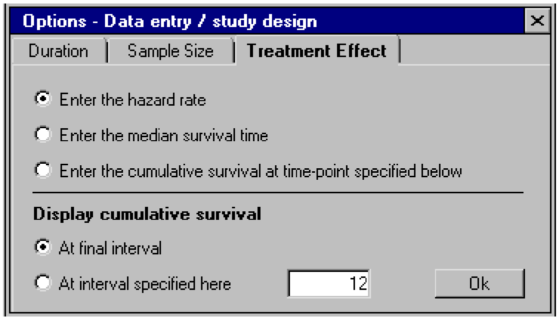
When the hazard rate is constant across all time intervals, any of three items—the hazard rate, the median survival time, or the cumulative survival through any point in time, is determined by either of the others. The program takes advantage of this and allows the user to enter any of the three values.

When the hazard rate is constant, the program displays three sets of text boxes—Hazard Rate, Median Survival, and Cumulative Survival. You may enter any one of these, and the program will display the other two.

❖ To select one of these three options, click *Treatment Effect.*

or

❖ Choose *Data entry/study design* from the Options menu.



● Enter the hazard rate
○ Enter the median survival time
○ Enter the cumulative survival at time-point specified below

Display cumulative survival

● At final interval
○ At interval specified here 12

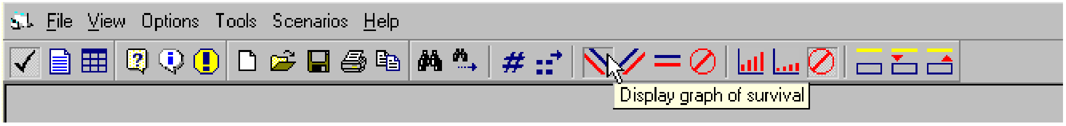
In this example, the user has entered a hazard rate of 0.05. The program displays a median survival time of 6.93 months and a cumulative survival of 0.09 at 24 months.

Tips

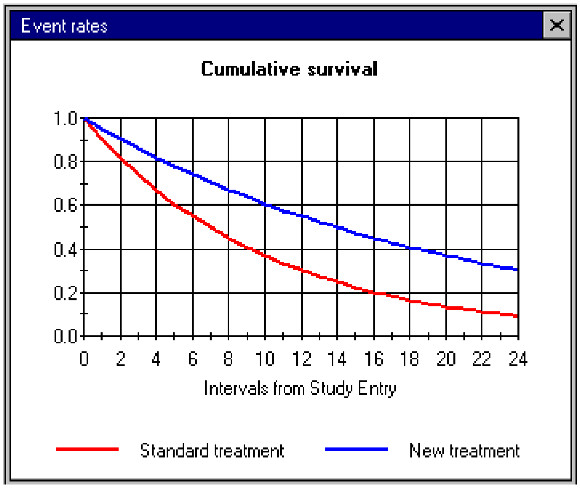
You may elect to enter the hazard rate for the first group and also for the second group, in which case the program will display the hazard ratio (the ratio of the two hazard rates). Alternatively, you may enter the hazard rate for the first group and the hazard ratio, in which case the program will compute the hazard rate for the second group. To switch between these options, select *Enter hazard ratio*.

The program assumes that the cumulative survival should be displayed (or entered) for the final interval. In this example, the study will run for 24 months, so the program displays the cumulative survival at 24 months. However, you may elect to work with the cumulative survival at any point in time. For example, you may be planning a study that will run for 36 months, but want to enter the cumulative survival rate at 60 months since this is the information that you have available. To set this option, click *Treatment Effect*. Then select the option for working with the cumulative survival at an alternate value, and enter that value.

The program will display a graph of cumulative survival. Use the icons on the toolbar to select any of cumulative survival, cumulative incidence, and hazard rates. Use these graphs to ensure that the data for treatment effect are reasonable and appropriate.



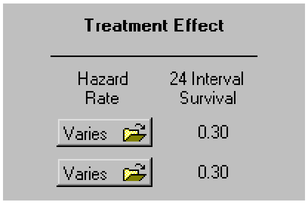
The graph should be consistent with your expectations. For example, if you enter a hazard rate of 0.05 per month and expect that this will yield 50% survival at 24 months, the graph should alert you if this is not the case.



The treatment effect shown is the treatment effect for which you compute power, so you should ensure that this effect is appropriate. In this example, the treatment effect which was entered as hazard rates of 0.05 versus 0.10 corresponds to cumulative survival of 0.09 versus 0.30, or a difference of 21% at 24 months. If it would be smaller to detect a smaller difference, or if it would not be important to detect a difference this small, then the treatment effect should be modified.

Treatment Effect When the Hazard Rate Varies

When the hazard rate varies by time interval, the hazard rate is shown as *Varies*.



Treatment Effect

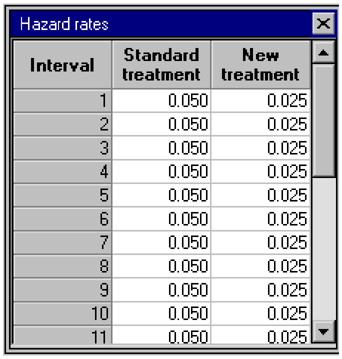
Hazard Rate	24 Interval Survival
Varies	0.30
Varies	0.30

- ❖ To open a grid, click *Varies*.
- ❖ Enter the hazard rate for each interval in the grid.
- ❖ Click the *Varies* to close the grid.

The program will display the cumulative survival at the final interval.

Tip

❖ To enter the same value for multiple cells, use CTRL-C to copy and then CTRL-V to paste.



Hazard rates

Interval	Standard treatment	New treatment
1	0.050	0.025
2	0.050	0.025
3	0.050	0.025
4	0.050	0.025
5	0.050	0.025
6	0.050	0.025
7	0.050	0.025
8	0.050	0.025
9	0.050	0.025
10	0.050	0.025
11	0.050	0.025

How Hazard Rates and the Hazard Ratio Affect Power

Power is driven in part by the number of events in the study. That is, the hazard rate must be high enough so that the event rate in the two groups can be compared. For this reason, if the hazard rate is very low, power will be adversely affected. As long as the hazard rates will yield a fair number of events, the primary factor affecting power will be the **hazard ratio,** or the ratio of hazard rates in one group as compared to the other group.

Attrition

In a survival study, patients can be lost for three reasons:

- They may reach the terminal event—for example, they may die.
- They may be censored because they are still being followed when the study ends.
- They may be lost for other reasons. For example, the patients may move out of the geographic area. Or, if the protocol requires that patients be maintained on a drug regimen, the patient may stop taking the drug and therefore be “lost” at that point.

The first two types of loss are accounted for automatically. That is, the number of patients that will be lost because they reach the criterion event (the number dying) is computed automatically from the hazard rates. Similarly, the number of patients who will be censored at the conclusion of the study is computed automatically based on the data from accrual and follow-up.

The third type of loss cannot be determined from the other values and must be provided by the user. This type of loss is referred to as drop out or attrition.

The program provides three options for attrition, and these are specified from the Procedures screen.

- If you have specified that there is no attrition, this section will not appear on the interactive screen.
- If you have specified that attrition is constant, the program will display one box for the attrition rate, which will be applied to each interval for both groups.
- If you have specified that attrition varies, the program will display a folder named Varies. Click this folder to open a grid, in which you may enter the attrition rate for each interval for each group.

How Attrition Affects Power

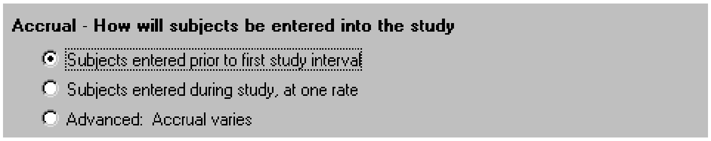
Attrition affects the number of patients in the study, and so the lower the attrition rate, the higher the power. The best case is when there is no attrition. If attrition exists, a low rate is preferable to a high one. If attrition varies, it will have more of an impact if it takes place early in the study, because this will result in the loss of more data. Beyond these simple rules, the impact of attrition in any given study depends on a host of other factors. For example, if the hazard rates are high, then most patients will die before they have a chance to drop out, and the impact of attrition may be low. By contrast, if the hazard rates are low and patients are expected to be alive (and followed) for many years, then even a modest attrition rate can severely impact the sample size and power.

Sample Size

The format of the sample size section depends on the type of accrual, which is set from the Procedures screen.

Sample Size When Accrual Is Prior to the Study

When accrual is prior to the study, use the option shown here prior to opening the module. Then, inside the module, enter the total number of subjects. There is no need to enter the accrual rate since all subjects are entered prior to the start of the study.

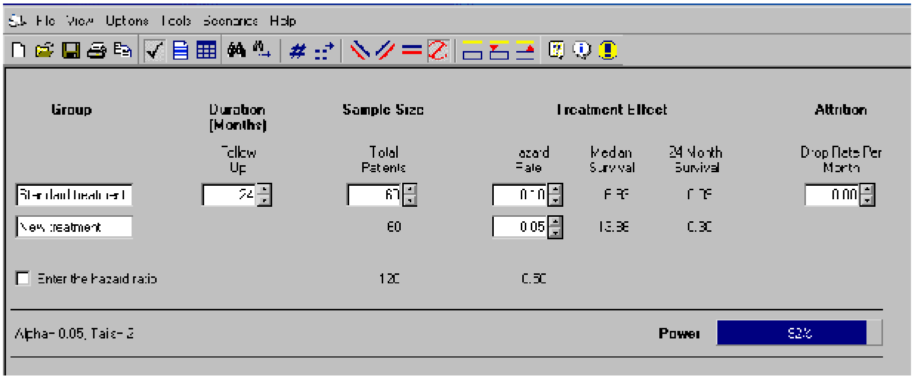


Accrual - How will subjects be entered into the study

● Subjects entered prior to first study interval

○ Subjects entered during study, at one rate

○ Advanced: Accrual varies



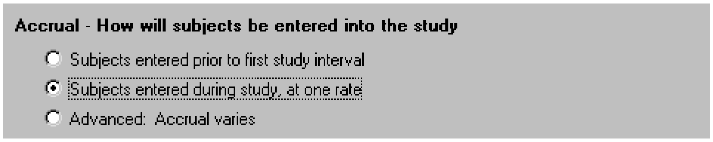
Group	Duration (Months)	Sample Size	Treatment Effect			Attrition
	Follow Up	Total Patients	Hazard Rate	Median Survival	24 Month Survival	Drop Rate Per Month
Standard treatment	24	60	0.10	6.93	0.09	0.00
New treatment		60	0.05	13.86	0.30	
□ Enter the hazard ratio		120	0.50			

Alpha= 0.05, Tails= 2

Power 92%

Sample Size When Accrual Is During the Study, at a Constant Rate

When accrual is during the study at a constant rate—for example, 10 per month for the first 6 months of a 24-month study—use the option shown here prior to opening the module. Then, inside the module, enter the duration of accrual and the number of subjects.

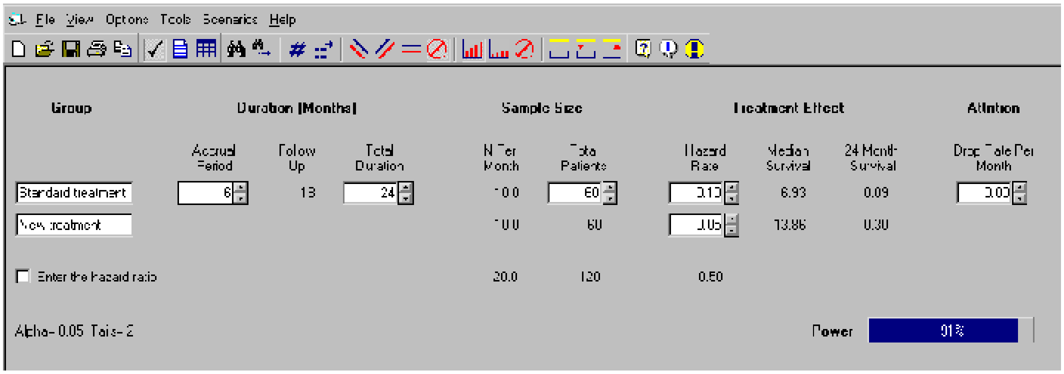


Accrual - How will subjects be entered into the study

○ Subjects entered prior to first study interval

● Subjects entered during study, at one rate

○ Advanced: Accrual varies



Group	Duration (Months)			Sample Size		Treatment Effect			Attrition
	Accrual Period	Follow Up	Total Duration	N Per Month	Total Patients	Hazard Rate	Median Survival	24 Month Survival	Drop Rate Per Month
Standard treatment	6	18	24	10.0	60	0.10	6.93	0.09	0.00
New treatment				10.0	60	0.05	13.86	0.30	
□ Enter the hazard ratio				20.0	120	0.50			

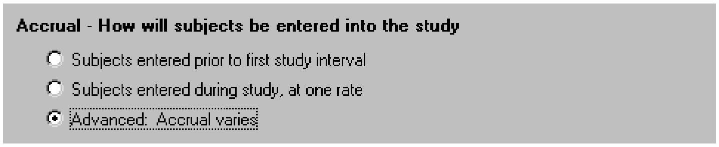
Alpha= 0.05 Tails= 2

Power 91%

When you enter the total number of subjects, the program computes the number per interval. Or, when you enter the number per interval, the program computes the total number. To switch between these options choose *Customize screen* from the Options menu. The number of intervals is set toward the left of the screen, under Duration.

Sample Size When Accrual Varies

When accrual varies from one interval to the next, or starts prior to the study and continues into the study, use the option shown here prior to opening the module. Then, inside the module, enter the duration of accrual and the number of subjects.



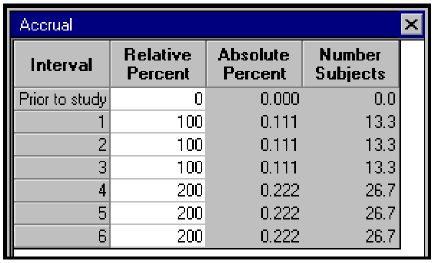
Accrual - How will subjects be entered into the study

○ Subjects entered prior to first study interval

○ Subjects entered during study, at one rate

● Advanced: Accrual varies

When accrual varies, you specify the accrual period—for example, six months—and the total sample size. The program opens a grid with a row for *Prior to study* and a row for each of the accrual intervals—six in this example. Enter the relative proportion of subjects entered in each interval. For example, 0, 100, 100, 100, 200, 200, 200, would signify that no subjects are entered prior to the study, that some would be entered for each of the first three months, and twice that number would be entered for each of the next three months. The program immediately shows these as absolute percentages and as number of patients. When the number of patients is modified, the actual number to be entered in each interval is then computed or modified automatically.



Accrual

Interval	Relative Percent	Absolute Percent	Number Subjects
Prior to study	0	0.000	0.0
1	100	0.111	13.3
2	100	0.111	13.3
3	100	0.111	13.3
4	200	0.222	26.7
5	200	0.222	26.7
6	200	0.222	26.7

Tips

The program will display a bar graph showing either the proportion of subjects entered during each interval or the number of subjects entered during each interval.

Spin Control may be used to modify the sample size quickly.

❖ To specify the size of the increment, click *N of Cases*.

or

❖ Click the *Find N* icon to have the program find the number of cases required for any given level of power.

By default, the program assumes that patients will be assigned to the two groups in equal numbers. However, you may specify some other ratio such as 1:2 or 2:3. Or, you may elect to enter the number of subjects separately for each group.

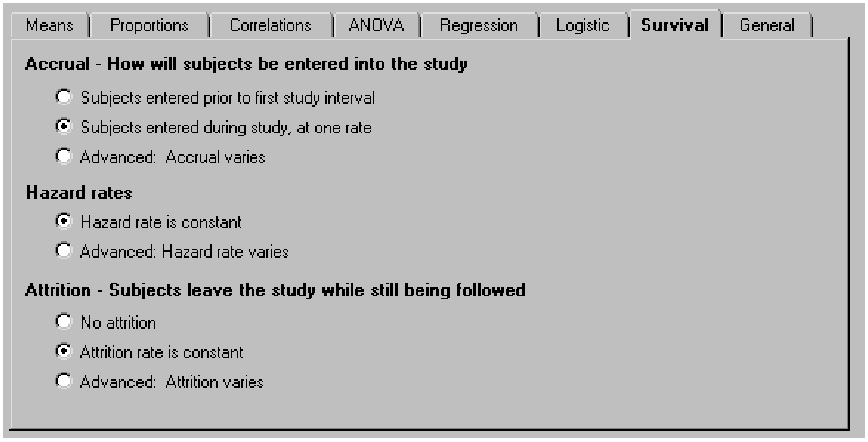
❖ To select these options, click *Sample Size* or choose *N-Cases* on the Options menu.

Example 1

Synopsis

A researcher is planning a study in which patients with cancer will be assigned randomly to one of two treatment groups. The standard treatment is surgery, and the new treatment is surgery plus radiation. The outcome is time to event, where Event is recurrence of the cancer as determined by diagnostic tests.

The accrual rate, the hazard rate, and the attrition rate will all be constant. These options are selected from the Procedures screen.



Accrual - How will subjects be entered into the study

○ Subjects entered prior to first study interval
● Subjects entered during study, at one rate
○ Advanced: Accrual varies

Hazard rates

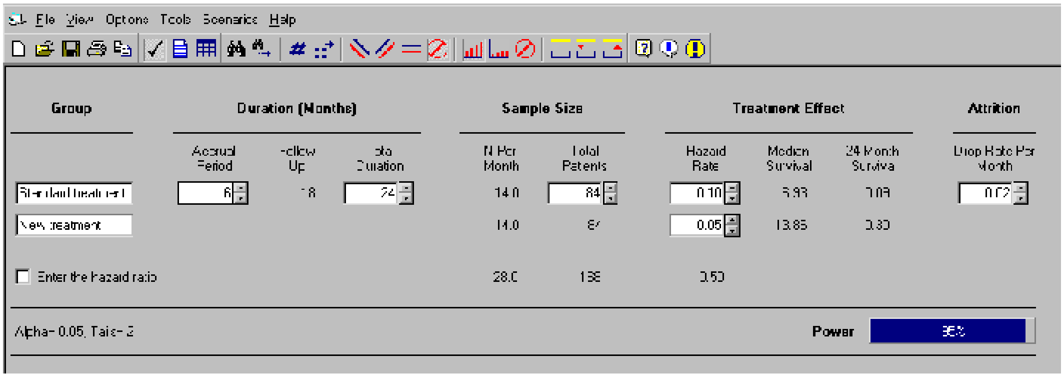
● Hazard rate is constant
○ Advanced: Hazard rate varies

Attrition - Subjects leave the study while still being followed

○ No attrition
● Attrition rate is constant
○ Advanced: Attrition varies

Interactive Computations

This set of options yields the following interactive screen.



Group	Duration (Months)			Sample Size		Treatment Effect			Attrition
	Accrual Period	Follow Up	Study Duration	N Per Month	Total Patients	Hazard Rate	Median Survival	24 Month Survival	Drop Rate Per Month
Standard treatment	6	18	24	14.0	84	0.10	6.93	0.09	0.02
New treatment				14.0	84	0.05	13.86	0.30	
□ Enter the hazard ratio				28.0	168	0.50			

Alpha= 0.05, Tails= 2

Power 85%

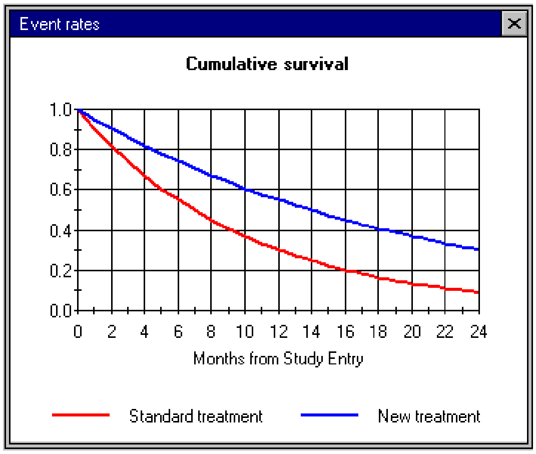
To customize the display:

- Choose *Customize screen* from the Options menu. Specify that subjects are *Patients* and the time interval is *Months*.

Duration. Under Duration, enter 6 months for the accrual period and an additional 18 months for follow-up, for a total study duration of 24 months. In other words, a patient entered on the first day of the study could be followed for as long as 24 months while a patient entered on the last day of the accrual period could be followed for as long as 18 months.

Hazard rates. Under Hazard Rates, enter an instantaneous hazard rate of 0.10 for the standard treatment and 0.05 for the new treatment. The program shows that this corresponds to median survival of 6.93 months for the standard treatment (50% of the patients will survive for 6.93 months) and 13.86 months for the new treatment. This also corresponds to cumulative survival rates at 24 months of 0.09 for the standard treatment and 0.30 for the new treatment. The program also shows the survival curves for the two groups graphically.

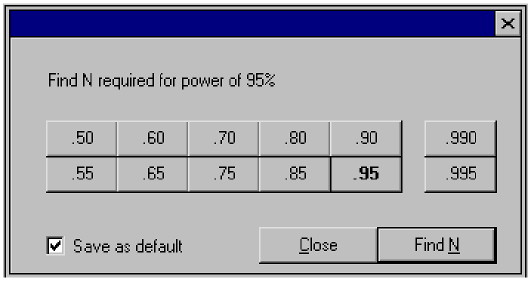
Alternatively, we could have entered the median survival or the cumulative survival Click *Treatment Effect* to display these options.



Attrition. In survival analysis, patients may be lost to follow-up for either of two reasons. First, they may be lost because the study ends while the patient is still being followed. The impact of this loss is computed automatically based on the study duration, and no entry is required. Second, they may be lost because they drop out of the study. For example, a patient might stop taking the drug or may move out of the country and be lost to follow-up. This second source of loss is addressed by attrition and is entered here as 2% per month.

Alpha and tails. Click *Alpha* and enter values for alpha and tails, 0.05 and 2, respectively.

Sample size. Click *Find N for power of 95%*. Under Sample Size, patients will be entered into each group at a rate of 14 per month for the duration of the accrual period (6 months). The program shows both the number entered per month (14) and the total number entered per group (84). Alternatively, you can modify the sample size directly. Also, you can specify that patients should enter the two groups in another ratio—for example, 2:1.



Find N required for power of 95%

☑ Save as default

Power. The program shows power. For the combination of accrual, sample size, hazard rates, attrition, alpha and tails, power is 95%. Modify any values on the screen and power changes accordingly.

To create a report:

❖ Click the Report icon on the toolbar. The program creates the following report:

Power for a test of the null hypothesis

One goal of the proposed study is to test the null hypothesis that the hazard rate, which is assumed to be constant across all study intervals, is identical in the two groups (standard treatment and new treatment).

Study design. This hypothesis will be tested in a study in which patients are entered and then followed until either (a) the terminal event occurs, or (b) they drop out of the study, or (c) the study ends and the patient is censored while still being actively followed. The study design calls for an accrual period of 6 months and a follow-up period of 18 months. In other words, the first patient to enter the study will be followed for a maximum of 24 months while the last patient to enter the study will be followed for a maximum of 18 months.

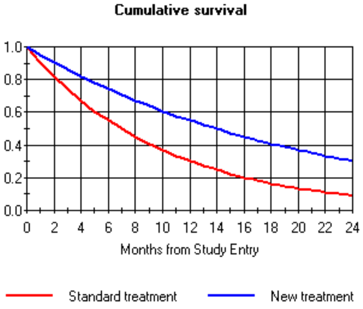
Effect size. To ensure that this effect size has been described accurately the program displays it in several formats. Computation of power is based on a hazard ratio of 0.50. Specifically, it assumes instantaneous hazard rates of 0.10 for the standard treatment group versus 0.05 for the new treatment group. Since the hazard rate is constant across intervals this is equivalent to median survival times of 6.93 months for the standard treatment group versus 13.86 months for the new treatment group. It is also equivalent to a cumulative survival at 24 months of 0.09 for the standard treatment group versus 0.30 for for the new treatment group (see graph). This effect was selected as the smallest effect that would be important to detect, in the sense that any smaller effect would not be of clinical or substantive significance. It is also assumed that this effect size is reasonable, in the sense that an effect of this magnitude could be anticipated in this field of research.

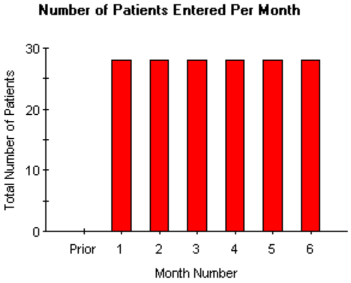
Sample size. Patients will be entered into the standard treatment group at the rate of 14.0 per month for 6 months yielding a total of 84 patients. Patients will be entered into the new treatment group at the rate of 14.0 per month for 6 months yielding a total of 84 patients.

Attrition. The computation assumes an attrition rate of 0.02 per month. This means that 2% of the patients who enter (for example) the second month of the study will drop out of the study during that month. This attrition (or drop-out) rate is separate from the censoring of patients that takes place when the study ends.

Alpha and Tails. The criterion for significance (alpha) has been set at 0.05. The test is 2-tailed, which means that an effect in either direction will be interpreted.

Power. For this study design, sample size, attrition rate, alpha and tails, and the population effect size described above, the study will have power of 95.2% to yield a statistically significant result.

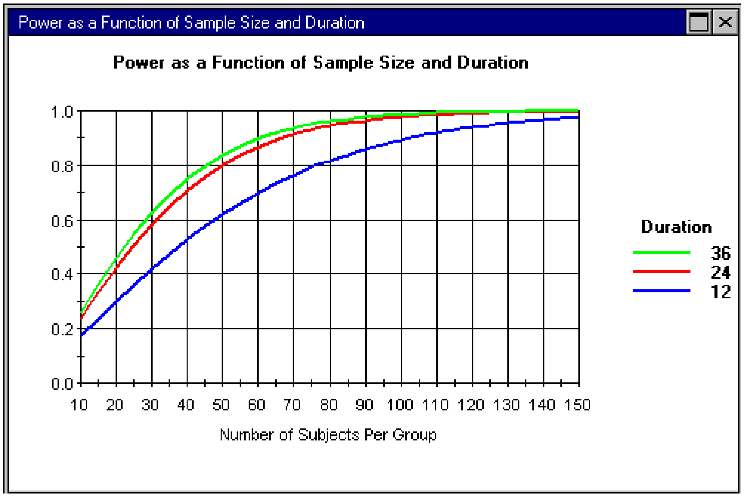




To create tables and graphs:

You may want to see how power would be affected if you changed the study design. For example, what would happen if the study duration were 12 months or 36 months rather than 24? This kind of question can be addressed interactively on the main screen, but the Tables and Graphs module provides a more comprehensive and faster approach.

❖ Click the *Tables and Graphs* icon on the toolbar. Add *Duration* as a factor and specify durations of 12, 24, and 36 months. The program creates the following graph.



The graph suggests that there is little to be gained by increasing the duration to 36 months. That is, you would run the study for an additional years and save little in sample size (only about 5 patients per group). The graph also shows that if you wanted to complete the study in 12 months, you would need an additional 30 patients per group to maintain the same level of power. On this basis, you may decide to stay with a 24-month study.

It is important to note that for another study, with different hazard rates, attrition rates, or hazard ratio, the relative position of these lines could shift drastically.

Example 2

Synopsis

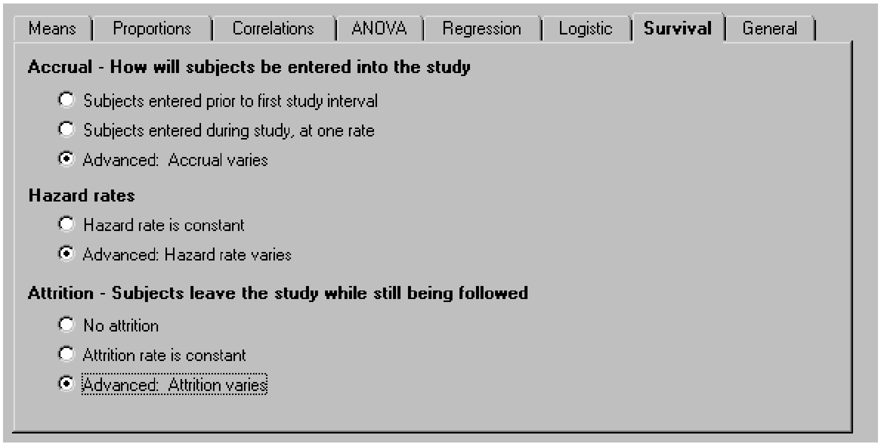
A researcher is planning a study in which patients with cancer will be assigned randomly to one of two treatment groups. The standard treatment is surgery, and the new treatment is a more radical form of surgery. The outcome is death from any cause.

The accrual rate will vary. The researcher anticipates that the staff will have trouble recruiting patients initially, but that the rate will increase after two months and will be stable after that.

The hazard rate will vary. Since only relatively healthy patients will be accepted for surgery, it is anticipated that few will die within six months of study entry. It is expected that the first deaths will occur at six months, and only at that point will the benefits of the new treatment become visible.

The attrition rate will vary. It is anticipated that few patients will drop out of the study in the first year, but 0.1 percent per month will drop out thereafter.

These options are selected from the Procedures screen.



Accrual - How will subjects be entered into the study

○ Subjects entered prior to first study interval
○ Subjects entered during study, at one rate
● Advanced: Accrual varies

Hazard rates

○ Hazard rate is constant
● Advanced: Hazard rate varies

Attrition - Subjects leave the study while still being followed

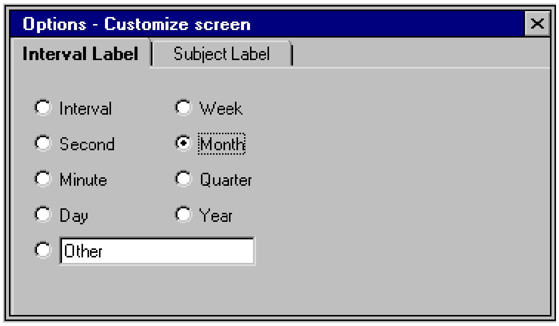
○ No attrition
○ Attrition rate is constant
● Advanced: Attrition varies

Interactive Computations

Time interval. The study will run for a total of four years. You need to treat this as 48 months rather than 4 years, because the accrual will vary by month, and you therefore need to provide information on a monthly basis.

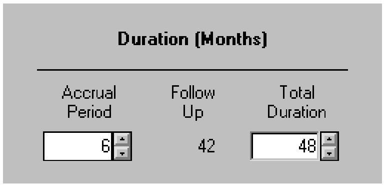
❖ To set the interval, choose *Customize screen* from the Options menu.

❖ Set *Months* as the interval and *Patients* as the subject label.



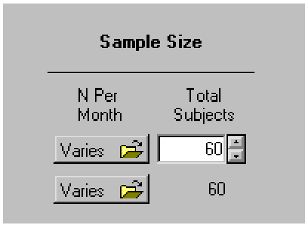
○ Interval ○ Week
○ Second ● Month
○ Minute ○ Quarter
○ Day ○ Year
○ Other

Accrual. The study will include an accrual period of 6 months and a follow-up period of 42 additional months. Enter these values, and the program shows a total duration of 48 months.



Accrual Period	Follow Up	Total Duration
6	42	48

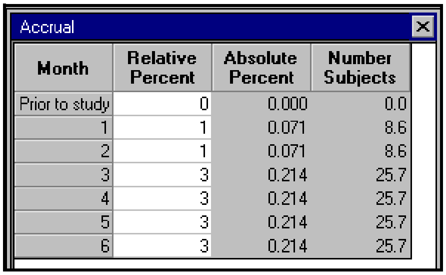
❖ To enter accrual data, click *Varies* under Sample size. The program displays a grid with seven rows—one for patients entered prior to the study and one for each of the six accrual periods.



Sample Size

N Per Month	Total Subjects
Varies	60
Varies	60

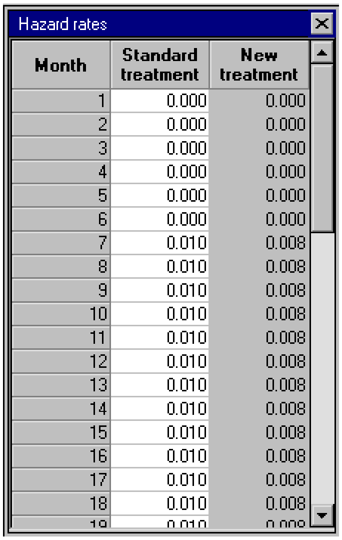
You want to enter a relatively small number of patients for two months and more for the next four months. Enter values of 1, 1, 3, 3, 3, 3,. These are relative values indicating that the accrual rate in the third month will be three times as high as that in the first month (values of 100, 100, 300, 300, 300, 300 would work as well). The program shows the corresponding percentages—7.1% in months 1 and 2 and 21.4% in later months. It also shows the actual number of patients to be entered in each month. Note that these numbers (actual sample size) are for the two groups combined and will change as the sample size changes. Click *Close* to hide the grid.



Accrual

Month	Relative Percent	Absolute Percent	Number Subjects
Prior to study	0	0.000	0.0
1	1	0.071	8.6
2	1	0.071	8.6
3	3	0.214	25.7
4	3	0.214	25.7
5	3	0.214	25.7
6	3	0.214	25.7

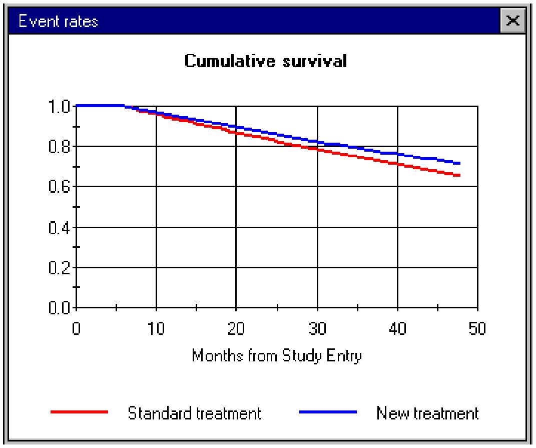
Hazard rates. The hazard rates will vary. Because one of the criteria for admission to the study is a relatively good prognosis, it is anticipated that no patients will die for at least six months. For the standard group, enter 0 in the first row, then copy this to the first six rows. Enter 0.01 for the 7th row and copy this to the next 42 rows. For the second group, you could enter data for each row. However, since the hazard rate for the second group is expected to be proportional to the first group, you will simply define a link between the two. Select *Enter hazard ratio* and provide a value of 0.8. In the grid, the program grays-out the column for the second group to indicate that these values will be computed, and then inserts the values—0 for the first 6 months and 0.008 thereafter. To hide the grid, click *Close.*



Hazard rates

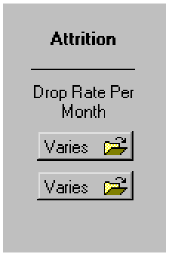
Month	Standard treatment	New treatment
1	0.000	0.000
2	0.000	0.000
3	0.000	0.000
4	0.000	0.000
5	0.000	0.000
6	0.000	0.000
7	0.010	0.008
8	0.010	0.008
9	0.010	0.008
10	0.010	0.008
11	0.010	0.008
12	0.010	0.008
13	0.010	0.008
14	0.010	0.008
15	0.010	0.008
16	0.010	0.008
17	0.010	0.008
18	0.010	0.008
19	0.010	0.008

The program shows this data graphically. The first graph shows cumulative survival and the second shows hazard rates. Use the toolbar to switch between different views of the graph to ensure that the data is correct. Here, for example, the survival is 100% for the initial 6 months.



Attrition. The attrition rate will vary. It is anticipated that few patients will drop out of the study in the first year, but that 0.1 percent per month will drop out thereafter.

❖ To enter attrition data, click *Varies* under Attrition.

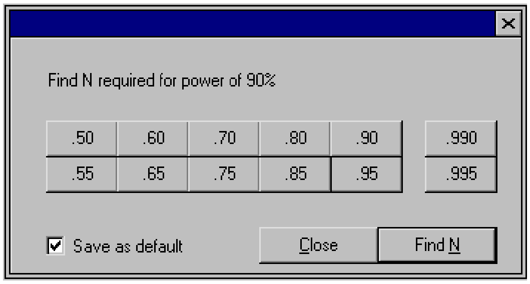


❖ Enter 0 in the first row and copy this to all cells for both groups—standard and new treatments—for the first 12 months. Enter 0.001 for the 13th row and copy that value to all subsequent rows for all columns.

Alpha and tails. Set alpha at 0.05. The test is two-tailed.

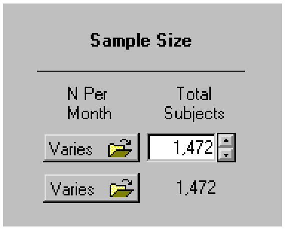
Sample size. You want to find the sample size that would be required to yield power of 90%.

❖ Click *Find N.*


Find N required for power of 90%

☑ Save as default

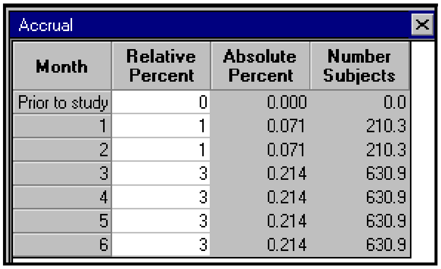
The program shows that a sample of 1,472 per group will be required.



Sample Size

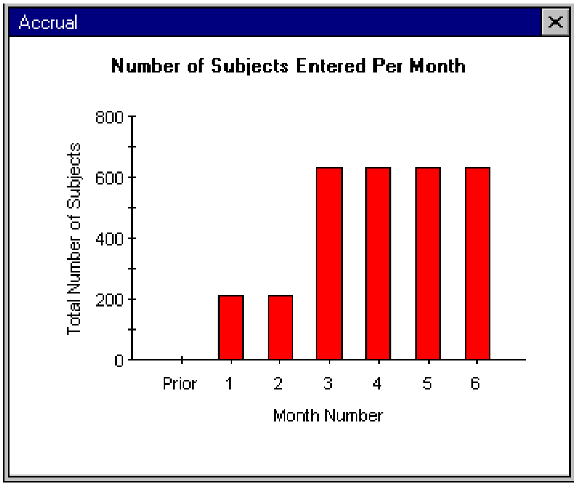
N Per Month	Total Subjects
Varies	1,472
Varies	1,472

This corresponds to 210.2 patients per month initially and 630.9 patients per month thereafter. This data can also be displayed as a bar graph.



Accrual

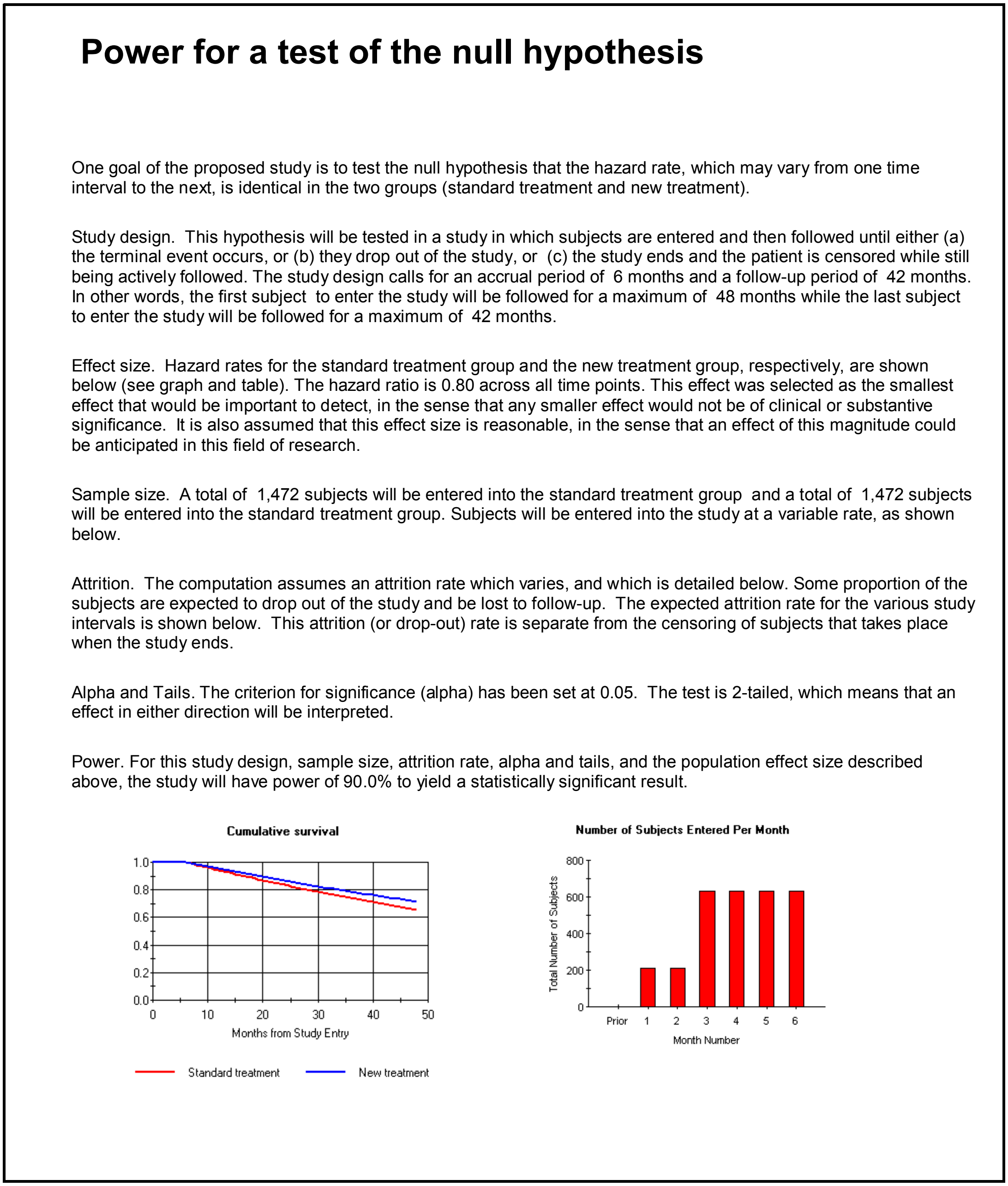
Month	Relative Percent	Absolute Percent	Number Subjects
Prior to study	0	0.000	0.0
1	1	0.071	210.3
2	1	0.071	210.3
3	3	0.214	630.9
4	3	0.214	630.9
5	3	0.214	630.9
6	3	0.214	630.9



To create a report:

❖ Click the *Report* icon on the toolbar.

The program creates the following report:



Power for a test of the null hypothesis

One goal of the proposed study is to test the null hypothesis that the hazard rate, which may vary from one time interval to the next, is identical in the two groups (standard treatment and new treatment).

Study design. This hypothesis will be tested in a study in which subjects are entered and then followed until either (a) the terminal event occurs, or (b) they drop out of the study, or (c) the study ends and the patient is censored while still being actively followed. The study design calls for an accrual period of 6 months and a follow-up period of 42 months. In other words, the first subject to enter the study will be followed for a maximum of 48 months while the last subject to enter the study will be followed for a maximum of 42 months.

Effect size. Hazard rates for the standard treatment group and the new treatment group, respectively, are shown below (see graph and table). The hazard ratio is 0.80 across all time points. This effect was selected as the smallest effect that would be important to detect, in the sense that any smaller effect would not be of clinical or substantive significance. It is also assumed that this effect size is reasonable, in the sense that an effect of this magnitude could be anticipated in this field of research.

Sample size. A total of 1,472 subjects will be entered into the standard treatment group and a total of 1,472 subjects will be entered into the standard treatment group. Subjects will be entered into the study at a variable rate, as shown below.

Attrition. The computation assumes an attrition rate which varies, and which is detailed below. Some proportion of the subjects are expected to drop out of the study and be lost to follow-up. The expected attrition rate for the various study intervals is shown below. This attrition (or drop-out) rate is separate from the censoring of subjects that takes place when the study ends.

Alpha and Tails. The criterion for significance (alpha) has been set at 0.05. The test is 2-tailed, which means that an effect in either direction will be interpreted.

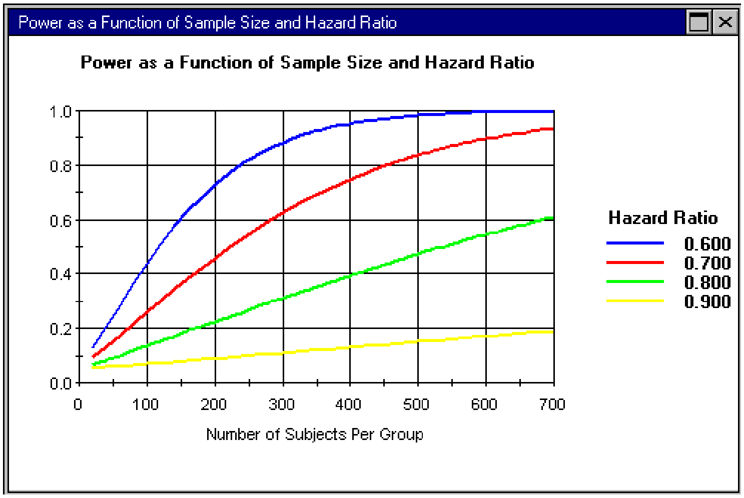
Power. For this study design, sample size, attrition rate, alpha and tails, and the population effect size described above, the study will have power of 90.0% to yield a statistically significant result.

To create tables and graphs:

You may want to see how power would be affected if you changed the study design. For example, what would happen if the hazard ratio was lower or higher? This kind of question can be addressed interactively on the main screen, but the Tables and Graphs module provides a more comprehensive and faster approach.

- Click the *Tables and Graphs* icon on the Toolbar. Add *Hazard Ratio* as a factor and specify ratios of 0.6, 0.7, 0.8, and 0.9.

The graph shows that with this sample size you will have power exceeding 0.99 if the hazard ratio is more compelling (0.7 or 0.6). Power, however, will be quite low (below 40%) if the hazard ratio is 0.9.



22 Equivalence Tests (Means)

Application

Equivalence tests are used when we want to prove that two groups are equivalent. The Equivalence Tests module is used to compare two means. (Another module is used to compare the event rates in two groups.)

The Equivalence Tests module is used, for example, if the study goal is to show that a newer version of an asthma drug is equivalent to the standard version of the drug when response is assessed using a continuous scale. In this example, patients are asked to rate their degree of relief on a 100-point scale. If the standard drug yields a mean relief score of 70 (where higher scores indicate more relief), then our goal may be to show that the new drug also yields a mean score of 70.

It is important to note that we can never prove that the response rate in two groups is *precisely* identical, since this would require an infinite sample size. Rather, our goal is to prove that the mean in the two groups is **comparable**, or **identical within limits**. For example, we may decide that the new drug is comparable if it yields a mean rate within five scale points of the standard treatment. (How we define this **acceptable difference** will depend on the substantive goals of the application and is discussed below.)

It is also important to note that the absence of a statistically significant effect can never be taken as evidence that two treatments are equally effective. In this example, if we pose the null hypothesis that the treatments are equally effective, test the null hypothesis, and obtain a p-value of 0.8 we *cannot* accept this as evidence of equality because the lack of significance may be due to inadequate sample size.

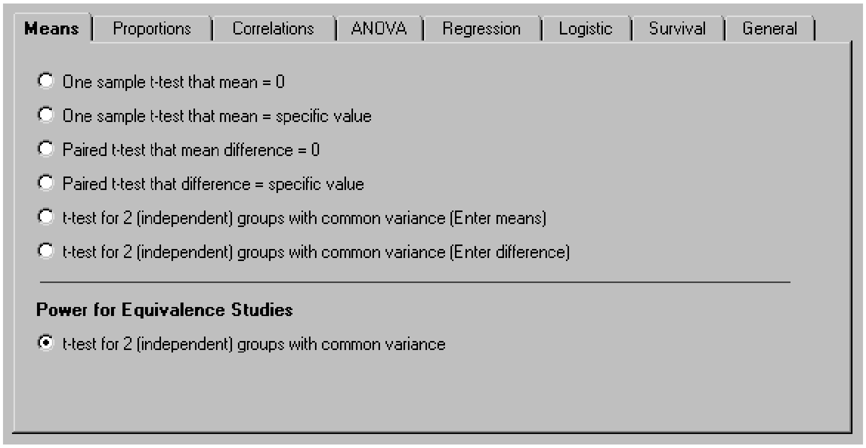
Rather, we would pose the null hypothesis that the mean score for the new drug is five points lower than the response rate for the standard drug, and we would test this null hypothesis. If this null hypothesis can be rejected, then we conclude that the mean for the new drug is comparable (within five points) to the standard drug.

As always, the power analysis that precedes the study must mirror the analysis that will follow the study. Therefore, in this example, we need to compute the power to reject the null hypothesis that the new drug is worse by five scale points.

Selecting the Procedure

To display the available procedures, choose *New analysis* from the File menu.

- ❖ Click the *Means* tab.
- ❖ Under Power for Equivalence Studies, select *t-test for 2 independent groups with common variance*.
- ❖ Click *OK* to proceed to the module.

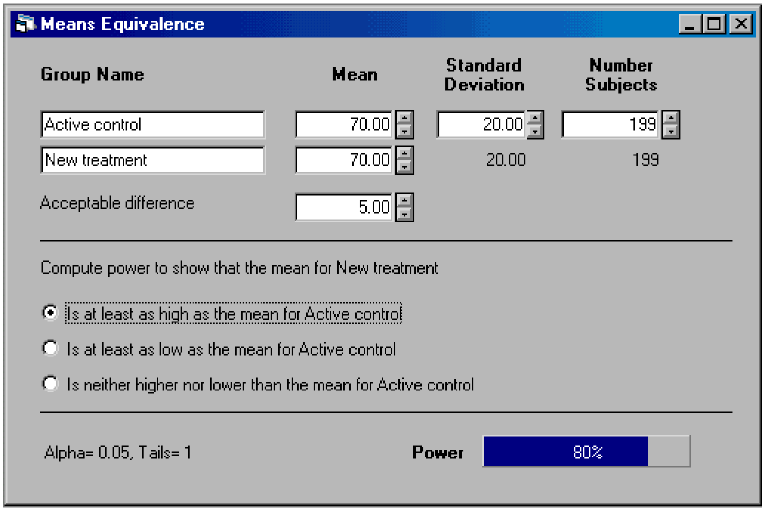


○ One sample t-test that mean = 0
○ One sample t-test that mean = specific value
○ Paired t-test that mean difference = 0
○ Paired t-test that difference = specific value
○ t-test for 2 (independent) groups with common variance (Enter means)
○ t-test for 2 (independent) groups with common variance (Enter difference)

Power for Equivalence Studies

● t-test for 2 (independent) groups with common variance

The program displays the following interactive screen:



Group Name	Mean	Standard Deviation	Number Subjects
Active control	70.00	20.00	199
New treatment	70.00	20.00	199

Acceptable difference 5.00

Compute power to show that the mean for New treatment

● Is at least as high as the mean for Active control
○ Is at least as low as the mean for Active control
○ Is neither higher nor lower than the mean for Active control

Alpha= 0.05, Tails= 1 **Power** 80%

Labels

Initially, the two groups are named *Active control* and *New treatment*. You can modify these names as needed.

Mean and Standard Deviations

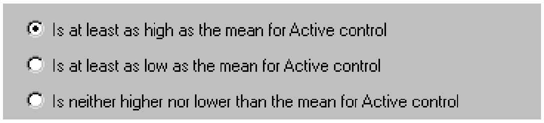
Enter the expected event mean for the two groups. For example, if the standard treatment is expected to yield a mean score of 70 and the new treatment is also expected to yield a mean score of 70, enter 0.70 for each. Enter the common within-group standard deviation (in this example, 20).

Acceptable Difference

As noted, it is never possible to prove that two treatments are precisely equal; you can prove only that they are **comparable**, or **equivalent within limits**. These limits are defined here. If the study goal is to show that the new treatment is comparable to the standard treatment within five scale points, type 5.0. The value entered here is typically the most important factor in determining power and therefore must be selected carefully.

Hypothesis to Be Tested

The program allows you test any of three hypotheses, which are selected from the following panel:



● Is at least as high as the mean for Active control

○ Is at least as low as the mean for Active control

○ Is neither higher nor lower than the mean for Active control

- The first option is used if a high score on the scale indicates a better outcome (for example, degree of relief) and we want to show that the mean score for the new treatment is as high as that for the standard treatment.
- The second option is used if a low score on the scale indicates a better outcome (for example, number of side effects) and we want to show that the mean for the new treatment is as low as the event rate for the standard treatment.
- The third option is used if, for example, we want to show that the mean for the new treatment is neither lower nor higher than the rate for the standard treatment. For example, if we are looking at the impact of drugs on blood pressure, we want to ensure that the mean blood pressure for the new treatment as compared with the standard treatment is neither lower nor higher.

How This Setting Affects Power

If the mean for the two groups is identical (70 and 70 in this example), then power will be the same for any of these three options. However, if the expected mean for the groups differs, then power will be quite different for option 1 as compared with option 2. Option 3 will always be the lower of options 1 and 2.

Alpha and Tails

The program shows alpha and tails. To modify, click *Alpha*.

The meaning of alpha in equivalence tests is the same as in the standard test of the null hypothesis. In this example, an alpha of 0.05 means that the standard of proof is set at five scale points, and we have only a 5% likelihood of rejecting the null hypothesis when, in fact, the new treatment is worse by five scale points.

The test is always one-tailed; therefore, the program does not allow the user to modify the setting for tails. When we want to show that the mean for the new treatment is at least as high as that for the standard treatment, we need to reject only the possibility that the mean is lower; thus, the test is one-tailed. When we want to show that the mean for the new treatment is at least as low as that for the standard treatment, we need to reject only the possibility that the mean is higher; thus, the test is one-tailed. When we want to test both of these and show that the mean is neither higher nor lower, we actually do each test separately, and each is done as a one-tailed test. No adjustment is made to alpha because it is not possible for the two (opposite) null hypotheses to be rejected in the same study.

Sample Size

Enter the sample size, or use the Find N icon to locate the sample size required for a given level of power.

Power

In this example, our study will have a power level of 80% to show that the mean for *New treatment* is at least as high as the mean for *Active control*. This assumes that the means for the *Active control* and *New treatment* populations are precisely equal (at 70), with a common within-group standard deviation of 20, that a difference of five points or less is unimportant, that the sample size in the two groups will be 199 and 199, and that alpha (one-tailed) is set at 0.05.

Formally, the null hypothesis is that the mean for *New treatment* is five points lower than the mean for *Active control*, and the study has a power of 80.1% to reject this null hypothesis. Equivalently, the likelihood is 80.1% that the 95.0% confidence interval for the mean difference will exclude a difference of five points in favor of *Active control*.

Example

Synopsis

Patients who arrive at hospital emergency rooms suffering with acute asthma are typically treated with a nebulizer for 15 minutes. After this treatment, the patients are asked to rate their degree of relief on a 100-point scale, with 100 being the best possible outcome for being able to breathe normally. This treatment yields a mean score of 70 (standard deviation of 20) on this scale. However, the drug has some side effects that patients would prefer to avoid.

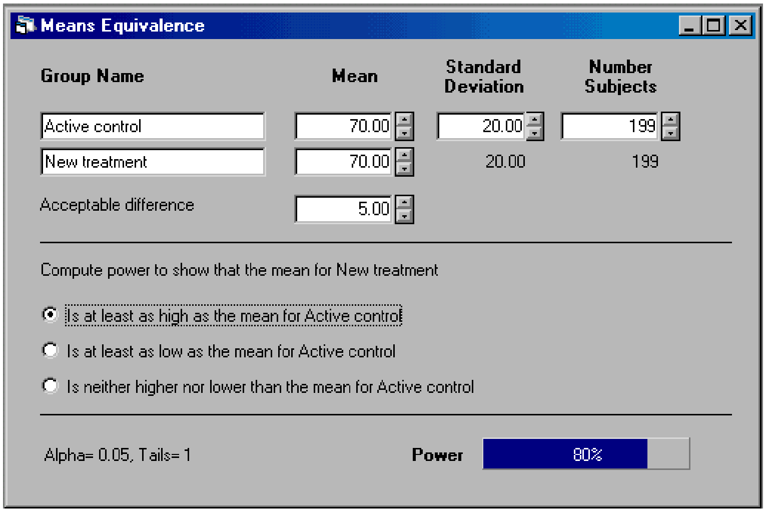
A drug company has developed an alternative formula for the medication. The new formula has fewer side effects and, for this reason, would be preferable. However, it must first be established that the drug is at least as effective as the standard formulation. It is decided that the new drug would be clinically useful if the mean relief score falls within five scale points of the standard drug.

To display the available procedures, choose *New analysis* from the File menu.

❖ Click the *Means* tab.

❖ Under Power for Equivalence Studies, select *t-test for 2 independent groups with common variance*.

The program displays the following interactive screen:



Group Name	Mean	Standard Deviation	Number Subjects
Active control	70.00	20.00	199
New treatment	70.00	20.00	199
Acceptable difference	5.00		

Compute power to show that the mean for New treatment

- ● Is at least as high as the mean for Active control
- ○ Is at least as low as the mean for Active control
- ○ Is neither higher nor lower than the mean for Active control

Alpha= 0.05, Tails= 1

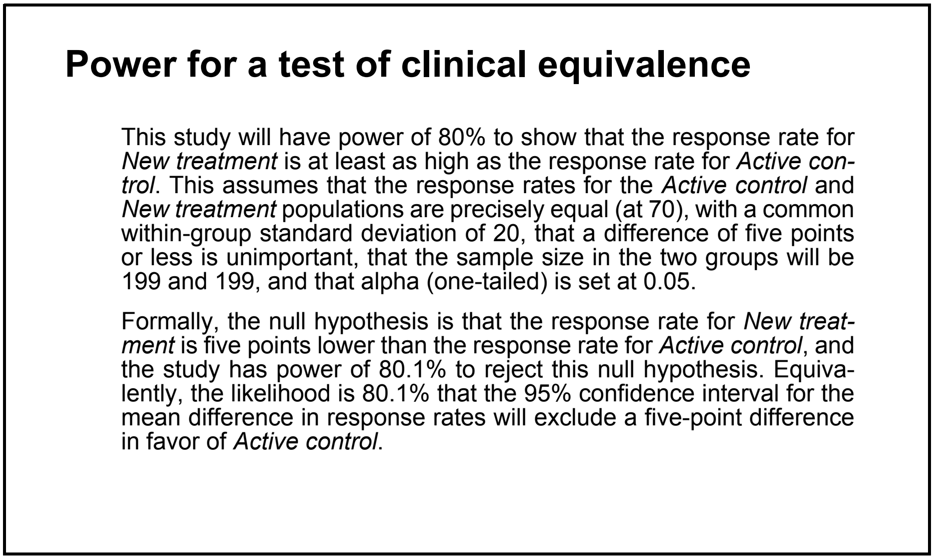
Power 80%

- ❖ Enter a mean of 70.00 for each of the groups.
- ❖ Enter a common within-group standard deviation of 20.
- ❖ Enter an acceptable difference of five points.
- ❖ Click *Alpha* and enter the desired value (0.05 in this example). The test must be one-tailed.
- ❖ Click the *Find N* icon and select a power level of 0.80.

The program shows that a sample of 199 patients per group will yield a power of 80%.

To create a report:

- ❖ Click the *Report* icon to generate the following report:



Power for a test of clinical equivalence

This study will have power of 80% to show that the response rate for *New treatment* is at least as high as the response rate for *Active control*. This assumes that the response rates for the *Active control* and *New treatment* populations are precisely equal (at 70), with a common within-group standard deviation of 20, that a difference of five points or less is unimportant, that the sample size in the two groups will be 199 and 199, and that alpha (one-tailed) is set at 0.05.

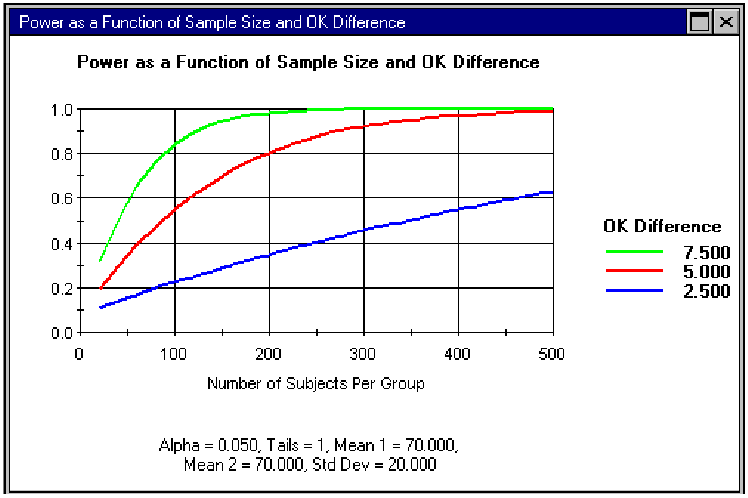
Formally, the null hypothesis is that the response rate for *New treatment* is five points lower than the response rate for *Active control*, and the study has power of 80.1% to reject this null hypothesis. Equivalently, the likelihood is 80.1% that the 95% confidence interval for the mean difference in response rates will exclude a five-point difference in favor of *Active control*.

To create tables and graphs:

At this point, we have established that we would need 199 patients per group to have adequate power to establish equivalence (if equivalence is defined with a margin of five scale points). We want to know how the sample size would be affected if we applied a stricter or more liberal criterion. That is, what if we required that the new treatment be as effective within 2.5 scale points, within 5 points (as before) or within 7.5 scale points?

- ❖ Click the *Table and graphs* icon.
- ❖ Click *Modify table*, and click the *Criterion* tab.
- ❖ Enter values of 2.5, 5.0, and 7.5 for the criteria.

The program displays the following graph:


Power as a Function of Sample Size and OK Difference
Power as a Function of Sample Size and OK Difference
1.0
0.8
0.6
0.4
0.2
0.0
0
100
200
300
400
500
Number of Subjects Per Group
OK Difference
🟩 7.500
🟥 5.000
🟦 2.500

Alpha = 0.050, Tails = 1, Mean 1 = 70.000,
Mean 2 = 70.000, Std Dev = 20.000

The middle line corresponds to the criterion of 0.05 and, as before, we see that we would need about 200 per group to yield a power level of 80%. The upper line corresponds to a criterion of 7.5 scale points. If we were willing to adopt this as our criterion for equivalence, we would need only about 90 patients per group to have the same level of power. The lower line corresponds to a criterion of 2.5 points. The graph shows that even with a sample of 500 per group, the power level would not be much higher than 60%.

We may decide to work with the criterion of 7.5 points to cut the sample substantially, or we may decide to stay with the 5-points criterion. It should be clear, however, that the selection of the criterion is critical and that even small variations in this number will have profound implications for power.

23 Equivalence Tests (Proportions)

Application

Equivalence tests are used when we want to prove that two groups are equivalent. This procedure is used when comparing two proportions, such as the response rate in two groups. (Another procedure is used when comparing the means in two groups.)

Use this procedure if, for example, the study goal is to show that a generic version of a drug is equivalent to the standard version of the drug. If the standard drug yields a response rate of 40%, then the goal may be to show that the new drug also yields a response rate of 40%.

It is important to note that we can never prove that the response rate in two groups is *precisely* identical because this would require an infinite sample size. Rather, our goal is to prove that the response rate in the two groups is **comparable**, or **equivalent within limits**. For example, we might decide that the new drug is comparable if it yields a response rate within five percentage points of the standard treatment. (How we define this **acceptable difference** will depend on the substantive goals of the application and is discussed below.)

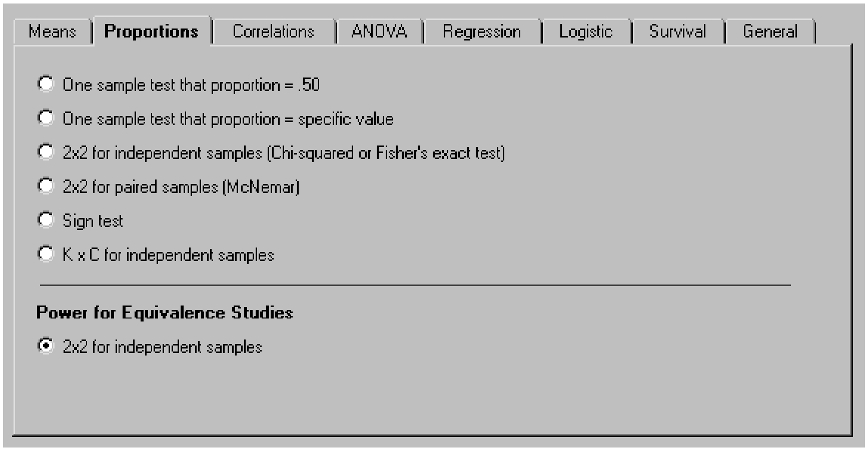
It is also important to note that the absence of a statistically significant effect can never be taken as evidence that two treatments are equally effective. In this example, if we pose the null hypothesis that the treatments are equally effective, test the null hypothesis, and obtain a p-value of 0.80, we *cannot* accept this as evidence of equality because the lack of significance may be due to inadequate sample size.

Rather, we would pose the null hypothesis that the response rate for the new drug is five percentage points lower than the response rate for the standard drug, and we would test this null hypothesis. If it can be rejected, then we conclude that the response rate for the new drug is comparable (within five points) to the standard drug.

As always, the power analysis that precedes the study must mirror the analysis that will follow the study. Therefore, in this example, we need to compute the power to reject the null hypothesis that the new drug is worse by five percentage points.

Selecting the Procedure

- To display the available procedures, choose *New analysis* from the File menu.
- Click the *Proportions* tab.
- Under Power for Equivalence Studies, select *2x2 for independent samples*.
- Click *OK* to proceed to the module.



○ One sample test that proportion = .50

○ One sample test that proportion = specific value

○ 2x2 for independent samples (Chi-squared or Fisher's exact test)

○ 2x2 for paired samples (McNemar)

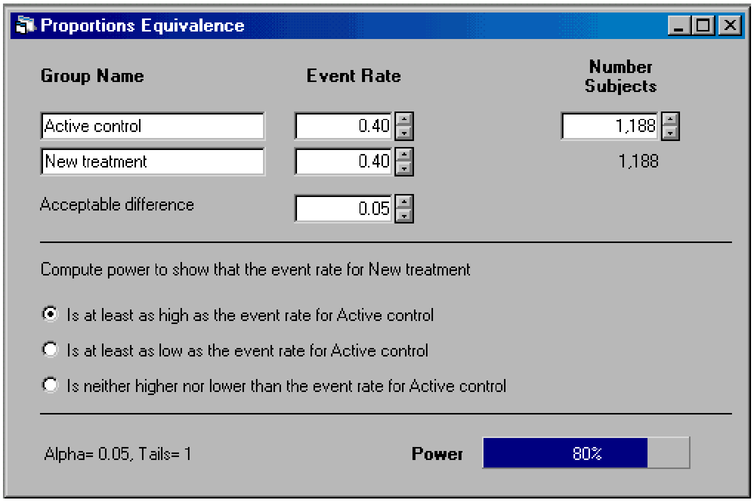
○ Sign test

○ K x C for independent samples

Power for Equivalence Studies

● 2x2 for independent samples

The program displays the following interactive screen:



Group Name	Event Rate	Number Subjects
Active control	0.40	1,188
New treatment	0.40	1,188
Acceptable difference	0.05	

Compute power to show that the event rate for New treatment

● Is at least as high as the event rate for Active control

○ Is at least as low as the event rate for Active control

○ Is neither higher nor lower than the event rate for Active control

Alpha= 0.05, Tails= 1

Power 80%

Labels

Initially, the two groups are named *Active control* and *New treatment*. You can modify these names as needed.

Customize Screen

Initially, the event rate is labeled *Event Rate*. It may be helpful to change this to *Response Rate* or some other label that reflects the study goals. To do this, choose *Customize screen* from the Options menu.

Event Rate

Enter the expected event rate for the two groups. For example, if the standard treatment is expected to yield a 40% response rate and the new treatment is also expected to yield a 40% response rate, type 0.40 for each.

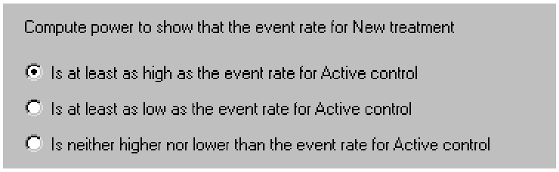
Acceptable Difference

As noted, it is never possible to prove that two treatments are precisely equal; you can prove only that they are **comparable**, or **equivalent within limits**. These limits are defined here. If the study goal is to show that the new treatment is comparable to the standard treatment within five percentage points, type 0.05.

The value entered here is typically the most important factor in determining power and, therefore, must be selected carefully. If we are working with a drug to prevent hay fever and the response rate is in the range of 40%, then an acceptable difference might be five percentage points. In contrast, if we are working with a drug to prevent tissue rejection following a transplant and the success rate is in the range of 95%, then an acceptable difference may be as little as one percentage point or less.

Hypothesis to Be Tested

The program allows you test any of three hypotheses, which are selected from the following panel:



Compute power to show that the event rate for New treatment

● Is at least as high as the event rate for Active control

○ Is at least as low as the event rate for Active control

○ Is neither higher nor lower than the event rate for Active control

- The first option is used if, for example, you are looking at response rates and want to show that the response rate for the new treatment is as high as that for the standard treatment.
- The second option is used if, for example, you are looking at adverse events and want to show that the event rate for the new treatment is as low as the event rate for the standard treatment.
- The third option is used if, for example, you want to show that the response rate for the new treatment is neither lower nor higher than the rate for the standard treatment.

How This Setting Affects Power

If the event rate in the two groups is identical (in this example, 0.40 and 0.40), then power will be the same for any of these three options. However, if the expected event rate for the groups differs, then power will be quite different for the first option compared with the second option. The third option will always be the lower of the first and second options.

Alpha and Tails

The program shows values for alpha and tails. To modify the value for alpha, click *Alpha*.

Note. The meaning of alpha in equivalence tests is the same as in the standard test of the null hypothesis. In this example, an alpha of 0.05 means that the standard of proof is set at 0.05 and that there is only a 5% likelihood of rejecting the null hypothesis when, in fact, the new treatment is worse by five percentage points.

The test is always one-tailed; therefore, the program does not allow the user to modify the setting for tails. When we want to show that the event rate for the new treatment is at least as high as that for the standard treatment, we need to reject only the possibility that the event rate is lower; thus, the test is one-tailed. When we want to show that the event rate for the new treatment is at least as low as that for the standard treatment, we need to reject only the possibility that event rate is higher; thus, the test is one-tailed. When we want to test both of these and show that the event rate is neither higher nor lower, we actually do each test separately, and each is done as a one-tailed test. No adjustment is made to alpha because it is not possible for the two (opposite) null hypotheses to be rejected in the same study.

Sample Size

Enter the sample size, or use the Find N icon to locate the sample size required for a given level of power.

Power

The program shows the level of power. In this example, the study will have power of 80% to show that the response rate for *New treatment* is at least as high as the response rate for *Active control*. This assumes that the response rates for the *Active control* and *New treatment* populations are precisely equal (at 40%), that a difference of five points or less is unimportant, that the sample size in the two groups will be 1188 and 1188, and that alpha (one-tailed) is set at 0.05.

Formally, the null hypothesis is that the response rate for *New treatment* is five percentage points lower than the response rate for *Active control*, and the study has power of 80% to reject the null hypothesis. Equivalently, the likelihood is 80% that the 95% confidence interval for the difference in response rates will exclude a five-point difference in favor of *Active control*.

Example 1

Synopsis

Patients who arrive at hospital emergency rooms suffering from acute asthma are typically treated with a nebulizer for 15 minutes. With response defined as the patient being able to leave the ER comfortably within an hour of treatment, this treatment proves effective in approximately 80% of patients treated. However, this drug has some side effects that patients would prefer to avoid.

A drug company has developed an alternative formula for the medication. The new formula has the advantage of having fewer side effects and, therefore, would be preferable. However, it must first be established that the drug is at least as effective as the standard formula. It is decided that the new drug would be clinically useful if the response rate falls within five percentage points of the standard treatment.

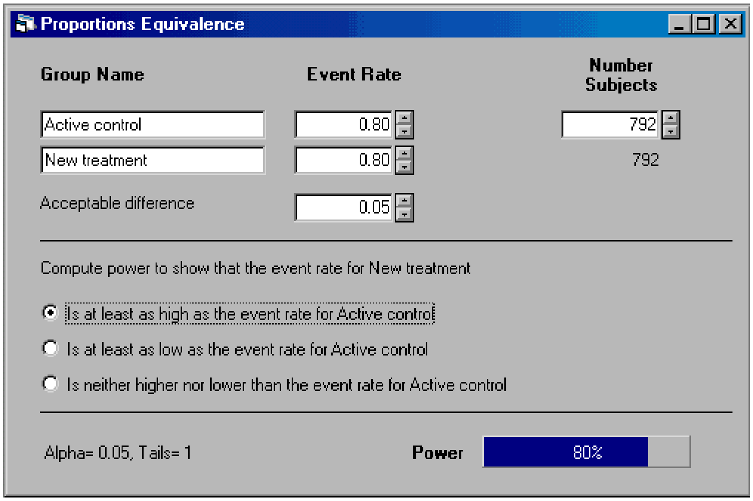
Interactive Computation

❖ Choose *New analysis* from the File menu.

❖ Click the *Proportions* tab.

❖ Under Power for Equivalence Studies, select *2x2 for independent samples*, and click *OK*.

The program displays the following interactive screen:



Proportions Equivalence

Group Name	Event Rate	Number Subjects
Active control	0.80	792
New treatment	0.80	792
Acceptable difference	0.05	

Compute power to show that the event rate for New treatment

● Is at least as high as the event rate for Active control

○ Is at least as low as the event rate for Active control

○ Is neither higher nor lower than the event rate for Active control

Alpha= 0.05, Tails= 1

Power 80%

❖ Choose *Customize screen* from the Options menu. Type *Response* as the label for events and *Patient* as the label for subjects.

❖ Enter response rates of 0.80 for each group.

❖ Enter an acceptable difference of 0.05.

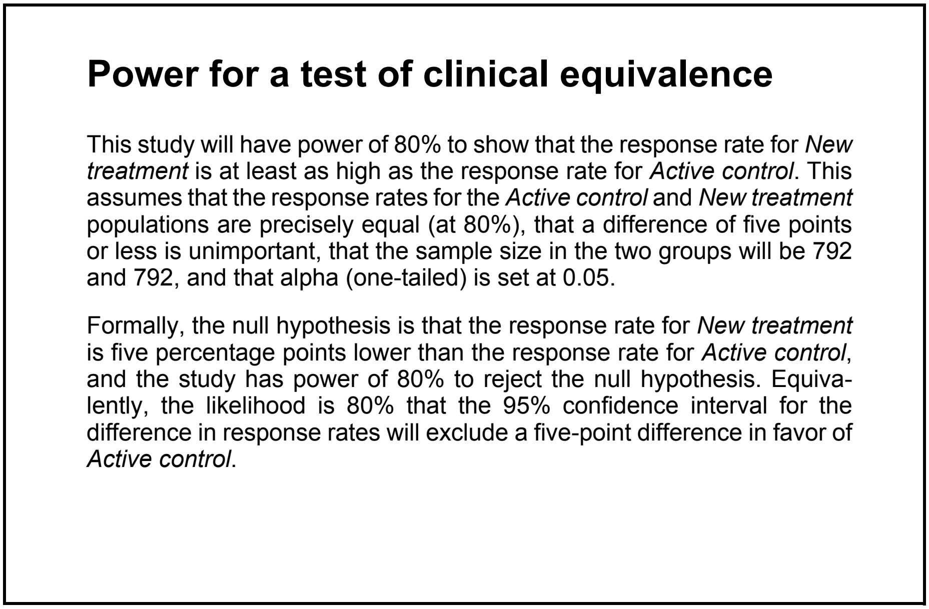
❖ Click *Alpha* and select the desired value (in this example, 0.05). The test must be one-tailed.

❖ Click the *Find N* icon and select a power level of 0.80.

The program shows that a sample of 792 patients per group will yield power of 80%.

To create a report:

- Click the *Report* icon to generate the following report:



Power for a test of clinical equivalence

This study will have power of 80% to show that the response rate for *New treatment* is at least as high as the response rate for *Active control*. This assumes that the response rates for the *Active control* and *New treatment* populations are precisely equal (at 80%), that a difference of five points or less is unimportant, that the sample size in the two groups will be 792 and 792, and that alpha (one-tailed) is set at 0.05.

Formally, the null hypothesis is that the response rate for *New treatment* is five percentage points lower than the response rate for *Active control*, and the study has power of 80% to reject the null hypothesis. Equivalently, the likelihood is 80% that the 95% confidence interval for the difference in response rates will exclude a five-point difference in favor of *Active control*.

To create tables and graphs:

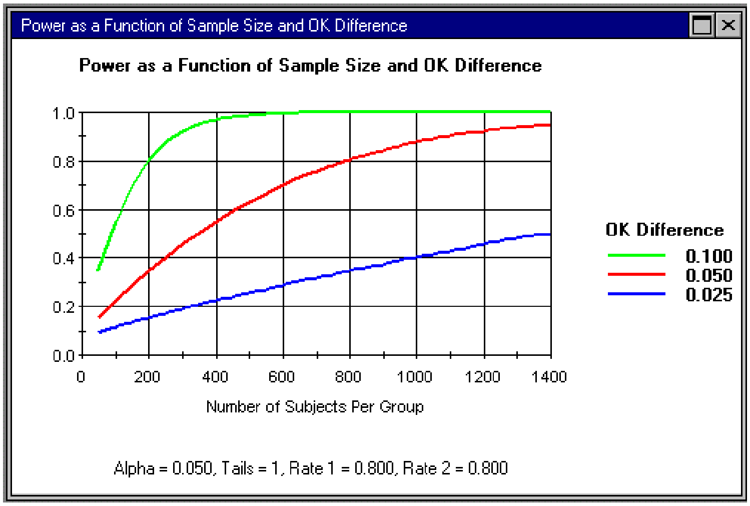
At this point, we have established that we would need 792 patients per group to have adequate power to establish equivalence if equivalence is defined with a five-point margin. We want to know how the sample size would be affected if we applied a stricter or a more liberal criterion. That is, what if we required that the new treatment be as effective within 2.5 percentage points, within 5 points (as before), or within 10 percentage points?

- Click the *Tables and graphs* icon.

- Click *Modify table* and click the *Criterion* tab.

- Enter values of 0.025, 0.050, and 0.010 for the criteria.

The program displays the following graph:



Alpha = 0.050, Tails = 1, Rate 1 = 0.800, Rate 2 = 0.800

The middle line corresponds to the criterion of 0.05 and, as before, we see that we would need about 800 per group to yield power of 80%. The upper line corresponds to a criterion of 10 percentage points. If we were willing to adopt this as our criterion for equivalence, we would need only about 200 patients per group to have the same level of power. The lower line corresponds to a criterion of 0.025 points. The graph shows that even with a sample of 1400 per group, power would be no higher than 50%. We may decide to work with the 10% criterion to cut the sample substantially, or we may decide to stay with the 5% criterion. It should be clear, however, that the selection of the criterion is critical and that even small variations in this number will have profound implications for power.

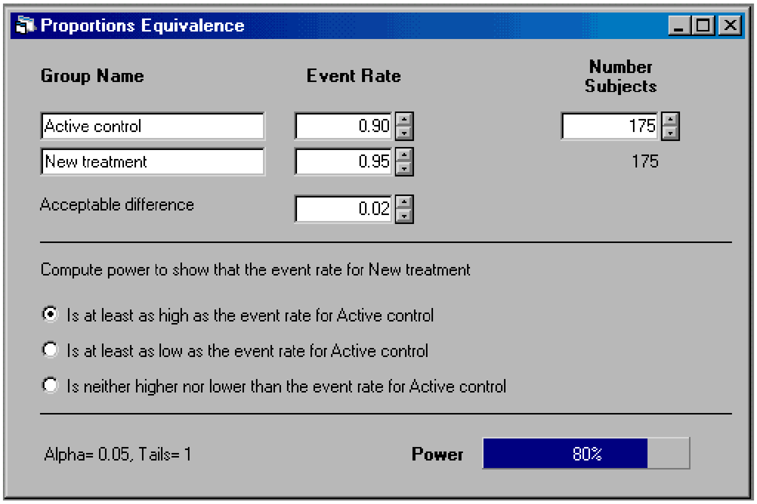
Example 2

A drug currently being used for treatment of athlete's foot is effective in approximately 90% of patients. The drug company has developed a new formula for the drug. The company believes that this formula will be effective against some additional strains and will therefore increase the effectiveness to 95% of patients. However, for regulatory purposes, the company needs to show only that the new drug is as effective as the standard drug. For the purposes of the test, this is defined as within two percentage points.

- Choose *New analysis* from the File menu.
- Click the *Proportions* tab.

- Under Power for Equivalence Studies, select *2x2 for independent samples*, and click *OK*.

  The program displays the following interactive screen:



Group Name	Event Rate	Number Subjects
Active control	0.90	175
New treatment	0.95	175
Acceptable difference	0.02	

Compute power to show that the event rate for New treatment

● Is at least as high as the event rate for Active control

○ Is at least as low as the event rate for Active control

○ Is neither higher nor lower than the event rate for Active control

Alpha= 0.05, Tails= 1

Power 80%

- Choose *Customize screen* from the Options menu. Type *Response* as the label for events and *Patient* as the label for subjects.

- Enter a response rate of 0.90 for the active control.

- Enter a response rate of 0.95 for the new treatment.

- Enter an acceptable difference of 0.02.

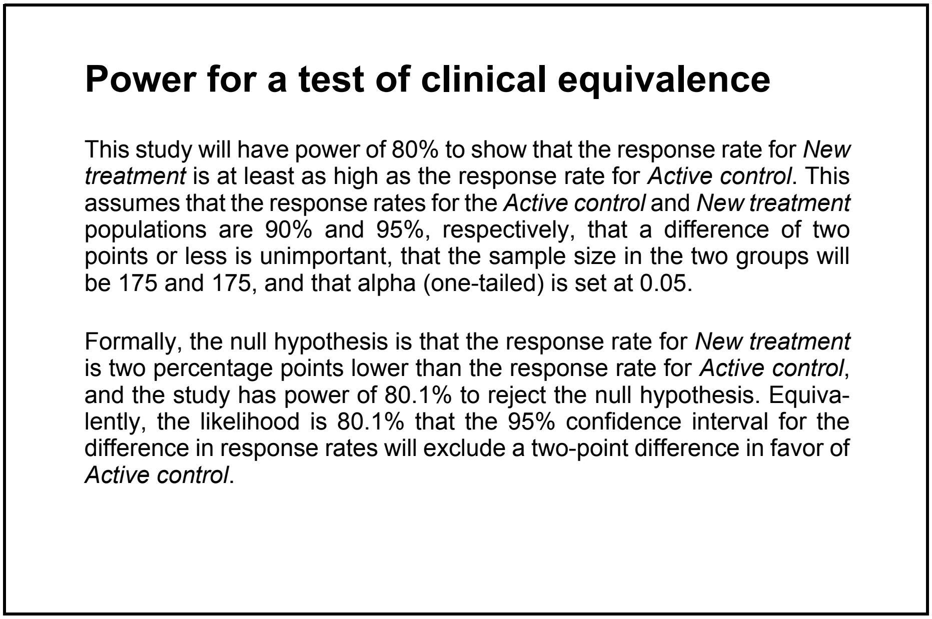
- Click *Alpha* and select the desired value (in this example, 0.05). The test must be one-tailed.

- Click the *Find N* icon and select a power level of 0.80.

  The program shows that a sample of 175 patients per group will yield power of 80%.

To create a report:

- Click the *Report* icon to generate the following report:



Power for a test of clinical equivalence

This study will have power of 80% to show that the response rate for *New treatment* is at least as high as the response rate for *Active control*. This assumes that the response rates for the *Active control* and *New treatment* populations are 90% and 95%, respectively, that a difference of two points or less is unimportant, that the sample size in the two groups will be 175 and 175, and that alpha (one-tailed) is set at 0.05.

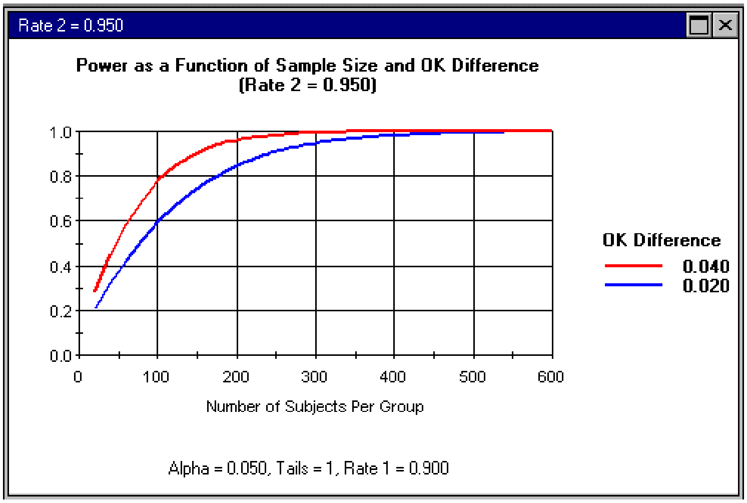
Formally, the null hypothesis is that the response rate for *New treatment* is two percentage points lower than the response rate for *Active control*, and the study has power of 80.1% to reject the null hypothesis. Equivalently, the likelihood is 80.1% that the 95% confidence interval for the difference in response rates will exclude a two-point difference in favor of *Active control*.

To create tables and graphs:

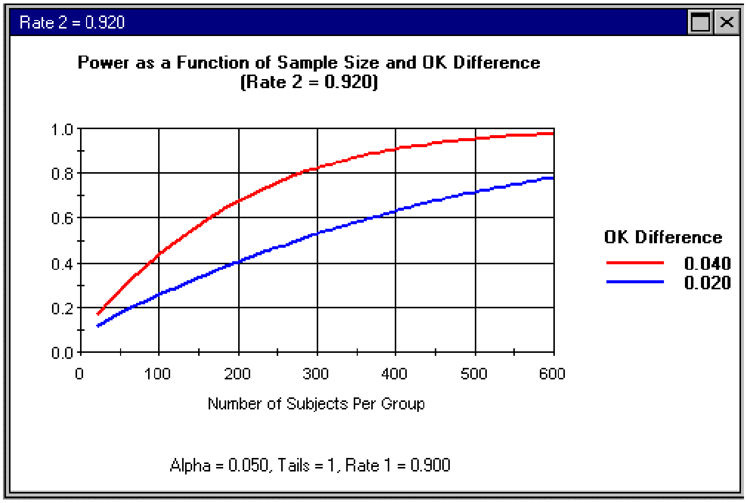
At this point, we have established that we would need 175 patients per group to have adequate power to establish equivalence if equivalence is defined with a two-point margin. We want to know how the sample size would be affected if we applied a more liberal criterion of 0.04. We also want to know how power would be affected if the actual response rate for the new treatment is not 95% but only 92%.

- Click the *Tables and graphs* icon.
- Click *Modify table*, and click the *Criterion* tab.
- Enter values of 0.02 and 0.04.
- Click the *Event rates* tab and enter values of 0.92 and 0.95 for the event rate in the second group.

The program displays the following two graphs:



Alpha = 0.050, Tails = 1, Rate 1 = 0.900



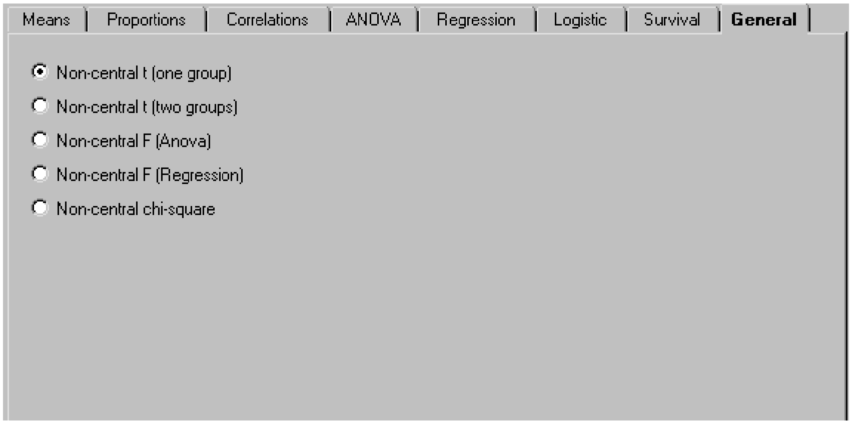
Alpha = 0.050, Tails = 1, Rate 1 = 0.900

These graphs show the following:

- If we are willing to assume that the event rate in the second group is 0.95, then we will need a sample size in the range of 100 per group (if we elect to use the more liberal criterion of 0.04) to 175 per group (for the more conservative criterion of 0.02).
- However, if we want to ensure that power is adequate even if the event rate for the second group is as low as 0.92, then the sample size would fall in the range of 300 per group for the more liberal criterion, or 600 per group for the more conservative criterion.

24 General Case

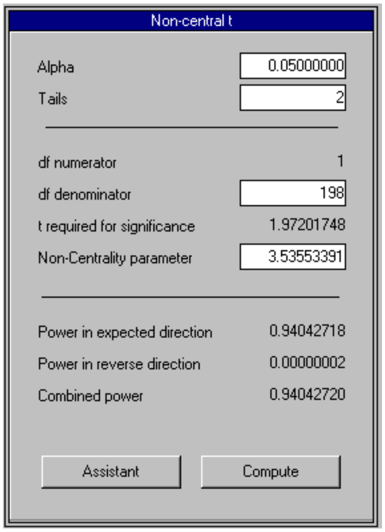
By selecting the General panel, the advanced user can compute power for procedures not otherwise included in the program.



● Non-central t (one group)
○ Non-central t (two groups)
○ Non-central F (Anova)
○ Non-central F (Regression)
○ Non-central chi-square

General Case (Non-Central T)

Figure 24.1 Non-central t (one group)



Non-central t

Alpha 0.05000000

Tails 2

df numerator 1

df denominator 198

t required for significance 1.97201748

Non-Centrality parameter 3.53553391

Power in expected direction 0.94042718

Power in reverse direction 0.00000002

Combined power 0.94042720

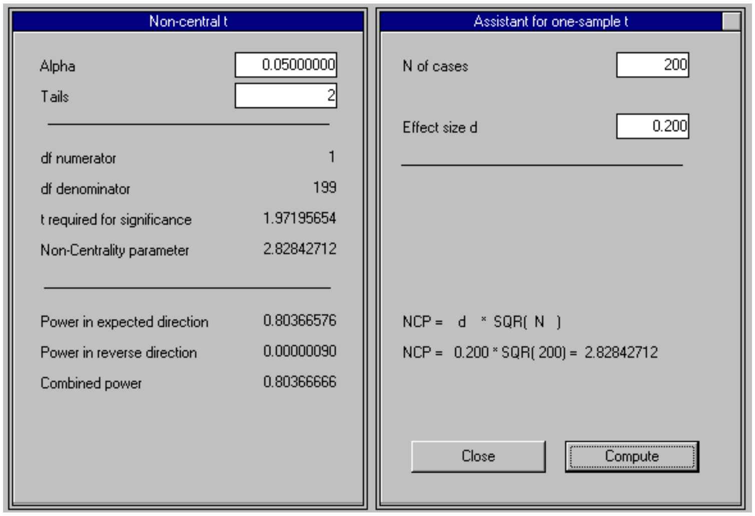
The user can enter alpha, tails, df$_{\text{Denominator}}$, and the non-centrality parameter (NCP). The program displays the t-value required for significance and power.

For a one-sample test, NCP $= d\sqrt{N}$, and df$_{\text{Denominator}}$ $= N - 1$.

For a two-sample test, NCP $= \dfrac{d\sqrt{n'}}{\sqrt{2}}$, where n' is the harmonic mean of N1 and N2, and df$_{\text{Denominator}}$ $= Ntotal - 2$.

The program also offers assistant panels. When the assistant is active, values for degrees of freedom (df) and NCP must be entered through the assistant. To enter these values directly, close the assistant.

Figure 24.2 Non-central t (one group) with assistant



Alpha 0.05000000
Tails 2

df numerator 1
df denominator 199
t required for significance 1.97195654
Non-Centrality parameter 2.82842712

Power in expected direction 0.80366576
Power in reverse direction 0.00000090
Combined power 0.80366666

N of cases 200

Effect size d 0.200

NCP = d * SQR(N)
NCP = 0.200 * SQR(200) = 2.82842712

The assistant panel for one-sample t-tests is shown in Figure 24.2.

❖ Enter the number of cases.

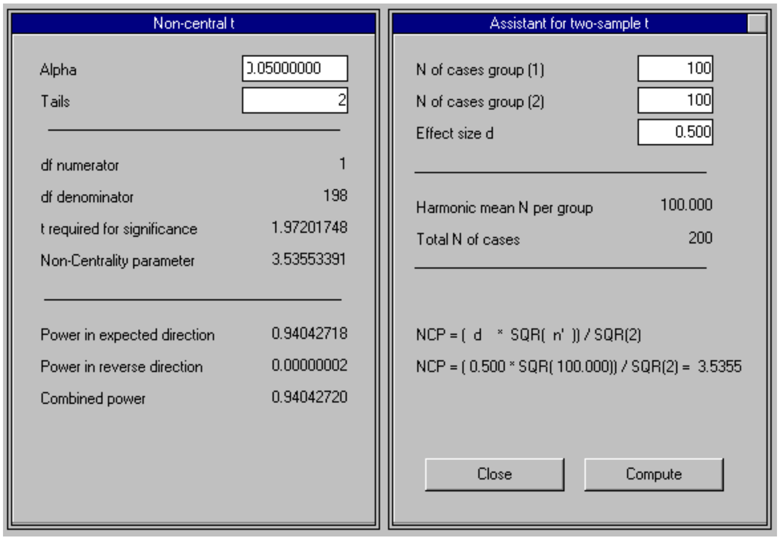
❖ Enter the effect size, d.

❖ Enter alpha and tails.

The program displays:

- The computed non-centrality parameter
- The required t
- Power

Figure 24.3 Non-central t (two-group) with assistant



```
Non-central t

Alpha                              0.05000000
Tails                                       2

df numerator                                1
df denominator                            198
t required for significance        1.97201748
Non-Centrality parameter           3.53553391

Power in expected direction        0.94042718
Power in reverse direction         0.00000002
Combined power                     0.94042720
```

```
Assistant for two-sample t

N of cases group (1)               100
N of cases group (2)               100
Effect size d                    0.500

Harmonic mean N per group      100.000
Total N of cases                   200

NCP = ( d   * SQR( n' )) / SQR(2)
NCP = ( 0.500 * SQR( 100.000)) / SQR(2) =  3.5355
```

The assistant panel for two-sample t-tests is shown in Figure 24.3.

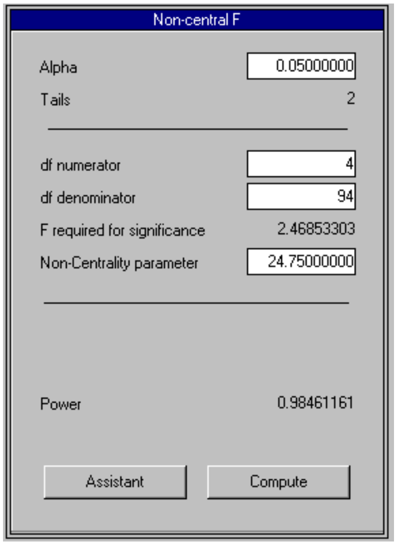
- Enter the number of cases.
- Enter the effect size, d.
- Enter alpha and tails.

The program displays:

- The computed non-centrality parameter
- The required t
- Power

General Case (Non-Central F)

Figure 24.4 Non-central F (ANOVA)



Non-central F

Alpha 0.05000000

Tails 2

df numerator 4

df denominator 94

F required for significance 2.46853303

Non-Centrality parameter 24.75000000

Power 0.98461161

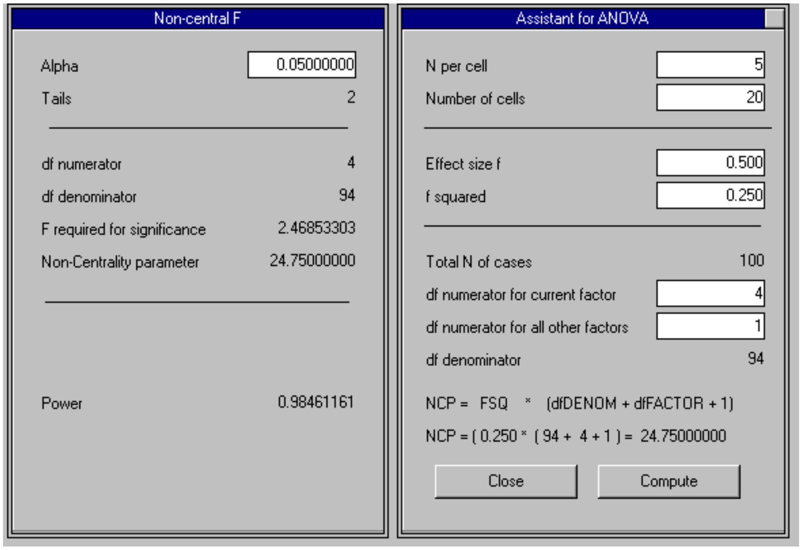
The general case for power based on non-central F is shown in Figure 24.4.

The user can enter alpha, tails, $df_{\text{Numerator}}$, $df_{\text{Denominator}}$, and the non-centrality parameter (NCP). The program displays the F-value required for significance and power.

NCP is computed as $f^2 * (df_{\text{Numerator}} + df_{\text{Error}} + 1)$. In the case of a oneway ANOVA or a multiple regression with only one set of variables, this is equivalent to $f^2 * \text{Ntotal}$. In other cases, however, the two formulas are not identical because the term in parentheses does not include the degrees of freedom (df) associated with other factors or covariates (in a factorial ANOVA) or other sets of variables (in a multiple regression).

The program also offers assistant panels for ANOVA and multiple regression. In these panels, the user can enter data that are used in intermediate computations and transferred to the first panel. When the assistant is active, values for df and NCP must be entered through the assistant. To enter these values directly, close the assistant.

Figure 24.5 Non-central F (ANOVA) and assistant



Alpha 0.05000000
Tails 2
df numerator 4
df denominator 94
F required for significance 2.46853303
Non-Centrality parameter 24.75000000
Power 0.98461161

N per cell 5
Number of cells 20
Effect size f 0.500
f squared 0.250
Total N of cases 100
df numerator for current factor 4
df numerator for all other factors 1
df denominator 94
NCP = FSQ * (dfDENOM + dfFACTOR + 1)
NCP = (0.250 * (94 + 4 + 1) = 24.75000000

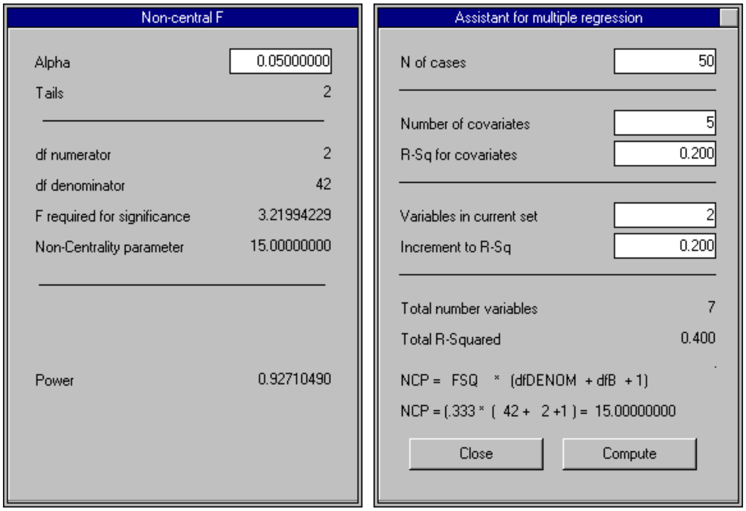
The assistant panel for ANOVA is shown in Figure 24.5.

- Enter alpha.
- Enter the number of cases per cell and the number of cells.
- Enter the effect size, f, or f^2 .
- Enter the $df_{\text{Numerator}}$ for the factor of interest.
- Enter the $df_{\text{Numerator}}$ for all other factors and covariates.
- The program computes the $df_{\text{Denominator}}$ as $\text{Ntotal} - df_{\text{Factor}} - df_{\text{Other}}$.
- Click *Compute*.

The program displays:

- The F-value required for significance
- The non-centrality parameter
- Power

Figure 24.6 Non-central F (multiple regression) and assistant



Non-central F

Alpha 0.05000000

Tails 2

df numerator 2

df denominator 42

F required for significance 3.21994229

Non-Centrality parameter 15.00000000

Power 0.92710490

Assistant for multiple regression

N of cases 50

Number of covariates 5

R-Sq for covariates 0.200

Variables in current set 2

Increment to R-Sq 0.200

Total number variables 7

Total R-Squared 0.400

NCP = FSQ * (dfDENOM + dfB + 1)

NCP = (.333 * (42 + 2 +1) = 15.00000000

The assistant panel for multiple regression is shown in Figure 24.6.

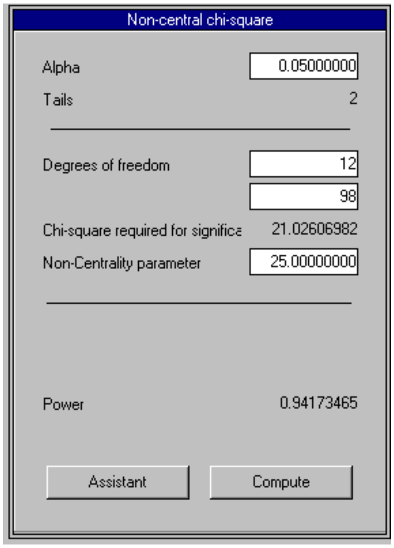
- ❖ Enter alpha.
- ❖ Enter the number of cases.
- ❖ Enter the number of covariates and the R^2 for these covariates.
- ❖ Enter the number of variables in the current set and the increment to R^2 for the current set.

The program displays:

- The F-value required for significance
- f^2
- The non-centrality parameter
- Power

General Case (Non-Central Chi-Square)

Figure 24.7 Non-central chi-square



Alpha 0.05000000
Tails 2

Degrees of freedom 12
98
Chi-square required for significa 21.02606982
Non-Centrality parameter 25.00000000

Power 0.94173465

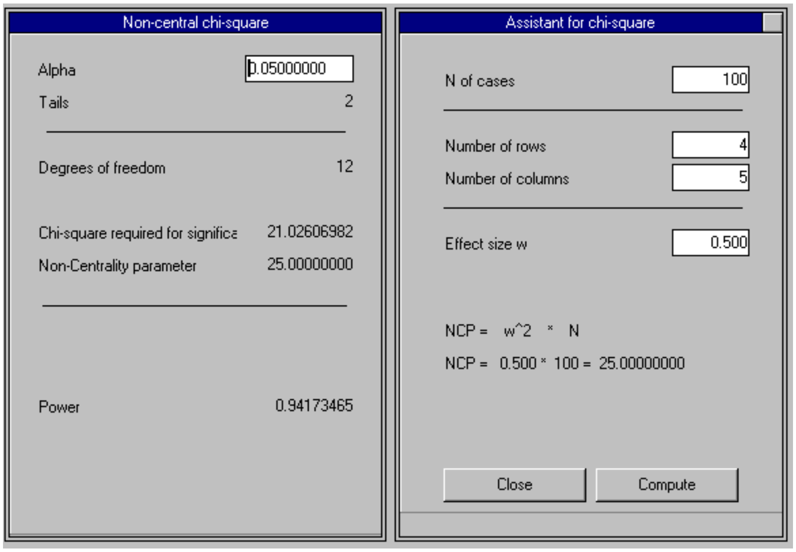
The general case for power based on non-central chi-square is shown in Figure 24.7.

The user can enter alpha, tails, degrees of freedom (df), and the non-centrality parameter (NCP). The program displays the chi-square value required for significance and power.

NCP is computed as $w^2 * N$.

The program allows the user to enter the NCP directly. The program also allows the user to invoke an assistant that accepts data in the form of a $K \times C$ crosstabulation, which it then uses to compute the NCP. The assistant shows the steps in the computation. When the assistant is active, values for df and NCP must be entered through the assistant. To enter these values directly, close the assistant.

Figure 24.8 Non-central chi-square and assistant



Non-central chi-square

Alpha 0.05000000

Tails 2

Degrees of freedom 12

Chi-square required for significa 21.02606982

Non-Centrality parameter 25.00000000

Power 0.94173465

Assistant for chi-square

N of cases 100

Number of rows 4

Number of columns 5

Effect size w 0.500

NCP = w^2 * N

NCP = 0.500 * 100 = 25.00000000

The assistant panel for non-central chi-square is shown in Figure 24.8.

- ❖ Enter alpha.
- ❖ Enter the number of cases.
- ❖ Enter the number of rows and columns in the $K \times C$ crosstabulation.
- ❖ Enter the effect size, w.

The program displays:

- The computed w^2
- The non-centrality parameter
- The chi-square value required for significance
- Power

Printing the General Case Panel

To print the panel, click the *Print* icon, or choose *Print* from the File menu. Only the panel on the left side is printed.

Copying Data to the Clipboard

To copy data to the clipboard, click the *Copy to clipboard* icon, or choose *Clipboard* from the File menu. Only the panel on the left side is copied to the clipboard.

25 Clustered Trials

Clustered trials are trials with a multiple-level design. For example, we assign hospitals at random to either of two conditions, and then all patients within each hospital are assigned to the same condition. Or, we assign schools at random to either of two conditions, and then all students within each school are assigned to the same condition.

How the use of Cluster Randomized Trials affects power

The logic of power analysis is fundamentally the same for studies that use simple random samples and for studies that use cluster random samples. For both, power is a function of the effect size, of the criterion for significance (alpha), and of the sampling distribution of the treatment effect. For purposes of power, the key difference between the two designs is in the last of these, the sampling distribution of the treatment effect. In the simple randomized trial there is only one source of error, the within-groups variance. By contrast, in the cluster randomized trial, there are two sources of error, within-clusters and between-clusters.

Consider a study where the outcome is the mean level of pain reported by patients subsequent to a surgical procedure. Standard care (Control) calls for patients to take pain medication according to one regimen, and the experimental condition (Treatment) calls for the patients to take this medication according to a new regimen.

Four hospitals will be assigned to the Control condition and four will be assigned to the Treatment condition. Power depends in part on the precision with which we can assess the infection rate in each condition (and the difference, which is the treatment effect). There will be two sources of error in our estimate.

One source of error is that the mean level of pain that we observe in a specific hospital is not the true mean in that hospital. If the true mean in Hospital A is 4 (on a 10-point scale) then we may observe a mean of 3.5 or 4.5 because of sampling error. The primary mechanism for reducing this error is to increase the sample size within hospitals.

The second source of error is that the true mean in these four hospitals is not the same as the true mean of all hospitals. The true mean varies from one hospital to the next, and so the mean in any sample of four hospitals (even if we could eliminate the first source of error) would not be the same as the mean across all possible hospitals. The primary mechanism for reducing this error is to increase the number of hospitals in the sample.

Note. The second source of error (between-studies variance) is a problem for cluster randomized trials but not for multi-center trials using simple random sampling. This is because in a simple multi-center trial every hospital includes patients assigned to both the Treated and the Control conditions. If a hospital happens to have a low risk or a high risk, this affects both conditions equally and therefore has no impact on the effect size (which is based on the difference between them). Therefore, under the usual assumptions (such as homogeneity of treatment effects across clusters) the between clusters variance has little or no impact on the error (or the power). By contrast, in a cluster randomized sample, each hospital is assigned entirely to one condition or the other. If a hospital happened to have a low mean or a high mean, this would affect one condition only, and therefore would have a direct impact on the effect size. More generally, this additional source of variance results in a larger error term and decreases the power of the test.

Implications for study planning

In planning a simple randomized trial we needed to address only one source of error, the dispersion of subjects from the population mean, and we do this by increasing the sample size. The question of allocating resources is straightforward, in the sense that there is only one option available (to increase the sample size).

In planning a cluster randomized trial, by contrast, we need to address the two sources of error. To reduce the error within hospitals we need to increase the N within hospitals. To reduce the error between hospitals we need to increase the number of clusters. Since there are now two mechanisms for reducing error (and increasing power) we need to consider how to best allocate resources between these options.

Decisions about allocation will depend largely on the extent to which the outcome varies between hospitals (clusters). If the risk of infection is consistent from one hospital to the next, then we might need only a few hospitals in our sample to obtain an accurate estimate of the infection rate. By contrast, if the risk varies substantially from one hospital to the next, then we might need to sample many hospitals to obtain the same level of accuracy.

For purposes of power it is useful to talk about the between-cluster variance as a proportion of the total variance. This proportion is called the intraclass correlation (*ICC*), denoted by ρ and defined as

$$\rho = \frac{\sigma_B^2}{\sigma_W^2 + \sigma_B^2}$$

where $\sigma^2{}_B$ is the variance between clusters, $\sigma^2{}_W$ is the variance within a cluster and ρ is the ICC. Power depends on

$$Power = 1 - F\left(c_\alpha, df, \lambda\right)$$

where *F(x,v,l)* is the cumulative distribution function of the test statistic at value *x*, with *v* degrees of freedom and non-centrality parameter *l*. In this expression, power is an increasing function of

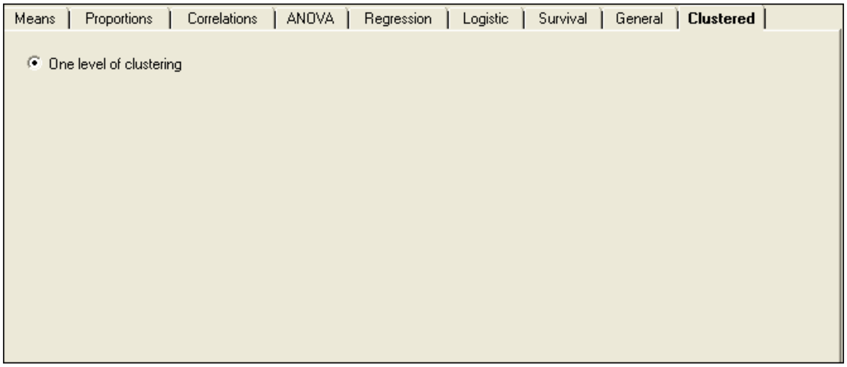
$$\lambda = \left(\sqrt{\frac{mn}{2}}\right)\left(\delta\sqrt{\frac{1}{1+(n-1)\rho}}\right),$$

where m is the number of clusters in each condition, *n* is the number of individuals in each cluster, δ is the effect size, and ρ is the *ICC*.

As outlined above, the researcher planning a cluster randomized trial must choose to allocate resources in various ways, for example by increasing the sample size within clusters, or by increasing the number of clusters. Since these mechanisms compete with each other for resources (a finite amount of time and money can be used either to increase the n within each cluster or the number of clusters), we must compare the impact of these different approaches and identify the most appropriate design for the study in question.

Selecting the Procedure

- Choose New analysis from the File menu.
- Click the Clustered tab.
- Click OK to proceed to the module.



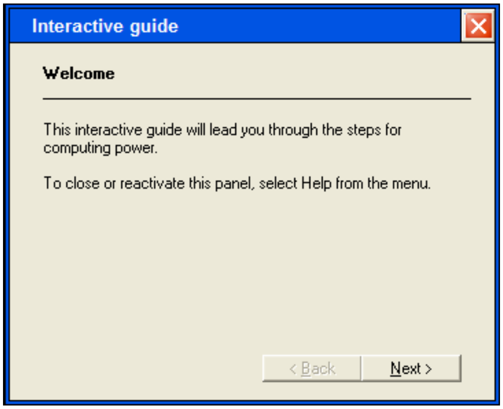
● One level of clustering

Interactive screen

Interactive guide

Click Help > Interactive guide.

The program displays a guide that will walk you through the full process of power analysis, as explained in the following text.



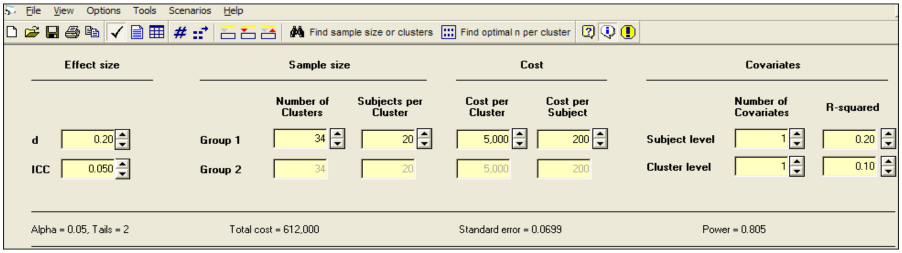
Welcome

This interactive guide will lead you through the steps for computing power.

To close or reactivate this panel, select Help from the menu.

Setting values on the interactive screen

The interactive screen is shown here



Effect size		Sample size	Number of Clusters	Subjects per Cluster	Cost per Cluster	Cost per Subject	Covariates	Number of Covariates	R-squared
d	0.20	Group 1	34	20	5,000	200	Subject level	1	0.20
ICC	0.050	Group 2	34	20	5,000	200	Cluster level	1	0.10

Alpha = 0.05, Tails = 2   Total cost = 612,000   Standard error = 0.0699   Power = 0.805

Name the groups

Initially, the program refers to the conditions as “Group 1” and “Group 2”. Click on either name (or click on Tools > Assign labels) and enter labels such as “Treated” and “Control”.

Name the clusters and subjects

Initially, the program refers to the two levels of sampling as “Cluster” and “Subject”. Click on either name (or click on Tools > Assign labels) and enter labels such as “Hospital” and “Patient” or “School” and “Student” (using the singular form rather than the plural).

Effect size

d is the standardized mean difference, defined as the raw difference between conditions divided by the standard deviation. The standard deviation in this case is computed within conditions and across clusters.

ICC is the intraclass correlation

For purposes of power it is useful to talk about the between-cluster variance as a proportion of the total variance. This proportion is called the intraclass correlation (ICC), denoted by ρ and defined as

$$\rho = \frac{\sigma_B^2}{\sigma_W^2 + \sigma_B^2}$$

The *ICC* reflects how the cluster means differ from each other within a condition. If they differ only a little (if the variance of the cluster means is only slightly more than the variance within clusters) then the *ICC* is relatively close to zero. If they differ by a lot (if the variance of the cluster means is substantially more than the variance within clusters) then the *ICC* is relatively far from zero.

The possible range of the *ICC* is 0.00 to 0.99, but in practice the *ICC* in any given field of research tends to fall within a more narrow range, and can be estimated based on prior studies. In some fields, the *ICC* might be expected to fall in the range of 0.01 to 0.05. In others, it would be expected to fall in the range of 0.10 to 0.30.

Sample size

There are two elements to the sample size – the number of clusters, and the number of subjects per cluster.

- Enter the number of clusters (for example, the number of hospitals or schools).
- Enter the number of subjects per cluster (for example, the number of patients per hospital, or the number of students per school).

Cost

The information in this section is optional. It is not needed to compute power. However, it is needed to find the optimal (most cost-effective) number of subjects per cluster (see below).

- Enter the cost of enrolling a new cluster (for example, the cost of enrolling a new hospital or school).
- Enter the cost of enrolling, treating, and following a new subject (for example, a new patient or student).

Covariates

- Subject-level. The study may have covariates at the subject level, For example, the patient's age or the student's pre-score may serve as a covariate to explain some of the outcome (and thus reduce the error term). If there are subject-level covariates, enter the number of covariates and the expected value of R^2 (the proportion of variance explained by the covariates).

- Cluster-level. The study may have covariates at the cluster level, For example, the hospital's mean success rate, or the student's mean SAT score may serve as a covariate to explain some of the outcome (and thus reduce the error term). If there are cluster-level covariates, enter the number of covariates and the expected value of R^2 (the proportion of variance explained by the covariates).

Alpha and tails

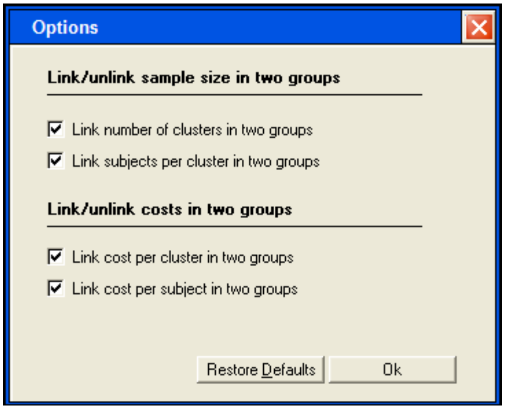
To modify these values, click on the values displayed, and the program will open a dialog box.

Linking/Unlinking the two conditions

By default, the program assumes that the number of clusters is the same in both conditions. If this is true, enter this value for the first condition, and the program will apply it to both conditions.

If the number is different for the two conditions, select Options > Link/Unlink groups and un-check "Link subjects per cluster in the two groups".

The same applies to the number of subjects within a cluster, to the cost of enrolling a new cluster, and to the cost of enrolling a new subject within a cluster. For each, the program allows you to link the values in the two groups, or to enter the value for each group independently of the other.



Options

Link/unlink sample size in two groups

☑ Link number of clusters in two groups

☑ Link subjects per cluster in two groups

Link/unlink costs in two groups

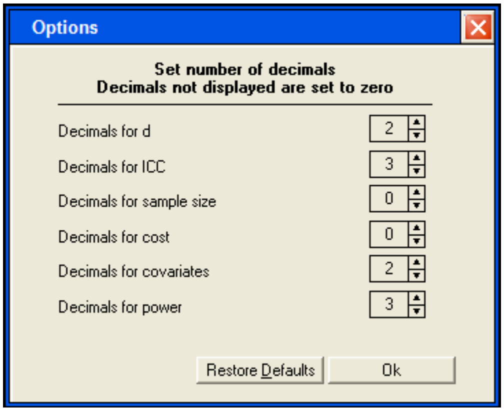
☑ Link cost per cluster in two groups

☑ Link cost per subject in two groups

Setting the number of decimals

By default the program displays two decimal places for d, three for the ICC, and so on for the other parameters. To set the number of decimals displayed –

Click Options > Decimals displayed.



Options

Set number of decimals
Decimals not displayed are set to zero

Decimals for d 2
Decimals for ICC 3
Decimals for sample size 0
Decimals for cost 0
Decimals for covariates 2
Decimals for power 3

Most values can be adjusted using a spin button. This button will always adjust the least significant decimal place. If the value displayed is 0.005, then each click will increase or decrease the value by 0.001. If the value displayed is 0.05, then each click will increase or decrease the value by 0.01.

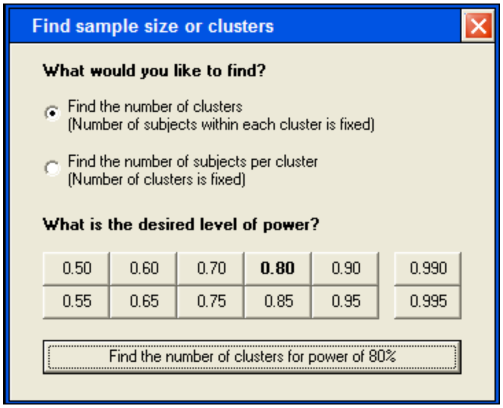
If you increase the number of decimals displayed (from 0.05 to 0.050) the value is not changed. If you decrease the number of decimals displayed (from 0.052 to 0.05) the value will be changed, and the new value will be displayed. Thus, the number displayed will always match the value that is actually being used in the computations.

Finding power and sample size

Once you have entered the effect size (*d*), the *ICC*, the number of clusters and the number of subjects, the number of covariates and R^2 for each level, alpha and tails, the program displays the power.

You can modify one (or more) of these values to see the impact on power. For example, you may modify the number of clusters until power reaches the desired level (such as 80% or 90%).

Or, the program can find the required sample size automatically as explained here.



Find sample size or clusters

What would you like to find?

● Find the number of clusters
(Number of subjects within each cluster is fixed)

○ Find the number of subjects per cluster
(Number of clusters is fixed)

What is the desired level of power?

To find the number of clusters

- Enter a value for sample size within clusters.
- Click “Find Sample size or clusters” on the toolbar.
- Select “Find the number of clusters”.
- Click a value such as 0.80.
- The program will display the number of clusters needed to yield power of 80%.

To find the number of subjects per cluster

- Enter a value for the number of clusters.
- Click “Find Sample size or clusters” on the toolbar.
- Select “Find the number of subjects per cluster”.
- Click a value such as 0.80.
- The program will display the number of subjects per cluster needed to yield power of 80%.

Note. In a standard (non-clustered) trial, as long as the effect size is not zero, power will always approach 1.0 as the sample size approaches infinity. By contrast, in a cluster randomized trial, with *ICC* greater than zero, the maximum power that can be reached is limited by the *ICC*, effect size, and covariates. For a given set of these values, it will be impossible for power to exceed some value, regardless of the number of subjects per cluster.

If (for example) the maximum power is 0.60 and you try to find the sample size that will yield power of 0.80, the program will issue a warning and explain that you need to increase the number of clusters (or other parameters).

Optimal design

For any set of parameters (the ICC, costs per cluster, cost per subject, cluster-level covariates, and subject-level covariates) there is a specific number of subjects per cluster that will yield the most cost-effective design.

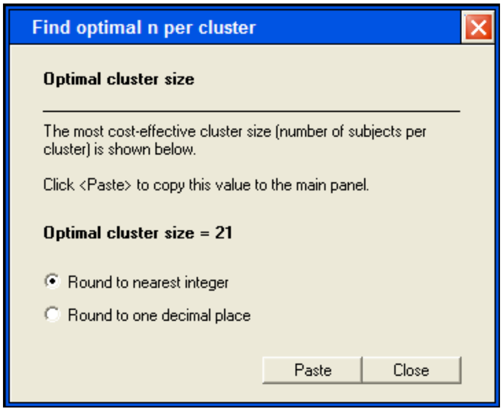
- When the *ICC* is high the balance will shift toward a low number per cluster (since power is dominated by the number of clusters rather than the number of subjects). When the *ICC* is low, the balance will shift toward a high *n* per cluster (since power is controlled by both).
- At the same time, the balance is affected by the relative costs of adding additional clusters vs. adding additional subjects within a cluster. If the ratio is large (say, $10,000 vs. $100) it will be cost effective to add subjects. If it is small (say $200 vs. $100) it will be cost effective to add clusters.
- The covariates also have an impact on the optimal sample size. Subject-level covariates serve to reduce the within-cluster error, and therefore reduce the need for a large *n* within clusters. As such, they shift the balance toward a lower n per cluster. By contrast, cluster-level covariates serve to reduce the between-cluster error, and therefore reduce the need for a large number of clusters. As such, they shift the balance toward a higher *n* per cluster.

Of course, these factors interact with each other. We need to balance the cost of a new cluster and its impact on power against the cost of a new subject *and its impact on power.*

Importantly, the relationship between subjects-per-cluster and cost is not monotonic. For example, suppose we start with a sample size of one per cluster, find the number of clusters needed for power of 90%, and compute the cost. Then we move to a sample size of two per cluster, find the number of clusters needed for power of 90%, and compute the cost, and so on. The cost will initially be high and will decline as the sample size is increased. At some point, however, the cost will begin to increase. The number of subjects per cluster when the cost curve reaches its lowest point is the optimal sample size.

The program can find the optimal sample size automatically, as follows.

- On the toolbar click “Find optimal *N* per cluster”.
- The program will open this dialog box and show the optimal *n*.



Optimal cluster size

The most cost-effective cluster size (number of subjects per cluster) is shown below.

Click <Paste> to copy this value to the main panel.

Optimal cluster size = 21

● Round to nearest integer

○ Round to one decimal place

- Select one of the options (Round to nearest integer or Round to one decimal place) and click Paste.
- The program will paste this value into the main screen, and show the corresponding power.

Tip. Now that you have the optimal n per cluster, click "Find sample size or clusters" and select the first option (Find the number of clusters).

- The program finds the number of clusters needed for the desired power.
- This design is the most cost-effective design that will yield the required power, given the values of the other parameters.

Understanding how the parameters affect power

The program computes power for a set of values provided by the user. Since these values reflect a set of decisions and assumptions, it is helpful to understand how each affects power. In particular (as explained below) it is often useful to see how power would be affected if some of these values were modified.

The effects size, *d*

The standardized mean difference (*d*) is the effect size. As *d* increases in absolute value (the further it gets from zero), the power increases.

The *ICC*, number of clusters, and *n* within clusters

In a non-clustered design the standard error is determined largely by the sample size, *n*. In a clustered design the standard error is determined by the *ICC*, the number of clusters, the n within clusters and the interaction among them. If the *ICC* is low, increasing either the number of clusters or the n within clusters will increase power. If the *ICC* is high, increasing the number of clusters will increase power, but increasing sample size within clusters will have a more limited effect.

The *ICC*

A higher value for the *ICC* will tend to reduce power. Specifically, it will tend to increase the importance of the number of clusters and diminish the importance of the sample size within clusters.

Number of clusters

Increasing the number of clusters will always increase the power. The number of clusters sets an upper limit on the potential power, and increasing the *n* per cluster will not increase power beyond this point.

Number of subjects within a cluster

When the *ICC* is low, the n per cluster can have a large impact on power. When the *ICC* is relatively high, the *n* per cluster has a relatively modest impact on power.

Subject-level covariates

The subject-level covariates reduce the within-cluster error. Increasing this R^2 will tend to increase power. The impact is similar to the impact of increasing the n within clusters. For example, the impact of this factor will be more important when the *ICC* is relatively low.

Cluster-level covariates

The cluster-level covariates reduce the between-cluster error. Increasing this R^2 will tend to increase power. The impact is similar to the impact of increasing the number of clusters. For example, the impact of this factor will be more important when the *ICC* is relatively high.

Alpha and tails

The role of alpha and tails in power for a clustered trial is the same as the role these play in a non-clustered trial. To wit –

As alpha is moved from 0.05 to 0.10, power will increase. As alpha is moved from 0.05 to 0.01 or 0.001, power will decrease.

The decision to use a one-tailed vs. a two-tailed test should be based on the nature of the research question. In the overwhelming majority of cases the two-tailed test is appropriate. This is true because even if we expect the effect to take a particular direction (we usually expect the treatment to improve the outcome) we would still interpret an effect that was statistically significant in the other direction.

That said, in the rare case where a one-tailed test is appropriate, it will yield higher power than a two-tailed test (provided, of course,that the true effect is in the expected direction).

Precision

In addition to displaying power, the program also displays the standard error of the estimate. In some cases, rather than focus only on the power of the test, we want to know how precisely we will be able to estimate the difference in means between the two treatments.

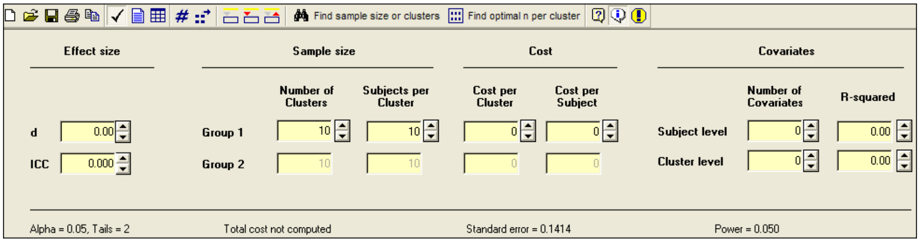
We can compute a confidence interval for the estimate by using the standard error. The 95% confidence interval will be given by the observed mean plus/minus t times the standard error, where *t* is based on the *t*-distribution with df equal to the number of clusters minus 2. With a large enough number of clusters, *t* will approach 1.96. Note that the standard error is the expected value for the standard error. In half the samples it will be smaller than the expected value, and in half it will be larger.

Cost

The program automatically shows the cost for the planned study (provided the user has entered costs for each cluster and each subject). This can be helpful when using the interactive screen to consider alternate versions of the study plans.

Example 1 – Patients within hospitals

Suppose a researcher is planning a study to test the impact of an intervention on the pain reported by patients following surgery for hernia repair. Patients in some hospitals (Control) will be given the standard set of instructions while patients in other hospitals (Treated) will be given a new set of instructions. Patients will be asked to record their pain (on a 10-point scale) for two weeks following surgery, and the mean pain reported by each patient will serve as that patient's score.pful when using the interactive screen to consider alternate versions of the study plans.

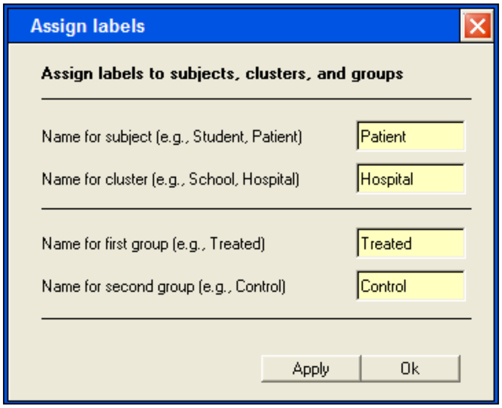


Effect size		Sample size			Cost		Covariates		
			Number of Clusters	Subjects per Cluster	Cost per Cluster	Cost per Subject		Number of Covariates	R-squared
d	0.00	Group 1	10	10	0	0	Subject level	0	0.00
ICC	0.000	Group 2	10	10	0	0	Cluster level	0	0.00

Alpha = 0.05, Tails = 2 Total cost not computed Standard error = 0.1414 Power = 0.050

Name the groups

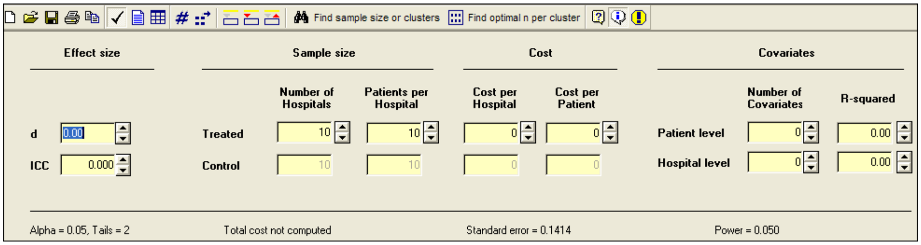
Click Tools > Assign labels and enter labels as follows.


Assign labels

Assign labels to subjects, clusters, and groups

Name for subject (e.g., Student, Patient)	Patient
Name for cluster (e.g., School, Hospital)	Hospital
Name for first group (e.g., Treated)	Treated
Name for second group (e.g., Control)	Control

Click Ok and these labels are applied to the main screen.



Effect size			Sample size			Cost		Covariates		
			Number of Hospitals	Patients per Hospital	Cost per Hospital	Cost per Patient			Number of Covariates	R-squared
d	0.00	Treated	10	10	0	0		Patient level	0	0.00
ICC	0.000	Control	10	10	0	0		Hospital level	0	0.00

Alpha = 0.05, Tails = 2 Total cost not computed Standard error = 0.1414 Power = 0.050

The effect size, *d*

A pilot study showed that the mean pain level is 6.0, with a standard deviation of 3.0.

We decide that a clinically important effect would be to reduce the mean to 4.0. This yields an effect size (*d*) of 6 minus 4 (that is, 2.0) divided by 3, or 0.67.

ICC

The ICC is expected to fall near 0.10.

Number of hospitals and patients

As a starting point, we set the number of hospitals at 10 and the number of patients per hospital at 10.

Costs

The cost of enrolling each hospital is set at $1,000, and the cost of enrolling (and following) each patient is set at $50.

Covariates

Hospital-level covariates

Each hospital has a protocol for preparing the patients to deal with recovery from surgery. We expect that more time will be helpful in itself, and will also serve as an indicator of the general care level provided. The amount of time (in minutes) spent with each patient will serve as a hospital-level covariate.

- For number of covariates enter 1.

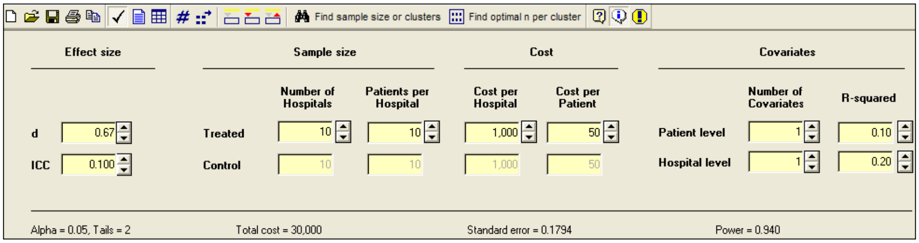
- For R^2 enter 0.20.

Patient-level covariates

Experience has shown that older patients tend to report more pain following this procedure. Therefore, we plan to use each patient's age as a covariate, and we expect that this will explain some 10% of the variance in pain scores.

- For number of covariates enter 1.
- For R^2 enter 0.10.

At this point the main screen looks like this

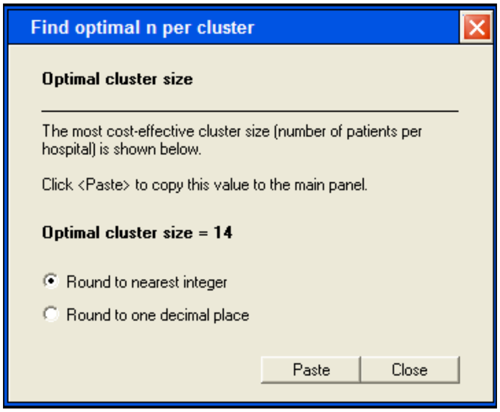


Effect size		Sample size			Cost		Covariates		
			Number of Hospitals	Patients per Hospital	Cost per Hospital	Cost per Patient		Number of Covariates	R-squared
d	0.67	Treated	10	10	1,000	50	Patient level	1	0.10
ICC	0.100	Control	10	10	1,000	50	Hospital level	1	0.20

Alpha = 0.05, Tails = 2 Total cost = 30,000 Standard error = 0.1794 Power = 0.940

Find the optimal number of patients per hospital

- Click 'Find optimal sample size'.



Find optimal n per cluster

Optimal cluster size

The most cost-effective cluster size (number of patients per hospital) is shown below.

Click <Paste> to copy this value to the main panel.

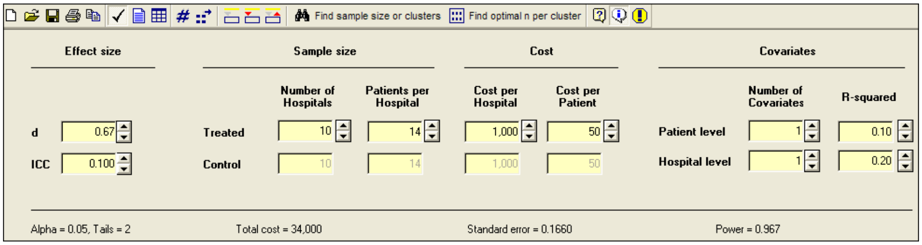
Optimal cluster size = 14

● Round to nearest integer

○ Round to one decimal place

- The program shows that the optimal n is 14.
- Click Paste to copy this number into the main screen.

The main screen now looks like this.



Effect size		Sample size	Number of Hospitals	Patients per Hospital	Cost per Hospital	Cost per Patient	Covariates	Number of Covariates	R-squared
d	0.67	Treated	10	14	1,000	50	Patient level	1	0.10
ICC	0.100	Control	10	14	1,000	50	Hospital level	1	0.20

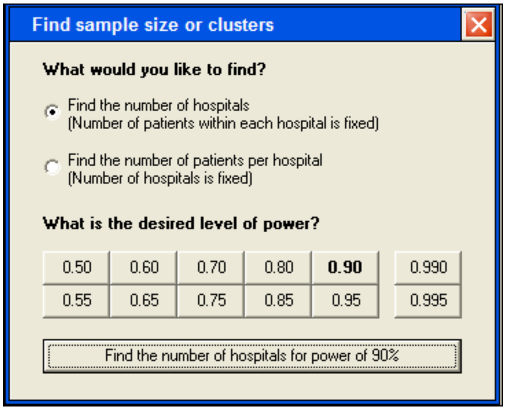
Alpha = 0.05, Tails = 2 Total cost = 34,000 Standard error = 0.1660 Power = 0.967

The number of clusters is still 10 (the arbitrary value we had set initially), the number of patients per hospital is 14 (the most cost-effective number). Power is shown as 0.967.

Find the number of clusters

Keeping the number of patients per hospital at 14, we need to find the number of hospitals that will yield power of 90%.

- Click “Find number of clusters”.
- Select the first option (“Find the number of hospitals”).
- Click 0.90.


Find sample size or clusters

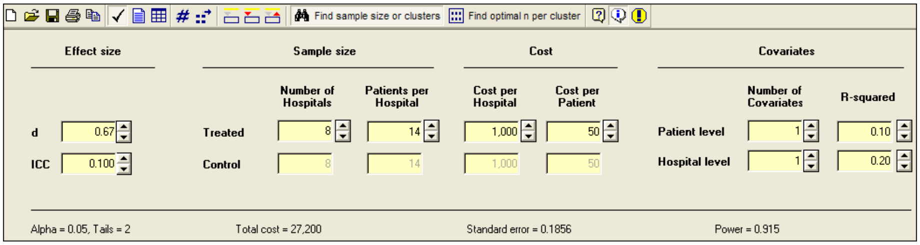
What would you like to find?

● Find the number of hospitals
(Number of patients within each hospital is fixed)

○ Find the number of patients per hospital
(Number of hospitals is fixed)

What is the desired level of power?

- The program sets the number of clusters at 8.
- Power is shown as 91.5%.
- The study cost is shown as 27,200.
- The standard error is shown as 0.186.



Effect size		Sample size			Cost		Covariates		
			Number of Hospitals	Patients per Hospital	Cost per Hospital	Cost per Patient		Number of Covariates	R-squared
d	0.67	Treated	8	14	1,000	50	Patient level	1	0.10
ICC	0.100	Control	8	14	1,000	50	Hospital level	1	0.20

Alpha = 0.05, Tails = 2 Total cost = 27,200 Standard error = 0.1856 Power = 0.915

Generate a report

To generate a report, click Report on the toolbar. The program generates the report shown here, which can be copied to Word™ or any Windows™ program.

Power for a test of the null hypothesis

One goal of the proposed study is to test the null hypothesis that the population means in the two groups (treated and control) are identical, or (equivalently) that the true effect size (d) is zero.

Study design. This hypothesis will be tested in a study that enrolls patients within hospitals.

Effect size. Power is computed for an effect size (d) of 0.67. The computations assume an intraclass correlation (ICC) of 0.100.

The standard deviation is assumed to be the same in the two groups.

Sample size. For each group we will enroll 8 hospitals with 14 patients per hospital for a total of 112 patients per group.

Patient-level covariates. There are 1 patient-level covariates. The R-squared between these covariates and outcome is assumed to be 0.10

Hospital-level covariates. There are 1 hospital-level covariates. The R-squared between these covariates and outcome is assumed to be 0.20

Alpha and Tails. The criterion for significance (alpha) has been set at 0.05. The test is 2-tailed, which means that an effect in either direction will be interpreted.

Power. Given these assumptions (for the effect size, ICC, and covariates), criteria (for alpha and tails), and plans (for the number of clusters and sample size within cluster), the study will have power of 91.5% to yield a statistically significant result.

Precison for estimating the effect size (d)

Precision. Given these same assumptions (for the ICC and covariates), and plans (for the number of clusters and sample size within cluster), the study will allow us to report the effect size (d)with a standard error of approximately 0.19.

Note that this is an expected (average) value for the standard error. The actual value in any given study will be lower or higher than this.

Disclaimer

This report is intended to help researchers use the program, and not to take the place of consultation with an expert statistician.

Cost

The projected costs for the treated group are as follows.

8 hospitals at 1,000 = 8,000.
14 patients per hospital (112 patients total) at 50 = 5,600.
Total cost for the treated group = 13,600.

The projected costs for the control group are as follows.

8 hospitals at 1,000 = 8,000.
14 patients per hospital (112 patients total) at 50 = 5,600.
Total cost for the control group = 13,600.

Total cost = 27,200.

Consider alternate assumptions

The power analysis is based on a series of decisions and assumptions. For example, we decide to “power” the study for an effect size of 0.67, to set alpha (two-tailed) at 0.05, and to require power of 90%. We assume that the ICC is 0.10 and that the proportion of variance explained by the hospital-level and patient-level covariates are 10% and 20%, respectively.

It is important to consider how power would be affected if some of these assumptions or decisions were changed. Or (from another perspective) it would be important to see what number of clusters would be needed to maintain power at 90% even if some of the assumptions or decisions were changed.

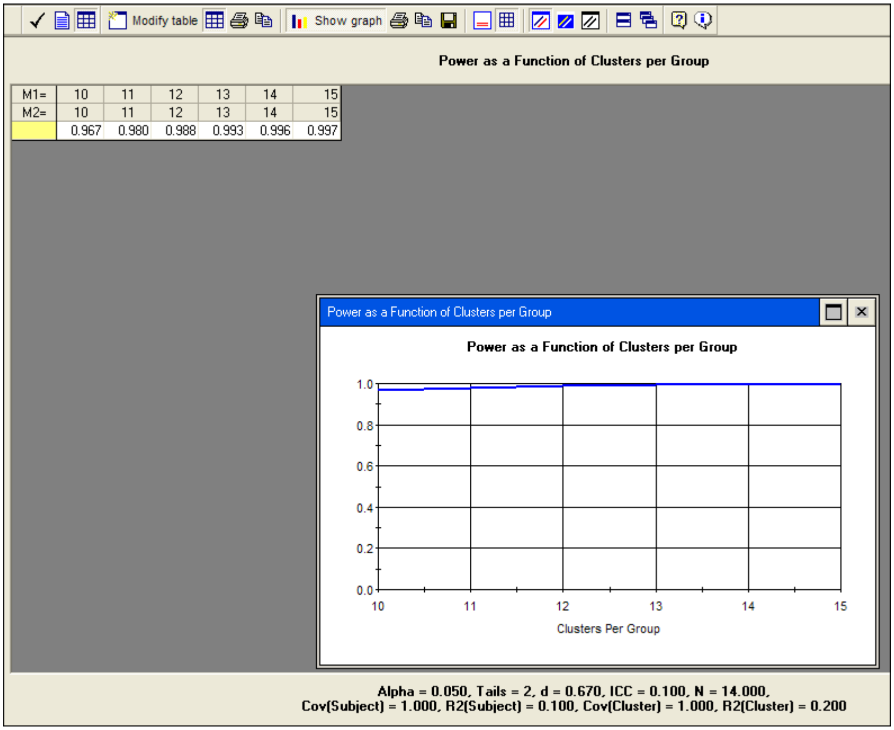
It is possible to do this working with the interactive screen. For example, if you change the *ICC* from 0.10 to 0.15, power moves from 91.5 to 84.2. Then, click “Find sample size” and the program shows that the number of clusters needed to maintain power of 90% increases from 8 to 10. The cost increases from 27,200 to 34,000.

Create a table

The program also allows you to look at these issues systematically by using tables and graphs. First, enter all the values for effect size, ICC, and so on, as above.

- Reset the *ICC* to 0.10.
- Then, click Tables on the toolbar.

The program immediately creates a table as shown here.



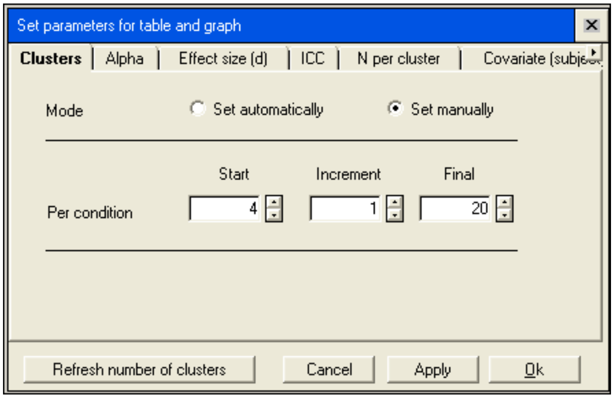
Power as a Function of Clusters per Group

M1=	10	11	12	13	14	15
M2=	10	11	12	13	14	15
	0.967	0.980	0.988	0.993	0.996	0.997

Power as a Function of Clusters per Group
Power as a Function of Clusters per Group
1.0
0.8
0.6
0.4
0.2
0.0
10
11
12
13
14
15
Clusters Per Group

Alpha = 0.050, Tails = 2, d = 0.670, ICC = 0.100, N = 14.000,
Cov(Subject) = 1.000, R2(Subject) = 0.100, Cov(Cluster) = 1.000, R2(Cluster) = 0.200

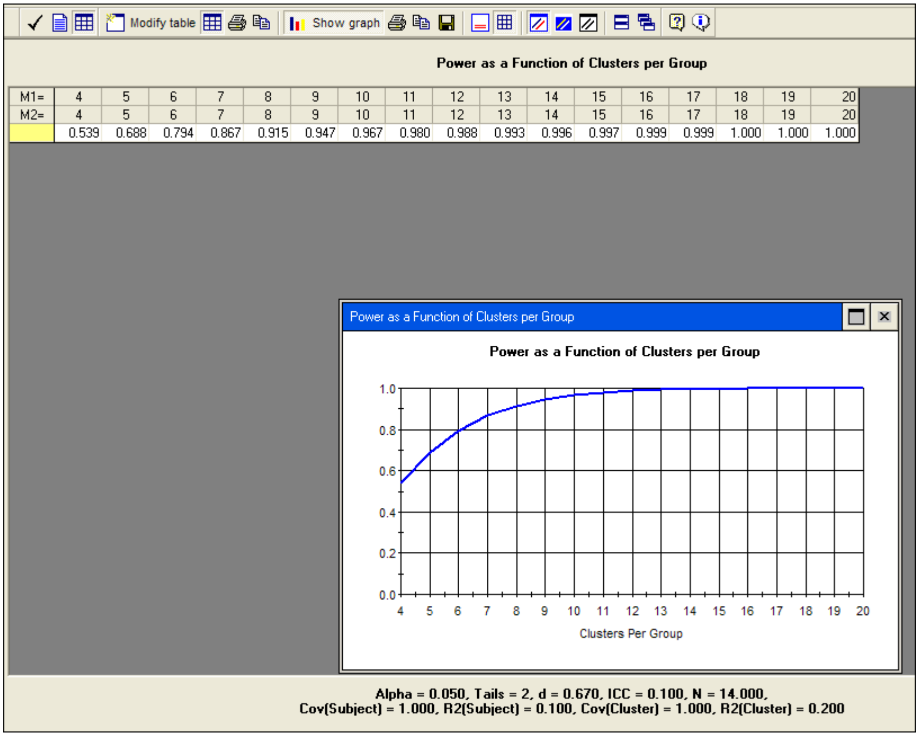
All parameters (the effect size, ICC, patient-level and hospital-level covariates, alpha, and tails) are taken from the interactive screen and displayed at the bottom of the table. The number of patients per hospital is taken from the interactive screen. The number of clusters varies from 10 to 15.



Mode ○ Set automatically ● Set manually

	Start	Increment	Final
Per condition	4	1	20

- Click Modify table.
- Select the tab for Clusters.
- Set the number of clusters to range from 4 to 20.
- Click Ok.



Power as a Function of Clusters per Group

M1=	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
M2=	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	0.539	0.688	0.794	0.867	0.915	0.947	0.967	0.980	0.988	0.993	0.996	0.997	0.999	0.999	1.000	1.000	1.000

Alpha = 0.050, Tails = 2, d = 0.670, ICC = 0.100, N = 14.000,
Cov(Subject) = 1.000, R2(Subject) = 0.100, Cov(Cluster) = 1.000, R2(Cluster) = 0.200

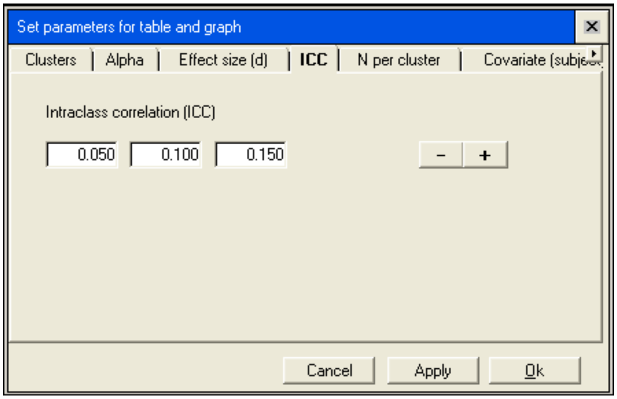
On the main screen, we had seen that we needed 8 hospitals to yield power of approximately 90%. Here, we see that we would need 6 hospitals to yield power of 80%, 8 hospitals to yield power of 90% (as before) and 9 hospitals to yield power of 95%. This provides a general sense of what our options would be if we wanted to think about lower or higher values of power.

These computations all assume that the *ICC* is 0.10. What would happen if the *ICC* was actually somewhat lower or higher than this? The program can vary the *ICC* systematically and show the results.

Click Modify table

- Select the tab for *ICC*.
- The value is shown as 0.10, which was taken from the interactive screen.
- Click "+" two times, to add two more values for the *ICC*. Enter values of 0.05, 0.10, and 0.15.

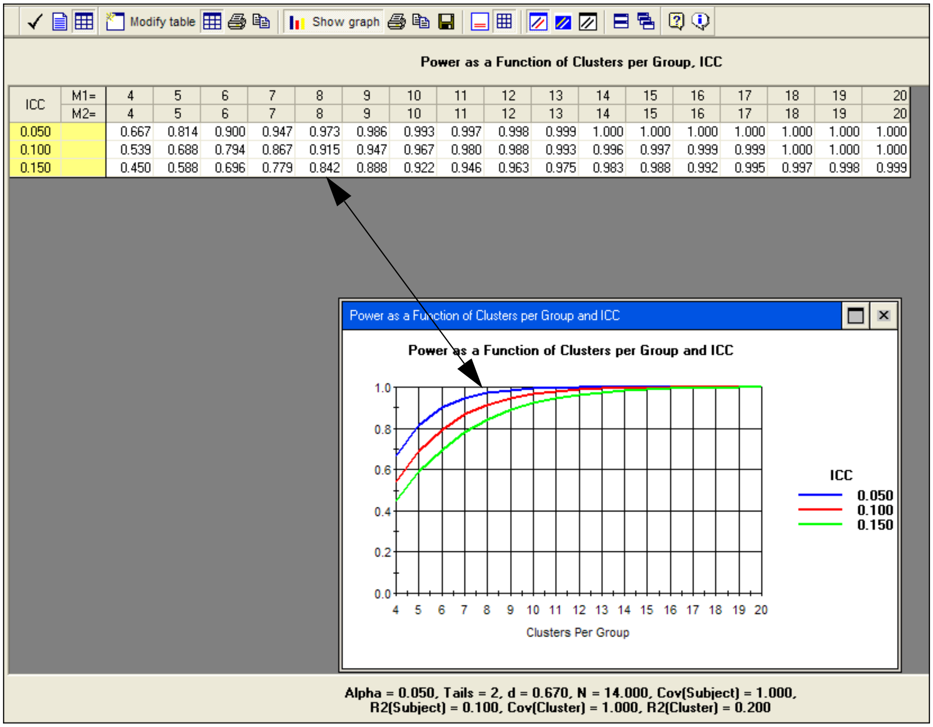
- Click OK.



Intraclass correlation (ICC)

0.050 0.100 0.150

Now, the graph shows three lines, one for each value of the *ICC*.



Power as a Function of Clusters per Group, ICC

ICC	M1=	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
	M2=	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.050		0.667	0.814	0.900	0.947	0.973	0.986	0.993	0.997	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.100		0.539	0.688	0.794	0.867	0.915	0.947	0.967	0.980	0.988	0.993	0.996	0.997	0.999	0.999	1.000	1.000	1.000
0.150		0.450	0.588	0.696	0.779	0.842	0.888	0.922	0.946	0.963	0.975	0.983	0.988	0.992	0.995	0.997	0.998	0.999

Alpha = 0.050, Tails = 2, d = 0.670, N = 14.000, Cov(Subject) = 1.000,
R2(Subject) = 0.100, Cov(Cluster) = 1.000, R2(Cluster) = 0.200

This table offers an overview of our options.

We can "power" the study based on the original *ICC* of 0.10, and set the number of hospitals at 8.

Then, assuming all the other parameters are correct –

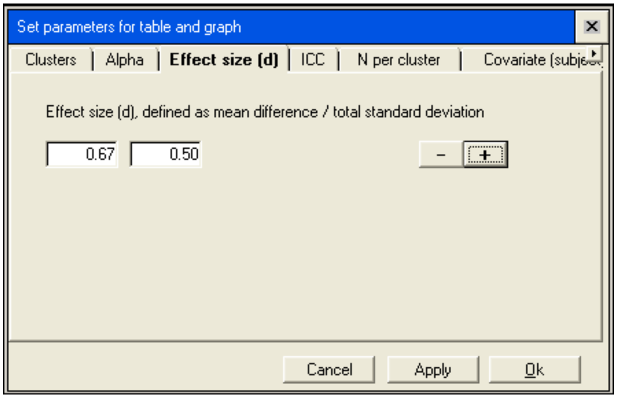
- If the *ICC* actually is 0.05, power will be 97%.
- If the *ICC* actually is 0.10, power will be 90% (as before).
- If the *ICC* actually is 0.15, power will be 84%.

Or, we may want to power the study based on the ICC of 0.15 (that is, the worst case among the values being considered). We would set the number of hospitals at 10, to yield power of 90% even for this ICC. Then –

- If the ICC actually is 0.05, power will be 99%
- If the ICC actually is 0.10, power will be 97%.
- If the ICC actually is 0.15, power will be 92%.

The program also allows us to take account of several factors simultaneously. For example, we might want to use these three values of the *ICC*, and also two values for the effect size,

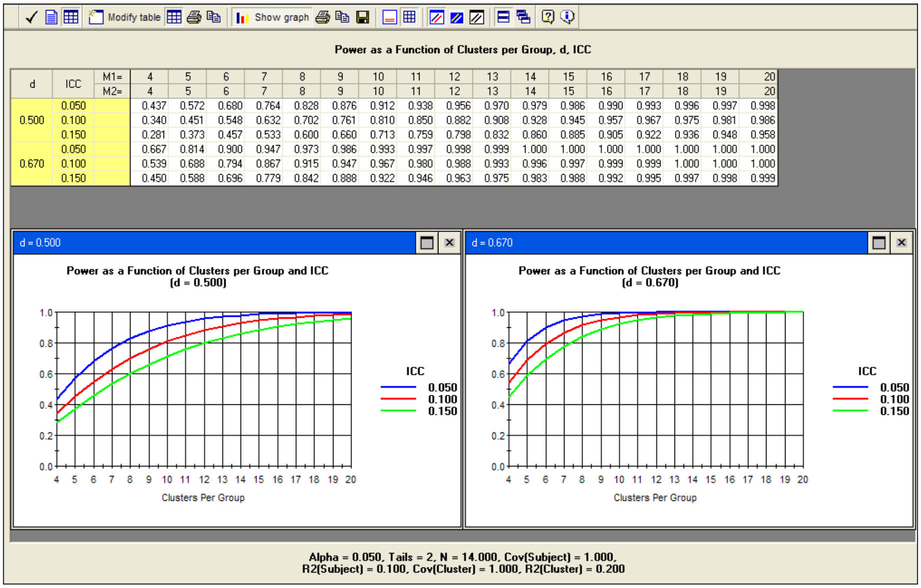
- Click Modify table.
- Select the tab for effect size.
- The value is shown as 0.67, which was taken from the interactive screen.
- Click "+" one time, to add one more value for d. Enter values of 0.67 and 0.50.
- Click OK.



Effect size (d), defined as mean difference / total standard deviation

0.67 0.50

The screen now looks like this (after clicking on Graph/Tile graphs).



Power as a Function of Clusters per Group, d, ICC

d	ICC	M1=	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
		M2=	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.500	0.050		0.437	0.572	0.680	0.764	0.828	0.876	0.912	0.938	0.956	0.970	0.979	0.986	0.990	0.993	0.996	0.997	0.998
	0.100		0.340	0.451	0.548	0.632	0.702	0.761	0.810	0.850	0.882	0.908	0.928	0.945	0.957	0.967	0.975	0.981	0.986
	0.150		0.281	0.373	0.457	0.533	0.600	0.660	0.713	0.759	0.798	0.832	0.860	0.885	0.905	0.922	0.936	0.948	0.958
0.670	0.050		0.667	0.814	0.900	0.947	0.973	0.986	0.993	0.997	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
	0.100		0.539	0.688	0.794	0.867	0.915	0.947	0.967	0.980	0.988	0.993	0.996	0.997	0.999	0.999	1.000	1.000	1.000
	0.150		0.450	0.588	0.696	0.779	0.842	0.888	0.922	0.946	0.963	0.975	0.983	0.988	0.992	0.995	0.997	0.998	0.999

Alpha = 0.050, Tails = 2, N = 14.000, Cov(Subject) = 1.000,
R2(Subject) = 0.100, Cov(Cluster) = 1.000, R2(Cluster) = 0.200

The graph at left is based on an effect size (d) of 0.50, and shows power for three values of the ICC. The graph at right is based on an effect size (d) of 0.67 (as before), and shows power for three values of the ICC.

If we want to power the study to ensure good power for an effect size of 0.50, we would use the graph at the left. To power the study for an effect size of 0.67, we would use the graph at right. In either case, we can see what happens if we want to plan for an ICC of 0.05, 0.10, or 0.15.

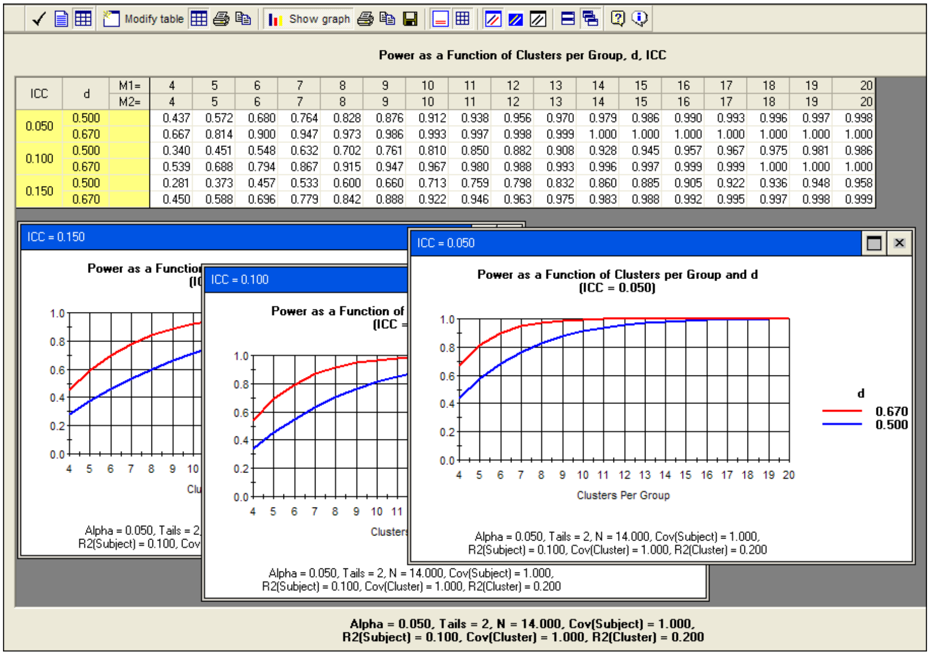
Customize the graphs

In this case each graph is based on one effect size (0.50 or 0.67), and the lines within the graph show the impact of the *ICC*. In some cases it would be helpful to have each graph reflect one *ICC*, and the lines within the graph show the impact of the effect size.

To make this change, proceed as follows.

The format of the graphs follows the sequence of columns in the table. In this table the sequence of columns is *d* followed by *ICC*, so each graph is based on one value of *d*, and the lines within a graph reflect the values of the *ICC*.

- Move one of the columns (grab the column heading that says d and move it to the right.



Power as a Function of Clusters per Group, d, ICC

ICC	d	M1=	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
		M2=	4	5	6	7	8	9	10	11	12	13	14	15	16	17	18	19	20
0.050	0.500		0.437	0.572	0.680	0.764	0.828	0.876	0.912	0.938	0.956	0.970	0.979	0.986	0.990	0.993	0.996	0.997	0.998
	0.670		0.667	0.814	0.900	0.947	0.973	0.986	0.993	0.997	0.998	0.999	1.000	1.000	1.000	1.000	1.000	1.000	1.000
0.100	0.500		0.340	0.451	0.548	0.632	0.702	0.761	0.810	0.850	0.882	0.908	0.928	0.945	0.957	0.967	0.975	0.981	0.986
	0.670		0.539	0.688	0.794	0.867	0.915	0.947	0.967	0.980	0.988	0.993	0.996	0.997	0.999	0.999	1.000	1.000	1.000
0.150	0.500		0.281	0.373	0.457	0.533	0.600	0.660	0.713	0.759	0.798	0.832	0.860	0.885	0.905	0.922	0.936	0.948	0.958
	0.670		0.450	0.588	0.696	0.779	0.842	0.888	0.922	0.946	0.963	0.975	0.983	0.988	0.992	0.995	0.997	0.998	0.999

Power as a Function of Clusters per Group and d
(ICC = 0.050)

Alpha = 0.050, Tails = 2, N = 14.000, Cov(Subject) = 1.000,
R2(Subject) = 0.100, Cov(Cluster) = 1.000, R2(Cluster) = 0.200

Alpha = 0.050, Tails = 2, N = 14.000, Cov(Subject) = 1.000,
R2(Subject) = 0.100, Cov(Cluster) = 1.000, R2(Cluster) = 0.200

- Now, the table looks like this.
- There is one graph for each *ICC*, and two lines within each graph, reflecting the two values of *d*.

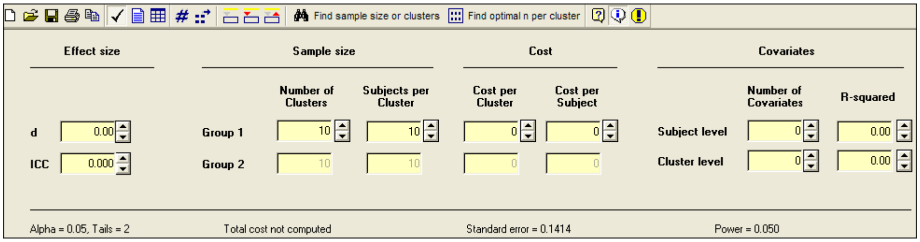
These graphs show that, for any value of the *ICC*, power drops some 10-15 points if we assume an effect size of 0.50 rather than 0.67. Put another way, to power the study for an effect size of 0.50, we would need to add about five hospitals. Using an *ICC* of 0.10 as an example, for power of 90%, with *d* = 0.67 we need 8 hospitals but with *d* = 0.50 we need 13.

Similarly, click Modify table to add any other factor(s) to the table and graphs.

Example 2 – Students within schools

Suppose a researcher is planning a study to test the impact of an intervention on the reading scores of students in the fifth grade.

Students in some schools (Control) will be given the standard curriculum while students in other schools will be given a revised curriculum (Treated). At the end of the school year, reading scores will be assessed using a standardized test.

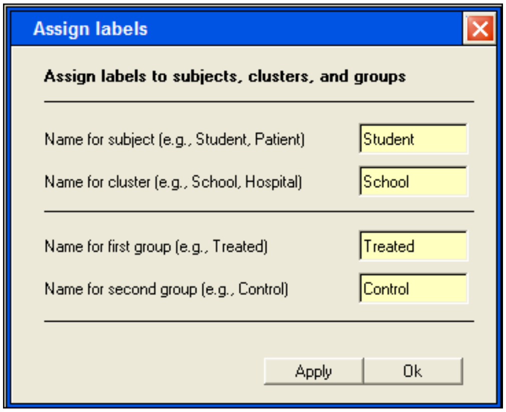


Effect size		Sample size	Number of Clusters	Subjects per Cluster	Cost per Cluster	Cost per Subject	Covariates	Number of Covariates	R-squared
d	0.00	Group 1	10	10	0	0	Subject level	0	0.00
ICC	0.000	Group 2	10	10	0	0	Cluster level	0	0.00

Alpha = 0.05, Tails = 2 Total cost not computed Standard error = 0.1414 Power = 0.050

Name the groups

Click Tools > Assign labels and enter labels as follows



Assign labels

Assign labels to subjects, clusters, and groups

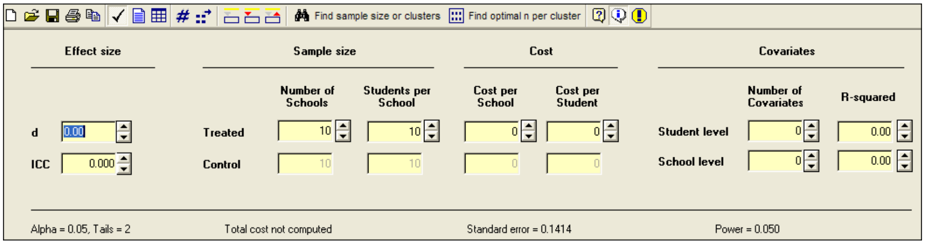
Name for subject (e.g., Student, Patient): Student

Name for cluster (e.g., School, Hospital): School

Name for first group (e.g., Treated): Treated

Name for second group (e.g., Control): Control

Click Ok and these labels are applied to the main screen.



Effect size

d 0.00

ICC 0.000

Sample size

	Number of Schools	Students per School
Treated	10	10
Control	10	10

Cost

	Cost per School	Cost per Student
Treated	0	0
Control	0	0

Covariates

	Number of Covariates	R-squared
Student level	0	0.00
School level	0	0.00

Alpha = 0.05, Tails = 2   Total cost not computed   Standard error = 0.1414   Power = 0.050

Effect size, *d*

A pilot study showed that the mean reading score is 70, with a standard deviation of 20.

We decide that a clinically important effect would be to increase the mean to 75. This yields an effect size (*d*) of 75 minus 70 (that is, 5) divided by 20, or 0.25.

ICC

The *ICC* is expected to fall near of 0.30.

Number of schools and students

As a starting point, we set the number of schools at 10 and the number of students per-school at 10.

Costs

The cost of enrolling each school is set at $2,500, and the cost of enrolling (and following) each student is set at $20.

Covariates

School-level covariates

The class takes the same standardized test every year. The mean score for the fifth grade class at the end of the prior year will serve as a school-level covariate.

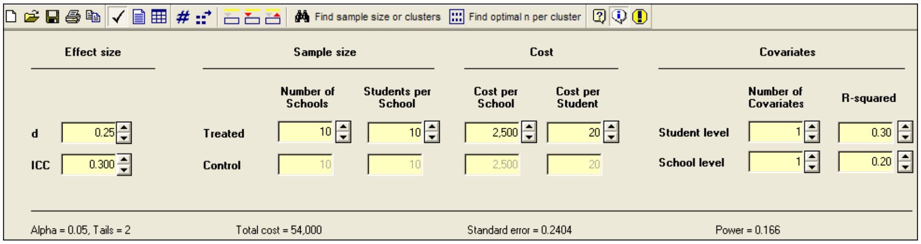
- For number of covariates enter 1.
- For R^2 enter 0.20.

Student-level covariates

For each student entering the fifth grade (and included in the study), the student's reading score from the end of the prior year (fourth grade) will serve as a student-level covariate.

- For number of covariates enter 1.
- For R^2 enter 0.30.

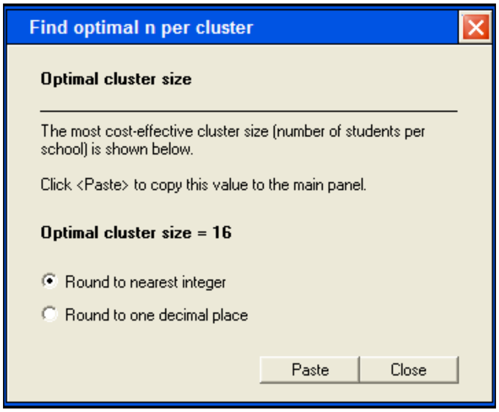
At this point the main screen looks like this



Effect size		Sample size			Cost		Covariates		
			Number of Schools	Students per School	Cost per School	Cost per Student		Number of Covariates	R-squared
d	0.25	Treated	10	10	2,500	20	Student level	1	0.30
ICC	0.300	Control	10	10	2,500	20	School level	1	0.20

Alpha = 0.05, Tails = 2 Total cost = 54,000 Standard error = 0.2404 Power = 0.166

Find the optimal number of students per school


Find optimal n per cluster

Optimal cluster size

The most cost-effective cluster size (number of students per school) is shown below.

Click <Paste> to copy this value to the main panel.

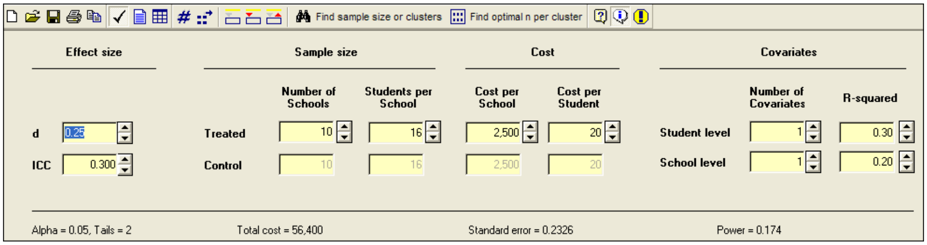
Optimal cluster size = 16

● Round to nearest integer

○ Round to one decimal place

- Click "Find optimal sample size".
- The program shows that the optimal n is 16.
- Click Paste to copy this number into the main screen.

The main screen now looks like this.



Effect size			Sample size			Cost		Covariates		
				Number of Schools	Students per School	Cost per School	Cost per Student		Number of Covariates	R-squared
d	0.25		Treated	10	16	2,500	20	Student level	1	0.30
ICC	0.300		Control	10	16	2,500	20	School level	1	0.20

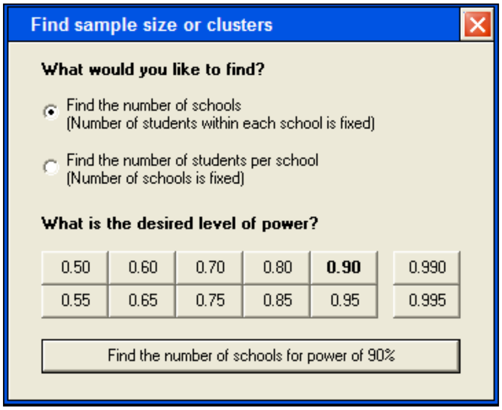
Alpha = 0.05, Tails = 2 Total cost = 56,400 Standard error = 0.2326 Power = 0.174

The number of schools is still 10 (the arbitrary value we had set initially), the number of students per school is 16 (the most cost-effective number). Power is shown as 0.174

Find the number of schools

Keeping the number of students per school at 16, we need to find the number of schools that will yield power of 90%.

- Click “Find number of schools”.
- Select the first option.
- Click 0.90.


Find sample size or clusters

What would you like to find?

● Find the number of schools
(Number of students within each school is fixed)

○ Find the number of students per school
(Number of schools is fixed)

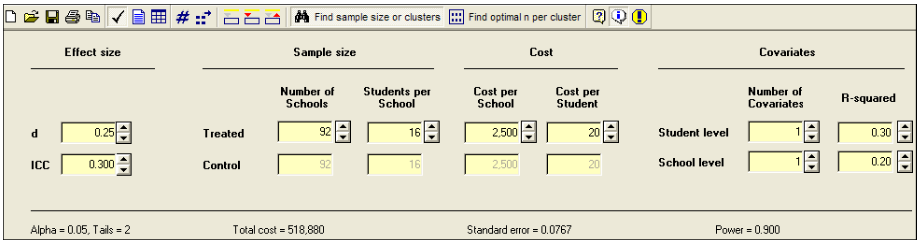
What is the desired level of power?

0.50	0.60	0.70	0.80	**0.90**	0.990
0.55	0.65	0.75	0.85	0.95	0.995

Find the number of schools for power of 90%

- The program sets the number of schools at 92.
- Power is shown as 90%.

- The study cost is shown as 518,880.
- The standard error is shown as 0.0767.



Effect size		Sample size	Number of Schools	Students per School	Number of Schools	Cost per School	Cost per Student	Covariates	Number of Covariates	R-squared
d	0.25	Treated	92	16		2,500	20	Student level	1	0.30
ICC	0.300	Control	92	16		2,500	20	School level	1	0.20

Alpha = 0.05, Tails = 2 Total cost = 518,880 Standard error = 0.0767 Power = 0.900

Generate a report

To generate a report, click Report on the toolbar. The program generates the report shown here, which can be copied to Word™ or any Windows™ program.

Power for a test of the null hypothesis

One goal of the proposed study is to test the null hypothesis that the population means

in the two groups (treated and control) are identical, or (equivalently) that the true effect size (d) is zero.

Study design. This hypothesis will be tested in a study that enrolls patients within hospitals.

Effect size. Power is computed for an effect size (d) of 0.25. The computations assume an intraclass correlation (ICC) of 0.300.

The standard deviation is assumed to be the same in the two groups.

Sample size. For each group we will enroll 92 hospitals with 16 patients per hospital for a total of 1,472 patients per group.

Patient-level covariates. There are 1 patient-level covariates. The R-squared between-these covariates and outcome is assumed to be 0.30

Hospital-level covariates. There are 1 hospital-level covariates. The R-squared between these covariates and outcome is assumed to be 0.20

Alpha and Tails. The criterion for significance (alpha) has been set at 0.05. The test is 2-tailed, which means that an effect in either direction will be interpreted.

Power. Given these assumptions (for the effect size, ICC, and covariates), criteria (for alpha and tails), and plans (for the number of clusters and sample size within cluster), the study will have power of 90.0% to yield a statistically significant result.

Precison for estimating the effect size (d)

Precision. Given these same assumptions (for the ICC and covariates), and plans (for the number of clusters and sample size within cluster), the study will allow us to report the effect size (d)with a standard error of approximately 0.08.

Note that this is an expected (average) value for the standard error. The actual value in any given study will be lower or higher than this.

Disclaimer

This report is intended to help researchers use the program, and not to take the place of consultation with an expert statistician.

Cost

The projected costs for the treated group are as follows.

92 hospitals at 2,500 = 230,000.
16 patients per hospital (1472 patients total) at 20 = 29,440.
Total cost for the treated group = 259,440.

The projected costs for the control group are as follows.

92 hospitals at 2,500 = 230,000.
16 patients per hospital (1472 patients total) at 20 = 29,440.
Total cost for the treated group = 259,440.

Total cost = 518,880.

Consider alternate assumptions

The power analysis is based on a series of decisions and assumptions. For example, we decide to "power" the study for an effect size of 0.25, to set alpha (two-tailed) at 0.05, and to require power of 90%. We assume that the ICC is 0.10 and that the proportion of variance explained by the hospital-level and patient-level covariates are 20% and 30%, respectively.

It is important to consider how power would be affected if some of these assumptions or decisions were changed. Or (from another perspective) it would be important to see what number of clusters would be needed to maintain power at 90% even if some of the assumptions or decisions were changed.

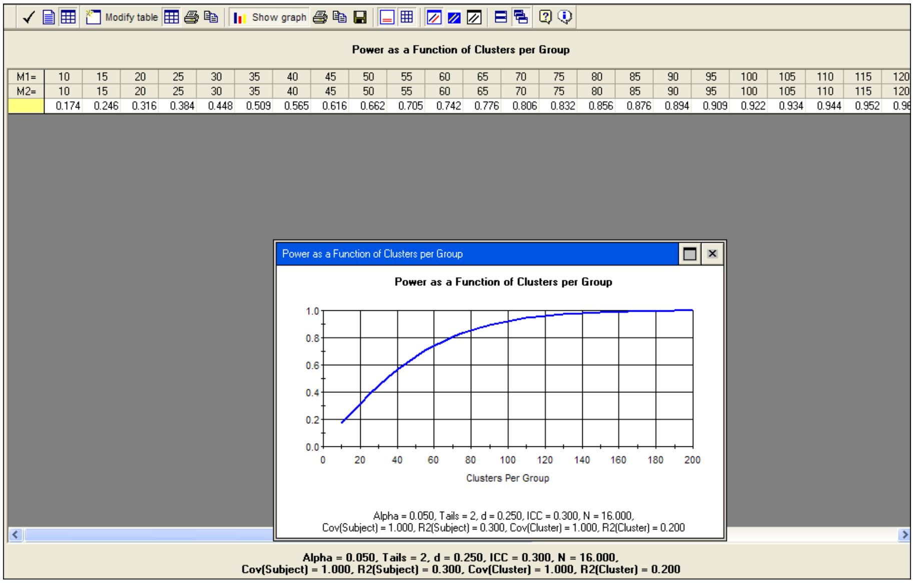
It is possible to do this working with the interactive screen. For example, if you change the *ICC* from 0.30 to 0.35, power moves from 90 to 86. Then, click “Find sample size” and the program shows that the number of clusters needed to maintain power of 90% increases from 92 to 105. The cost increases from 518,880 to 592,200.

Create a table

The program also allows you to look at these issues systematically by using tables and-graphs. First, enter all the values for effect size, ICC, and so on, as above.

- Reset the ICC to 0.30.
- Then, click “Tables” on the toolbar.

The program immediately creates a table as shown here.



Power as a Function of Clusters per Group

M1=	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100	105	110	115	120
M2=	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100	105	110	115	120
	0.174	0.246	0.316	0.384	0.448	0.509	0.565	0.616	0.662	0.705	0.742	0.776	0.806	0.832	0.856	0.876	0.894	0.909	0.922	0.934	0.944	0.952	0.96

Alpha = 0.050, Tails = 2, d = 0.250, ICC = 0.300, N = 16.000,
Cov(Subject) = 1.000, R2(Subject) = 0.300, Cov(Cluster) = 1.000, R2(Cluster) = 0.200

Alpha = 0.050, Tails = 2, d = 0.250, ICC = 0.300, N = 16.000,
Cov(Subject) = 1.000, R2(Subject) = 0.300, Cov(Cluster) = 1.000, R2(Cluster) = 0.200

All parameters (the effect size, *ICC*, student-level and school-level covariates, alpha, and tails) are taken from the interactive screen and displayed at the bottom of the table. The number of students per school is taken from the interactive screen. The number of schools varies from 10 to 200.

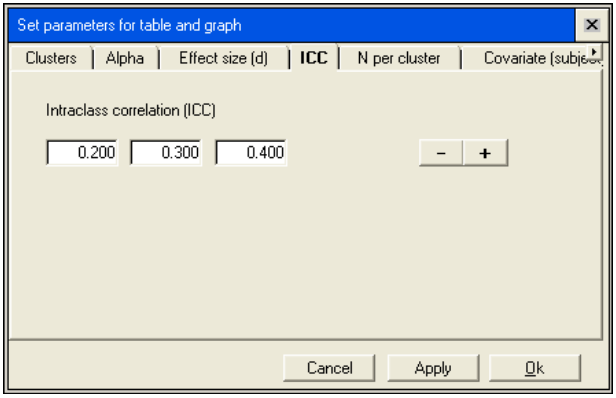
On the main screen, we had seen that we needed 92 schools to yield power of approximately 90%. Here, we see that we would need 70 schools to yield power of 80%, 92 schools to yield power of 90% (as before) and 115 schools to yield power of 95%. This provides a general sense of what our options would be if we wanted to think about lower or higher values of power.

These computations all assume that the *ICC* is 0.30. What would happen if the *ICC* was actually somewhat lower or higher than this? The program can vary the *ICC* systematically and show the results.

Click Modify table

- Select the tab for *ICC*.
- The value is shown as 0.30, which was taken from the interactive screen.

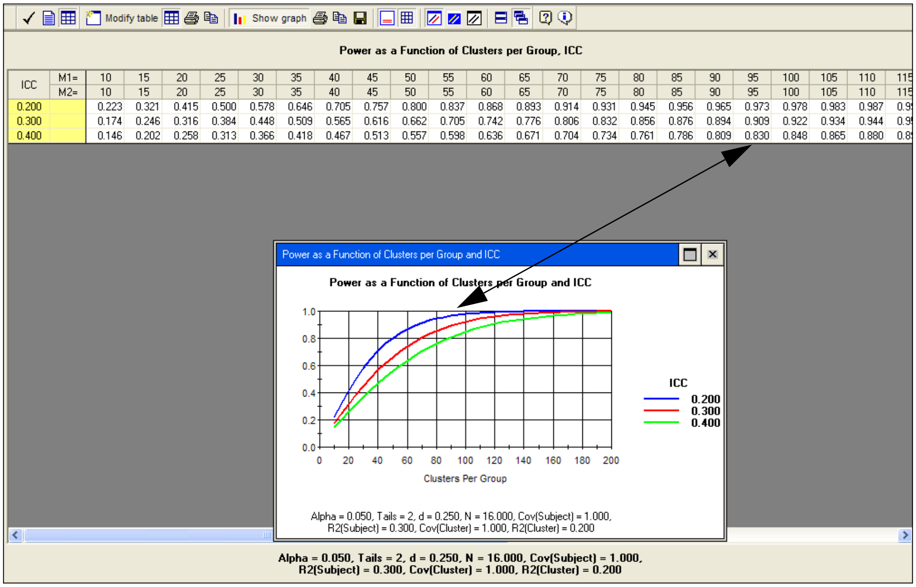
- Click "+" two times, to add two more values for the ICC. Enter values of 0.20, 0.30, and 0.40.
- Click OK.



Intraclass correlation (ICC)

0.200 0.300 0.400

Now, the graph shows three lines, one for each value of the ICC. This graph provides the following information.



Modify table
Show graph
Power as a Function of Clusters per Group and ICC
Power as a Function of Clusters per Group and ICC
1.0
0.8
0.6
0.4
0.2
0.0
0
20
40
60
80
100
120
140
160
180
200
Clusters Per Group
ICC
0.200
0.300
0.400
Alpha = 0.050, Tails = 2, d = 0.250, N = 16.000, Cov(Subject) = 1.000,
R2(Subject) = 0.300, Cov(Cluster) = 1.000, R2(Cluster) = 0.200

Power as a Function of Clusters per Group, ICC

ICC	M1=	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100	105	110	115
	M2=	10	15	20	25	30	35	40	45	50	55	60	65	70	75	80	85	90	95	100	105	110	115
0.200		0.223	0.321	0.415	0.500	0.578	0.646	0.705	0.757	0.800	0.837	0.868	0.893	0.914	0.931	0.945	0.956	0.965	0.973	0.978	0.983	0.987	0.9
0.300		0.174	0.246	0.316	0.384	0.448	0.509	0.565	0.616	0.662	0.705	0.742	0.776	0.806	0.832	0.856	0.876	0.894	0.909	0.922	0.934	0.944	0.9
0.400		0.146	0.202	0.258	0.313	0.366	0.418	0.467	0.513	0.557	0.598	0.636	0.671	0.704	0.734	0.761	0.786	0.809	0.830	0.848	0.865	0.880	0.8

Alpha = 0.050, Tails = 2, d = 0.250, N = 16.000, Cov(Subject) = 1.000,
R2(Subject) = 0.300, Cov(Cluster) = 1.000, R2(Cluster) = 0.200

This table offers an overview of our options.

We can "power" the study based on the original ICC of 0.30, and set the number of-schools at 92.

Then, assuming all the other parameters are correct –

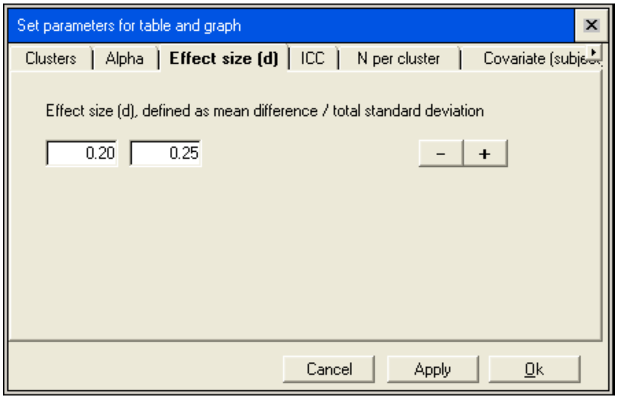
- If the *ICC* actually is 0.20, power will be 97%.
- If the *ICC* actually is 0.30, power will be 90% (as before).
- If the *ICC* actually is 0.40, power will be 81%.

Or, we may want to power the study based on the ICC of 0.40 (that is, the worst case among the values being considered). We would set the number of schools at 120, to yield power of 90% even for this ICC. Then –

- If the *ICC* actually is 0.20, power will be 99%
- If the *ICC* actually is 0.30, power will be 96%.
- If the *ICC* actually is 0.40, power will be 90%.

The program also allows us to take account of several factors simultaneously. For example, we might want to use these three values of the ICC, and also two values for the effect size,

- Click Modify table.
- Select the tab for effect size.
- The value is shown as 0.25, which was taken from the interactive screen.
- Click "+" one time, to add one more value for *d*. Enter values of 0.25 and 0.20.
- Click OK.

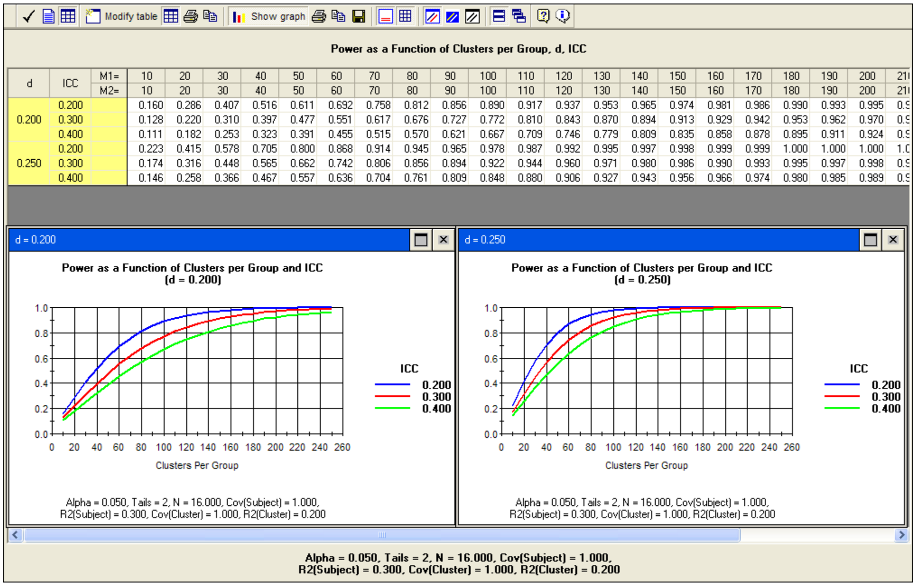


Set parameters for table and graph

Effect size (d), defined as mean difference / total standard deviation

0.20 0.25

The screen now looks like this (after arranging the position of the graphs).



Power as a Function of Clusters per Group, d, ICC

d	ICC	M1=	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200	21
		M2=	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200	21
	0.200		0.160	0.286	0.407	0.516	0.611	0.692	0.758	0.812	0.856	0.890	0.917	0.937	0.953	0.965	0.974	0.981	0.986	0.990	0.993	0.995	0.9
0.200	0.300		0.128	0.220	0.310	0.397	0.477	0.551	0.617	0.676	0.727	0.772	0.810	0.843	0.870	0.894	0.913	0.929	0.942	0.953	0.962	0.970	0.9
	0.400		0.111	0.182	0.253	0.323	0.391	0.455	0.515	0.570	0.621	0.667	0.709	0.746	0.779	0.809	0.835	0.858	0.878	0.895	0.911	0.924	0.9
	0.200		0.223	0.415	0.578	0.705	0.800	0.868	0.914	0.945	0.965	0.978	0.987	0.992	0.995	0.997	0.998	0.999	0.999	1.000	1.000	1.000	1.0
0.250	0.300		0.174	0.316	0.448	0.565	0.662	0.742	0.806	0.856	0.894	0.922	0.944	0.960	0.971	0.980	0.986	0.990	0.993	0.995	0.997	0.998	0.9
	0.400		0.146	0.258	0.366	0.467	0.557	0.636	0.704	0.761	0.809	0.848	0.880	0.906	0.927	0.943	0.956	0.966	0.974	0.980	0.985	0.989	0.9

d = 0.200

Power as a Function of Clusters per Group and ICC
(d = 0.200)

Alpha = 0.050, Tails = 2, N = 16.000, Cov(Subject) = 1.000,
R2(Subject) = 0.300, Cov(Cluster) = 1.000, R2(Cluster) = 0.200

d = 0.250

Power as a Function of Clusters per Group and ICC
(d = 0.250)

Alpha = 0.050, Tails = 2, N = 16.000, Cov(Subject) = 1.000,
R2(Subject) = 0.300, Cov(Cluster) = 1.000, R2(Cluster) = 0.200

Alpha = 0.050, Tails = 2, N = 16.000, Cov(Subject) = 1.000,
R2(Subject) = 0.300, Cov(Cluster) = 1.000, R2(Cluster) = 0.200

The graph at left is based on an effect size (*d*) of 0.20, and shows power for three values of the *ICC*. The graph at right is based on an effect size (d) of 0.25 (as before), and shows power for three values of the *ICC*.

If we want to power the study to ensure good power for an effect size of 0.20, we would use the graph at the left. To power the study for an effect size of 0.25, we would use the graph at right. In either case, we can see what happens if we want to plan for an *ICC* of 0.20, 0.30, or 0.40.

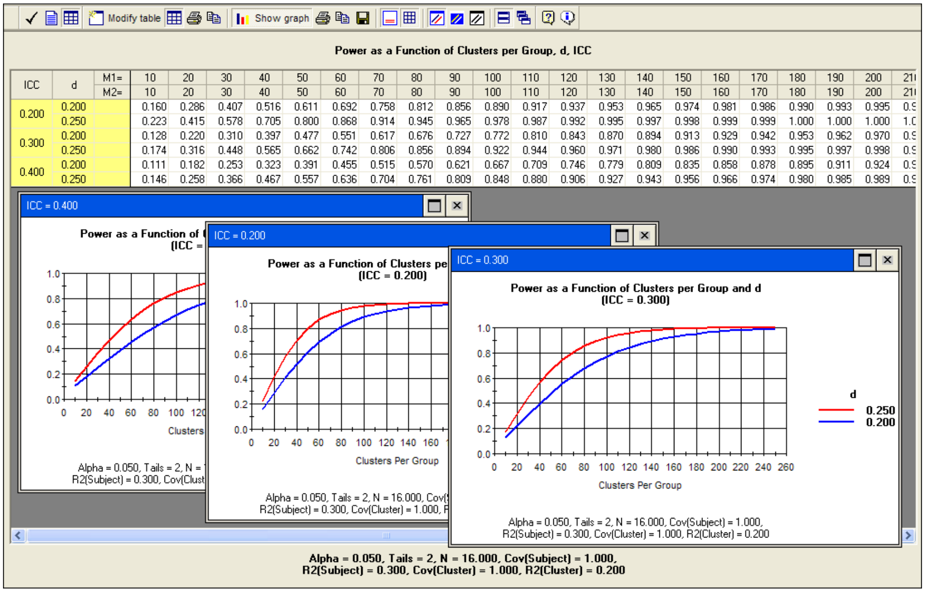
Customize the graphs

In this case each graph is based on one effect size (0.20 or 0.25), and the lines within the graph show the impact of the *ICC*. In some cases it would be helpful to have each graph reflect one *ICC*, and the lines within the graph show the impact of the effect size.

To make this change, proceed as follows.

The format of the graphs follows the sequence of columns in the table. In this table the sequence of columns is *d* followed by *ICC*, so each graph is based on one value of *d*, and the lines within a graph reflect the values of the *ICC*.

- Move one of the columns (grab the column heading that says *d* and move it to the right.



Power as a Function of Clusters per Group, d, ICC

ICC	d	M1=	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200	21
		M2=	10	20	30	40	50	60	70	80	90	100	110	120	130	140	150	160	170	180	190	200	21
0.200	0.200		0.160	0.286	0.407	0.516	0.611	0.692	0.758	0.812	0.856	0.890	0.917	0.937	0.953	0.965	0.974	0.981	0.986	0.990	0.993	0.995	0.9
	0.250		0.223	0.415	0.578	0.705	0.800	0.868	0.914	0.945	0.965	0.978	0.987	0.992	0.995	0.997	0.998	0.999	0.999	1.000	1.000	1.000	1.0
0.300	0.200		0.128	0.220	0.310	0.397	0.477	0.551	0.617	0.676	0.727	0.772	0.810	0.843	0.870	0.894	0.913	0.929	0.942	0.953	0.962	0.970	0.9
	0.250		0.174	0.316	0.448	0.565	0.662	0.742	0.806	0.856	0.894	0.922	0.944	0.960	0.971	0.980	0.986	0.990	0.993	0.995	0.997	0.998	0.9
0.400	0.200		0.111	0.182	0.253	0.323	0.391	0.455	0.515	0.570	0.621	0.667	0.709	0.746	0.779	0.809	0.835	0.858	0.878	0.895	0.911	0.924	0.9
	0.250		0.146	0.258	0.366	0.467	0.557	0.636	0.704	0.761	0.809	0.848	0.880	0.906	0.927	0.943	0.956	0.966	0.974	0.980	0.985	0.989	0.9

Alpha = 0.050, Tails = 2, N = 16.000, Cov(Subject) = 1.000, R2(Subject) = 0.300, Cov(Cluster) = 1.000, R2(Cluster) = 0.200

Alpha = 0.050, Tails = 2, N = 16.000, Cov(Subject) = 1.000,
R2(Subject) = 0.300, Cov(Cluster) = 1.000, R2(Cluster) = 0.200

- Now, the table looks like this.
- There is one graph for each ICC, and two lines within each graph, reflecting the wo values of *d*.

These graphs show that, for any value of the *ICC*, power drops some 20 points if we assume an effect size of 0.20 rather than 0.25. Put another way, to power the study for an effect size of 0.20 rather than 0.25, we would need to add about 50 schools. Using an *ICC* of 0.30 as an example, for power of 90%, with *d* = 0.25 we need 92 schools but with *d* = 0.20 we need 145.

Similarly, click Modify table to add any other factor(s) to the table and graphs.

Appendix A
Installation

Requirements

This program will run on Windows 2000, Windows XP, Windows Vista or Windows 7.

Installing

When you insert the CD, the installation should start automatically. If it does not, use the Windows Explorer to locate the file setup on the CD and double-click that filename

Uninstalling

The program folder will include an uninstall program. This program will prompt you before deleting any files that may be shared by other programs. Do not delete these files unless you are certain that they are not needed.

Troubleshooting

Before installing the program, close any other active programs, including the Microsoft Office Shortcut Bar.

Note

The following files are installed to the directory specified by the user: *PowerAndPrecision.pdf, PowerAndPrecision.chm, EasyInterfaceManual.pdf, and PowerAndPrecison V4.exe*.

The program installs and registers the following files into the Windows *system* directory (or the Windows *sys32* directory) and overwrites any earlier versions of the same files:

actbar2.ocx, ago4501.dll, c1awk.ocx, c1sizer.ocx, c4501v.dll, comdlg32.ocx, csinv32.ocx, csmete32.ocx, csspin32.ocx, cstext32.ocx, glu32.dll, houston.dll, inetwh32.dll, mfc42.dll, mmatrix.dll, msvbvm60.dll, msvcirt.dll, msvcrt.dll, olch2x32.ocx, opengl32.dll, powl6.dll, project10.dll, riched32.dll, richtx32.ocx, roboex32.dll, smthlp32.ocx, spr32x30.ocx, ssa3d30.ocx, v4501v.dll, vsflex7l.ocx, vsflex8l.ocx, vsview3.ocx, vsview6.ocx

The program installs and registers the following files into the Windows *system* directory (or the Windows *sys32* directory) but does *not* overwrite earlier versions of the same files:

comctl32.ocx, *comctl32.dll*

Additional files are written to the application directory.

All activities of the installation program are written to a log file in the program directory.

Appendix B
Troubleshooting

Following are explanations for problems that may occur.

Power will not rise above some point

When power is being computed for a two-sample study (such as a two-sample t-test) and cases are allocated unevenly to the two groups, the effective number of cases is driven primarily by the smaller of the two sample sizes. There is a maximum effective number of cases, defined as a function of the lesser number of cases, that will not be exceeded no matter how much the size of the other group is increased.

Graph lines appear jagged

Set the table to display more decimal places (3 is a good starting point).

Use a smaller increment for the number of cases (the program can set the range and increment automatically).

An increase in sample size results in a reduction in power

This is possible for exact tests of proportions because the increase in sample size may yield a more conservative value for alpha.

Appendix C
Computational Algorithms for Power

Computation of Power

The process of computing power varies from one statistical procedure to the next but always follows the same logic.

First, we determine what-value of the test statistic will be required to yield a significant effect. For example, if we are planning a t-test with $N = 20$ per group and alpha (2-tailed) set at 0.05, then a specific t-value will be required to yield a significant effect. This value is a function of the level of significance (alpha), the number of tails, and (for many tests) the sample size.

Second, we determine what proportion of studies is expected to yield a test statistic that meets this criterion. This value, which is the study's power, is a function of the effect size under the alternate hypothesis, and the sample size.

The specific computations required for these two steps are outlined here for each of the program's modules. References are to IMSL subroutines.

The operation of these algorithms can be observed directly by selecting the General panel from the procedures available (choose *New analysis* from the File menu). This allows the user to enter the non-centrality parameter (NCP) and degrees of freedom (df) directly or through "assistant" screens. In this module, the program displays the intermediate results of calculations so that the steps outlined here, and the computed value at each step, can be observed.

T-Test (One-Sample and Paired) with Estimated Variance

Required T and Power

The t-value required for significance (ReqT) is given by the central t-distribution for significance level alpha and/or tails and df = N – 1. The value is computed using the DTIN algorithm.

The effect size (d) is computed as

d = ABS(Mean Difference) / SD

The non-centrality parameter (NCP) is computed as NCP = Abs(d * Sqrt(N_{Cases})).

Power is given by the non-central t for NCP, ReqT, and df. The value is computed using the DTDF algorithm.

One-Sample Test and Paired T-Test

The same formula is applied for the one-sample t-test and the paired t-test. For the paired t-test, the program allows the user to enter the standard deviation for each time point and computes the standard deviation of the difference using the formula

SD_{Diff} = Sqrt($[SD(1)]^2 + [SD(2)]^2$ – (2 * Corr * SD(1) * SD(2)))

One-Tailed versus Two-Tailed Tests

For a one-tailed test, ReqT is found for alpha, and power is computed only for an effect in one direction (using the absolute value of d).

For a two-tailed test, ReqT is found for alpha/2. Power is computed for an effect in one direction using the absolute value of NCP, it is computed for an effect in the reverse direction using –1 * NCP, and the two values are summed.

T-Test (Two-Sample) with Estimated Variance

Required T and Power

The t-value required for significance (ReqT) is given by the central t-distribution for significance level alpha and/or tails and df = N1 + N2 – 2. The value is computed using the DTIN algorithm.

The effect size (d) is computed as d = ABS(Mean Difference) / SD_{Pooled}, where

SD_{Pooled} = Sqrt((((N1 – 1) * SD1 * SD1 + (N2 – 1) * SD2 * SD2) / (N1 + N2 – 2)))

The harmonic mean of the number of cases is given by

HarmonicN = (2 * N1 * N2) / (N1 + N2)

The non-centrality parameter is computed as

NCP = Abs((d * Sqrt(HarmonicN)) / Sqrt(2))

Power is given by the non-central t for NCP, ReqT, and df. The value is computed using the DTDF algorithm.

One-Tailed versus Two-Tailed Tests

For a one-tailed test, ReqT is found for alpha, and power is computed only for an effect in one direction (using the absolute value of d).

For a two-tailed test, ReqT is found for alpha/2. Power is computed for an effect in one direction using the absolute value of NCP, it is computed for an effect in the reverse direction using –1 * NCP, and the two values are summed.

Z-Test (One-Sample and Paired) with Known Variance

Required Z and Power

The z-value required for significance (Z_{req}) is given by the normal distribution for significance level alpha and/or tails. The value is computed using the DNORIN algorithm.

The effect size (d) is computed as

d = Abs(Mean Difference) / SD

The non-centrality parameter (NCP) is computed as

NCP = Abs(d * Sqrt(NCases))

Power is given by the normal distribution for NCP and Z_{req}. The value is computed using the DNORDF algorithm.

One-Sample Test and Paired T-Test

The same formula is applied for the one-sample t-test and the paired t-test. For the paired t-test, the program allows the user to enter the standard deviation for each time point and computes the standard deviation of the difference using the formula

SD_{Diff} = Sqrt($[SD(1)]^2$ + $[SD(2)]^2$ – (2 * Corr * SD(1) * SD(2)))

One-Tailed versus Two-Tailed Tests

For a one-tailed test, Z_{req} is found for alpha, and power is computed only for an effect in one direction (using the absolute value of d).

For a two-tailed test, Z_{req} is found for alpha/2. Power is computed for an effect in one direction using the absolute value of NCP, it is computed for an effect in the reverse direction using –1 * NCP, and the two values are summed.

In this description, we have used the term "non-central" to maintain the parallel with the t-test, although the normal distribution is symmetrical.

Z-Test (Two-Sample) with Known Variance

Required Z and Power

The z-value required for significance (Z_{req}) is given by the normal distribution for significance level alpha and/or tails. The value is computed using the DNORIN algorithm.

The effect size (d) is computed as

d = Abs(Mean Difference) / SDPooled

where

SDPooled = Sqrt((((N1 – 1) * SD1 * SD1 + (N2 – 1) * SD2 * SD2) / (N1 + N2 – 2)))

The harmonic mean of the number of cases is given by

HarmonicN = (2 * N1 * N2) / (N1 + N2)

The non-centrality parameter is computed as

NCP = Abs((d * Sqrt(HarmonicN)) / Sqrt(2))

Power is given by the normal distribution for NCP and Z_{req}. The value is computed using the DNORDF algorithm.

One-Tailed versus Two-Tailed Tests

For a one-tailed test, Z_{req} is found for alpha, and power is computed only for an effect in one direction (using the absolute value of NCP).

For a two-tailed test, Z_{req} is found for alpha/2. Power is computed for an effect in one direction using the absolute value of NCP, it is computed for an effect in the reverse direction using –1 * NCP, and the two values are summed.

Single Proportion versus a Constant

Single Proportion versus Constant: Arcsin Method

The arcsin transformation of a proportion is defined as

Arcsin = 2 * Atn(x / Sqrt(–x * x + 1))

where

x	=	Sqrt(p)
H1	=	Arcsin transformation of P1
H2	=	Arcsin transformation of P2
HDIFF	=	H1 – H2
Z_{req}	=	Z-value required for significance
N	=	N of cases
Z_{power}	=	Abs(HDIFF * Sqrt(2)) * Sqrt(N / 2) – Z_{req}
Power	=	Computed for Z_{power} based on normal distribution

One-Tailed versus Two-Tailed Tests

For a one-tailed test, Z_{req} is found for alpha, and power is computed only for an effect in one direction (using the absolute value of HDIFF).

For a two-tailed test, Z_{req} is found for alpha/2. Power is computed for an effect in one direction using the absolute value of HDIFF, it is computed for an effect in the reverse direction using –1 * Abs(HDIFF), and the two values are summed.

Single Proportion versus a Constant (Exact-test)

As noted above, in all procedures the program first finds the critical value required for significance under the null and then finds the proportion of studies (under the alternate) that will meet this criterion. In the exact-test for proportions, both steps in this process require a consideration of all possible discrete outcomes.

The program uses the binomial distribution to find the number of successes required to yield significance at the nominal value of alpha. After establishing the number of successes required to yield a significant effect, the program again uses the binomial distribution to find the proportion of cases expected to meet this criterion, given the "true" success rate under the alternate hypothesis.

Power for an effect in the other direction is computed by the analogous process—we find the first number of successes that will allow us to reject the hypothesis, and then determine what proportion of cases will meet this criterion.

One-Tailed and Two-Tailed Tests

For a one-tailed test, the criterion alpha used in the first step is the alpha specified by the user. The program computes power for an effect in either direction and reports only the higher of these two values. The program also reports the actual alpha.

For a two-tailed test, the criterion alpha used in the first step is alpha/2. The program computes power for an effect in either direction and sums these values. Similarly, actual alpha is the sum of actual alpha in either direction.

Two Independent Proportions

Two Proportions: Arcsin Method

The arcsin transformation of a proportion is defined as

Arcsin = 2 * Atn(x / Sqrt(–x * x + 1))

where

x = Sqrt(p)
H1 = Arcsin transformation of P1
H2 = Arcsin transformation of P2
HDIFF = H1 – H2
Z_{req} = Z-value required for significance
N' = (2 * (N1 * N2)) / (N1 + N2)
Z_{power} = Abs(HDIFF) * Sqrt(N' / 2) – Z_{req}
Power = Computed for Z_{power} based on abnormal distribution

One-Tailed versus Two-Tailed Tests

For a one-tailed test, Z_{req} is found for alpha, and power is computed only for an effect in one direction (using the absolute value of HDIFF).

For a two-tailed test, Z_{req} is found for alpha/2. Power is computed for an effect in one direction using the absolute value of HDIFF, it is computed for an effect in the reverse direction using –1 * Abs(HDIFF), and the two values are summed.

Two Proportions: Normal Approximations

Assume that P1 is the higher of the two proportions and P2 is the lesser of the two.

N' = (2 * N1 * N2) / (N1 + N2)
Z_{req} = Z required for significance, given alpha
$Mean_P$ = (P1 + P2) / 2
Q1 = 1 – P1
Q2 = 1 – P2
$Mean_Q$ = 1 – $Mean_P$
d' = 0.5 + Z_{req} * Sqrt(0.5 * N' * $Mean_P$ * $Mean_Q$)
Zpower = (d' – 0.5 * N' * (P1 – P2) – c) / Sqrt(0.25 * N' * (P1 * Q1 + P2 * Q2))

Power is found from the normal distribution for Zpower.

The program incorporates three versions of this approximation, which differ only in the definition of the correction factor, c:

- For the method identified simply as the normal approximation with unweighted $Mean_P$, c = 0.5.
- For the Kramer-Greenhouse method, c = – 0.5.
- For the Casagrande and Pike method, c = 0.

Computation of Mean_P under the Null

Power is always computed based on the variance expected under the null, and in a test of two proportions the variance under the null is a function of P1 and P2 (that is, the proportion of positive cases in either group).

The methods identified as the normal approximation, Kramer-Greenhouse, and Casagrande and Pike all use $(P1+P2)/2$ to compute the common population proportion (and variance) under the null.

A variant on this formula uses the weighted mean of P1 and P2 to compute the proportion under the null. This option is not generally appropriate, but the method is discussed in the literature and is offered here for special cases.

Mean_P = Abs((Cell(1, 3) * P1 + Cell(2, 3) * P2) / (Cell(1, 3) + Cell(2, 3)))

where the cell is identified by Cell(row,column). If N1 = N2, this will yield identical results to the method identified as the normal approximation with unweighted Mean_P.

One-Tailed versus Two-Tailed Tests

For a one-tailed test, Z_{req} is found for alpha, and power is computed only for an effect in one direction by assigning the higher proportion to P1 in the algorithm.

For a two-tailed test, Z_{req} is found for alpha/2. Power is computed for an effect in one direction using P1 > P2, it is computed for an effect in the reverse direction by assigning the lower proportion to P1 in the algorithm, and the two values are summed.

Two Proportions—Chi-Square Test

The user provides the sample size in each group and the proportion of "positive" cases in each group. The program uses this information to create the corresponding 2×2 table, corresponding to the alternate hypothesis.

The chi-square value required for significance is given by the DCHIIN algorithm for alpha and df = 1.

The non-centrality parameter (NCP) is defined as w^2 * N_{total}, where w is defined as in the $K \times C$ analysis. The NCP defined in this way is equivalent to the chi-square value computed for the 2×2 table, and the program works with this value, which is given by the CTTWO algorithm.

The program allows the user to select chi-square with the Yates correction. When this option is selected, the NCP is adjusted and is given the Yates-corrected chi-square value returned by the CTTWO algorithm.

Power is then given by the non-central chi-square distribution for df = 1, required chi-sq, and NCP. This value is computed by the DSNDF algorithm.

One-Tailed versus Two-Tailed Tests

The chi-square distribution is one-tailed (corresponding to a nondirectional test). When this computational option is selected, the program will compute power for a nondirectional test only.

Two Proportions: Fisher's Exact test

The user provides the number of cases in each row and the proportion of "positive" cases in each row under the alternate hypothesis.

The program iterates through every possible outcome for row 1. For example, if 10 cases will be assigned to row 1, it is possible to obtain 0, 1, 2, …, 10 successes in this row. A likelihood value is assigned to each row using the binomial distribution, taking into account the proportion of successes under the alternate.

Since the two rows are independent of each other, the probability of any given joint outcome is computed as the product of the likelihood for a given outcome in row 1 by a given outcome in row 2. For example, if the likelihood of 5 failures/5 successes in row 1 is 10% and the likelihood of 2 failures/8 successes in row 2 is 2%, the likelihood of drawing a sample with 5/5 in row 1 and 2/8 in row 2 is 0.10 * 0.02 = 0.002.

The program iterates through every possible combination of outcomes. For each combination, the program computes Fisher's exact-test and records the p-value. If the p-value is equal to or less than alpha, the likelihood of this outcome is added to a cumulative total; at the conclusion of the iterative process, this cumulative total gives the proportion of outcomes that yield significance—that is, power.

One-Tailed versus Two-Tailed Tests

The algorithm used to compute significance for any 2×2 table computes both a one-tailed and a two-tailed p-value.

When power is computed for a one-tailed test, the one-tailed p-value is used but is counted only if the effect is in the expected direction. When power is computed for a two-tailed test, the two-tailed p-value is used.

To speed up computations, the program does not compute the Fisher exact-test for 2×2 tables that are very unlikely to occur (defined as likelihood < 0.0000001). Even if this combination yielded a significant result, the increment to power would be 0.0000001 or less.

McNemar Test of Paired Proportions

The McNemar test is a special case of one proportion against a constant. Cases are assigned to the four cells:

	Negative	**Positive**
Negative	AA	AB
Positive	BA	BB

For example, we might specify that cases will fall into the four cells in the following proportions:

	Negative	**Positive**
Negative	15%	30%
Positive	20%	35%

Information about the treatment effect is provided by cells AB and BA, while cells AA and BB provide no information.

The effect size is based solely on cells AB+BA. Concretely, we work with 30% versus 20%, or 50% of the sample in this example. The proportion of cases falling into cell AB is computed based on this subsample as AB/(AB+BA). In this example, 30/50 = 60% fall into cell AB. Under the null hypothesis, this value will always be 50%.

The adjusted proportion in cell AB (60%) is tested against the constant (50%) using either of the formulas described for one proportion against a constant (the arcsin approximation or the exact-test). In either case, the number of cases is given by the N in these two cells.

Sign Test

The sign test is used to test the hypothesis that the proportion of cases falling into two groups is equal (that is, that the proportion in either group is 50%).

This test is a special case of a single sample proportion against a constant. The proportion of cases falling into cell A is tested against a constant of 50% using either of the formulas described for one proportion against a constant (the arcsin approximation or the exact-test). In either case, the number of cases is given by the full N.

K x C Crosstabulation

Computing Effect Size

The table provided by the user (which requires that the cells in a row sum to 1.0, and the proportion of cases falling into each row is given separately) is used to create a table in which the value in each cell is that cell's proportion of the total sample, so that the cells in the table (rather than the row) sum to 1.0, the row marginals sum to 1.0, and the column marginals sum to 1.0.

Under the null hypothesis, the expected value in each cell is defined as the product of that cell's marginals. For each cell, the discrepancy between the null and the alternate is computed as $((\text{PercentAlternate}-\text{PercentNull})^2/\text{PercentNull})$. The sum of these squared discrepancies yields w^2, and the square root of this value yields the effect size, w.

Computing Power

- The degrees of freedom is computed as (Rows – 1) * (Columns – 1).
- The chi-square value required for significance is given by the central chi-square distribution for alpha and df. This value is computed by the DCHIIN algorithm.
- The non-centrality parameter (NCP) is computed as $w^2 * N_{Total}$.
- Power is given by the non-central chi-square distribution for required chi-sq, NCP, and df. This value is computed by the DSNDF algorithm.

One-Tailed versus Two-Tailed Tests

The $K \times C$ test is nondirectional.

Correlations—One Sample

The test of a single correlation against a null hypothesis of 0 is computed by the exact method, which will yield the same value as the multiple regression procedure. Tests of a single correlation versus a non-zero constant and tests that two correlations are different from each other are carried out using the Fisher-Z transformation.

Pearson Correlation—One Sample versus Zero

The t-value required for significance (ReqT) is given by the central t-distribution for significance level alpha and/or tails and df = $N_{Total} - 2$. The value is computed using the DTIN algorithm.

The user provides the correlation coefficient, r, which is squared to yield r^2.
The non-centrality parameter (NCP) is computed as

NCP = t * Sqrt(NCases)

where

t = Sqrt(r^2 / (1 – r^2))

Power is given by the non-central t distribution for NCP, ReqT, and df. The value is computed using the DTDF algorithm.

One-Tailed versus Two-Tailed Tests

For a one-tailed test, ReqT is found for alpha, and power is computed only for an effect in one direction (using the absolute value of NCP).

For a two-tailed test, ReqT is found for alpha/2. Power is computed for an effect in one direction using the absolute value of NCP, it is computed for an effect in the reverse direction using –1 * NCP, and the two values are summed.

One Sample versus Constant Other than Zero

In this description, Zr is used to denote the Fisher-Z transformation of r.

- Zr_1 is the Fisher-Z transformation of the correlation under the alternate, and Zr_2 is the Fisher-Z transformation of the constant against which r_1 will be tested.
- The effect size, Q, is computed as Q = Abs($Zr_1 – Zr_2$).
- N is the number of cases.
- Z_{req} (the z-value required for significance) is Z_p, which is the z-value corresponding to power (not to be confused with Fisher's Z). Z_p is computed as

Z_p = Q * Sqrt((N – 3) / 1) – Z_{req}

Power is then computed as the area under the curve to the left of Z_p using the normal distribution.

One-Tailed versus Two-Tailed Tests

For a one-tailed test, Z_{req} is found for alpha, and power is computed only for an effect in one direction (using Q).

For a two-tailed test, Z_{req} is found for alpha/2. Power is computed for an effect in one direction using Q, it is computed for an effect in the reverse direction using –1 * Q, and the two values are summed.

Correlations—Two Sample

In this description, Zr is used to denote the Fisher-Z transformation of r. Zr_1 is the Fisher-Z transformation of the first correlation, and Zr_2 is the Fisher-Z transformation of the second correlation.

The effect size, Q, is computed as Q= Abs($Zr_1 – Zr_2$).

N' = ((2 * (N1 – 3) * (N2 – 3)) / (N1 + N2 – 6)) + 3

Z_{req} (the z-value required for significance) is Z_p, which is the z-value corresponding to power (not to be confused with Fisher's Z). Z_p is computed as

Z_p = Q * Sqrt((N' – 3) / 2) – Z_{req}

Power is then computed as the area under the curve to the left of Z_p using the normal distribution.

One-Tailed versus Two-Tailed Tests

For a one-tailed test, Z_{req} is found for alpha, and power is computed only for an effect in one direction (using Q).

For a two-tailed test, Z_{req} is found for alpha/2. Power is computed for an effect in one direction using Q, it is computed for an effect in the reverse direction using –1 * Q, and the two values are summed.

Analysis of Variance

The program will compute power for a one-way analysis of variance, or a balanced factorial ANOVA, for a fixed effects model.

Computing the Effect Size (f)

The effect size (f) is computed in one of four ways:

The program allows the user to enter f directly, in which case the value given for SD_{Within} has no impact on f.

The user can enter the $SD_{Between}$ levels for a factor, in which case the program computes

$f = SD_{Between} / SD_{Within}$

The user can enter the mean for each level in a factor, in which case the program computes the $SD_{Between}$ levels for a factor, and then computes

$f = SD_{Between} / SD_{Within}$

Finally, the user can enter the range of means (that is, the single lowest and highest means) and the pattern of dispersion for the remaining means. In this case, we compute the effect size (d) for the two extreme groups, using the same method as for a t-test.

$d = (Mean_{HI} – Mean_{LO}) / SD_{Within}$

Then, this value is adjusted to take account of the pattern of remaining means.

If the other means are clustered at the center,

$f = d * Sqrt(1 / (2 * k))$

If the means are spread uniformly over the range,

f = (d / 2) * Sqrt((k + 1) / (3 * (k – 1)))

If the means are at the extremes, we need to distinguish between two cases:

- If the number of groups is even, f = 0.5 * d.
- If the number of groups is odd, f = d * (Sqrt((k * k) – 1)) / (2 * k).

Computing the F-Value Required for Significance

The F-value required for significance (ReqF) is given by the central F distribution for significance level alpha and DF_1, DF_2 where DF_1 is degrees of freedom for the current factor (or interaction) and

$DF_2 = N_{total} – DF_{Factor} – DF_{OtherFactors} – DF_{Interactions} – DF_{Covariates}$

This value is obtained by the DFIN algorithm.

For each factor or interaction, the value of f computed above is squared to yield f^2, which is used in computing the non-centrality parameter (NCP):

$NCP = f^2*(DF_1 + DF_2 + 1)$

Equivalently,

$NCP = f^2*(N_{total} – DF_{other})$

where DF_{other} is the df associated with factors (or interactions) other than the current one, or with covariates.

Power is given by the non-central F distribution for NCP, ReqF, DF_1, and DF_2. The value is computed using the DFFNCD algorithm.

Note. The computation of NCP is based on the effect size and the DF's that are available for the current-test—that is, DF_{error} and $DF_{Numerator}$. DF's that have been assigned to other factors or interactions are excluded from this computation. It is possible to compute NCP as $f^2 * N_{total}$. But this would seem to be appropriate only when other factors and interactions are assumed to explain no variance and are therefore pooled with the error term. The formula used by this program is appropriate for the vast majority of cases and will be slightly conservative otherwise.

The algorithm adopted for computing power for ANOVA (and ANCOVA) is consistent with the algorithm used for multiple regression, where computation of NCP is similarly based on DF_{error} and DF for the current factor, with DF for other factors excluded.

One-Tailed versus Two-Tailed Tests

Analysis of variance is nondirectional.

Analysis of Covariance

Power computation for analysis of covariance (ANCOVA) proceeds exactly as for ANOVA except that the effect size and df are modified.

The effect size is adjusted as follows:

$$f_{adjusted} = f \,/\, \mathrm{Sqrt}(1 - R^2_{cov})$$

Additionally, the degrees of freedom attributed to the covariate are removed from DF_{error}, which affects the computation of ReqF and the NCP. Recall that the NCP is computed as $f^2 * (DF_1 + DF_2 + 1)$. The inclusion of a covariate will increase the first term but decrease the second term (since DF_2 will be lowered). Unless the sample size is quite small, the impact on f^2 will tend to overwhelm the impact on DF_2.

Multiple Regression

Definition of Terms

N_{total}	Total number of cases
Set A	Set of covariates
K_A	Number of variables in set A
I^2_A	Increment to R^2 for set A
Set B	Set for which power will be computed
K_B	Number of variables in set B
I^2_B	Increment to R^2 for set A
Set C	Set of additional covariates, entered subsequent to set B
K_C	Number of variables in set C
I^2_C	Increment to R^2 for set C

Note that the number of variables in set A and/or C may be set to 0, so this construction can be used to define any possible multiple regression analysis.

The program displays the multiple regression as a series of lines, each representing a set of variables. For each line it gives power for the increment due to that line and also for the cumulative impact of all variables through the current line.

- For analysis of line 1, set A is defined as nil, set B is defined as line 1, and set C is defined as lines 2–10.
- For analysis of line 2, set A is defined as line 1, set B is defined as line 2, and set C is defined as lines 3–10.
- For analysis of line 3, set A is defined as lines 1–2, set B is defined as line 3, and set C is defined as lines 4–10. The same approach is used for lines 4–10.

The program also allows the user to designate a set that consists of more than one line, say lines 4–6. In this case, set A is defined as lines 1–3, set B is lines 4–6, and set C is lines 7–10.

Model 1 versus Model 2 Error

Under model 2 error (the default), all sets (including set C) are assumed to be in the model at every point. Thus, for example, when power is computed for line 2 in a five-line model, the R^2_{Total} is based on sets A, B, and C (in effect, the cumulative R^2 through line 5), and the Df_{error} assumes that all five lines are in the model. Under model 1 error, set C is always nil. In other words, the analysis of each line assumes that all subsequent lines are empty. (Model 1/model 2 error should not be confused with 1-tail/2-tail tests or with type 1/type 2 error).

Computation of Power

$$R^2_{Total} = I^2_A + I^2_B + I^2_C$$
$$F^2 = I^2_B / (1 - R^2_{Total})$$

The F-value required for significance (ReqF) is given by the central F distribution for significance level alpha and DF_1, DF_2 where

$$DF_1 = K_B$$
$$DF_2 = N_{Cases} - K_A - K_B - K_C - 1$$

This value is obtained by the DFIN algorithm.

The non-centrality parameter is computed as

$$NCP = f^2 * (DF_1 + DF_2 + 1)$$

Equivalently,

NCP = $f^{2} * (N_{Total} - K_A - K_C)$

Power is given by the non-central F distribution for NCP, ReqF, DF_1, and DF_2. The value is computed using the DFFNCD algorithm.

One-Tailed versus Two-Tailed Tests

The multiple regression test is nondirectional.

Algorithms for Power (Logistic Regression)

We assume that the covariate vector $x = (x_1, x_2, \ldots)'$ can be partitioned into two groups of covariates $x_1, x_2, \ldots, x_{md}$, which have a finite number of values and $x_{md} + 1 \ldots, x_{md\,+\,mc}$ which, conditional on the value of $x_1, x_2, \ldots, x_{md}$, have a normal distribution.

We then describe an algorithm for calculating the power of the Wald test of hypotheses of the form $H_0 : A(B - B_0) = 0$, where B is the coefficient vector of the model.

Description of the Algorithm

Assumptions and Notation

Suppose that the vector $x_d = (x_1, x_2, \ldots, x_{md})'$ has K possible distinct-values, and let $w_{1,}\, w_{2\ldots,}\, w_k$ be the relative frequency of each pattern of values. We assume further that for each of these K values of x_d there is a mean vector, $\mu_s = (\mu_{1s}, \ldots, \mu_{m_c}, s)'$, and a covariance matrix, $\sum s$, which specify the distribution of $x_{md} + 1, \ldots, x_{md} + {}_{mc}$ when x_d has its s_{th} value.

We assume that the Wald test statistic

$$T = (\hat{\beta} - \beta_0)' A'(A \hat{I} + A') + A(\hat{\beta} - \beta_0)$$

will be used to test H_0, where $\hat{I}$ is the empirical information matrix evaluated at $\hat{\beta}$, the maximum likelihood estimate of β.

Let b be the value of B under the alternate hypothesis, then the distribution of T, is that of a non-central chi-square with degrees of freedom equal to the dimension of A and non-centrality parameter ($b - B_0)' A'(AI + A') + A(b - B_0)$, where I is the information matrix for the n observations evaluated at b. We then find power and sample size using this non-centrality parameter.

Computational Algorithm

We show how to find I for the logistic model.

Under the logistic model, the information matrix is

$$nE\left(xx'\frac{\exp(B'x)}{(1+\exp(B'X))^2} \right)$$

Let,

$$f(u) = \left(\frac{\exp(u)}{(1+\exp(u))^2} \right)$$

Then the expected value of I is simply

$$nE(xx'f(B'x))$$

We show how this is accomplished without multidimensional numerical quadrature below. We need to calculate the expectation of $x'x$ multiplied times a random variable that depends on $B'x$.

Recall that the first m_d covariates are discrete and have K distinct values, each with probabilities $w_1, w_{2 \ldots ,} w_k$, and the distribution of $x'x$ is different for each of these values. Let $E_1, E_2, \ldots$ be the expectations with respect to these distributions.

Therefore,

$$E(xx'f(B'x)) = \sum_{i=1}^{m} w_l E_l(xx'f(B'x))$$

We then need to compute $Ei(x'xf(B'x))$. In the description that follows, the subscript l is suppressed.

First, note that

$$E(x'xf(B'x)) = E(f(B'x)E(x'x|B'x)) =$$
$$E(f(B'x)(E(x|B'x)'E(x|B'x)+V(x\,|B'x)))$$

Let u be defined as the vector $x_1,\ldots,x_d,\mu_1,\ldots,mu_{mc})'$ and augment $\sum$ by a matrix whose first m_d columns and rows are zero and whose last $m_c * m_c$ submatrix is the old value of $\sum$.

Then

$$E(x'|B'x) = u + \sum B\left(B'\sum B\right)^{-1}\left(B'x - B'u\right)$$

and

$$V\left(x'|B'x\right) = \sum - \left(\sum B\right)\left(B'\sum B\right)^{-1}\left(\sum B\right)'$$

Let $\sigma = \sqrt{B'\sum B},\ \tau = B'u,\ z = (B'x - \tau),\ and\ \gamma = \sum B/\sigma$

Note that z has a standard normal distribution.

Then, using standard formulas for multivariate normal conditional distributions,

$$E\left(x'|B'x\right) = u + \gamma z$$

and

$$V(x'|B'x) = \sum - \gamma\gamma'$$

Define $e_i(\tau,\sigma) = E(z^i f(\sigma z + \tau)),\ for = 0,1,2$ where z has a standard normal distribution. This value will be found by numerical integration.

Substituting 2 and 3 into 2, we get

$$E(x'xf(B'x)) = (uu' + \sum - \gamma\gamma')e_0(\tau,\sigma) + (u\gamma' + \gamma u')e_1(\tau,\sigma) + \gamma\gamma' e_2(\tau,\sigma)$$

Algorithms for Power (Survival Analysis)

The log-rank statistic is given by

$$\frac{\left(\sum_t O_t - e_t\right)^2}{\sum_t e_t\left(1 - e_t\right)}$$

where the sum is over the death times t in the study. We assume that study period is divided into m periods with a uniform censoring distribution during each period and constant hazard and drop-out rate during each period. We assume that e_i and $E(o_t)$ are approximately constant throughout the period so that the non-centrality parameter can be approximated by

$$\frac{\sum_{i=1}^{m} d_i\left(E(o_t)_i - e_i\right)^2}{\sum_{i=1}^{m} d_i e_i (1 - e_i)}$$

We calculate d_i, the expected number of deaths in the i_{th} period, based on the censoring distribution and the constant hazard and drop out rate during the i_{th} period. We calculate e_i as the number at risk in the middle of the i_{th} interval in group 1 over the total number at risk. This is also calculated from the survival and censoring distributions. We approximate $e(O_t)$ similarly to e_i except that the number at risk in each treatment group is weighted by the hazard in that group. The non-central chi-square distribution is then used to calculate power or sample size.

This formula is deceptively simple but may be used even when the accrual rate, the hazard rate, and/or the attrition rate, vary from one time interval to the next. Additionally, it is remarkably accurate, yielding results that are, for all intents and purposes, identical to those obtained through time-consuming simulations.

Algorithms for Power (Equivalence of Means)

To compute power for a test that two means are equal, the program uses the same algorithm as for the test that two means are not equal. The difference is in the computation of the standardized difference, d.

In the traditional case (to prove that the groups differ), d is computed as

d= (Mean1 – Mean2)/SD

where Mean1 and Mean2 are the means of the two populations and SD is the common within group standard deviation.

By contrast, for the test of equivalence, d is computed as

d# = ((Mean1 – Mean3) – (Mean1 – Mean2)) / sd

where

Mean3 = Mean1 – Diff

and Diff is the acceptable difference.

Algorithms for Power (Equivalence of Proportions)

To compute power for a test that two proportions are equal, the program uses the same algorithm as for the test that two proportions are not equal. The difference is in the computation of the standardized difference, d. (Where the program offers a number of options for the latter, only one of these, the weighted normal approximation, is available for the test of equivalence.)

In the traditional case (to prove that the groups differ), d is computed as

d= (P1 – P2)

where P1 and P2 are the event rates in the two populations.

By contrast, for the test of equivalence, d is computed as

d = (P1 – P3) – (P – P2)

where

P3 = P1 – Diff

and Diff is the acceptable difference.

Appendix D
Computational Algorithms for Precision

T-Test for Means (One-Sample) with Estimated Variance

The width of the confidence interval computed for a *completed* study depends on the confidence level, the standard deviation, and the sample size. The width of the confidence interval for a *planned* study depends on these values and on the sampling distribution of the standard deviation as well.

The t-value corresponding to the confidence level (tCI) is given by the central t-distribution for (1 – $\mathrm{CI}_{\mathrm{Level}}$) / Tails for df = N – 1. The value is computed using the DTIN algorithm.

The standard error of the mean is given by

SE = SD1 / Sqrt(N)

The lower and upper limits of the confidence interval are given by

LowerLimit = Mean1 – t_{CI} * SE * 50% Tolerance Factor
UpperLimit = Mean1 + t_{CI} * SE * 50% Tolerance Factor

where the 50% Tolerance Factor = Sqrt($\mathrm{CHI}_{\mathrm{Req}}$ / df) for df = N – 1. This value is given by the function DCHIIN for 1 – Tolerance level, and df, and takes account of the sampling distribution of the standard deviation.

The value computed in this way is the median value. Half the studies will yield a confidence interval wider than the expected value and roughly half will yield a narrower interval.

One-Tailed versus Two-Tailed Tests

Assuming the 95% confidence level, the t-value computed for a one-tailed test would be based on 1 – 0.95 (that is, 0.05). The t-value corresponding to a two-tailed test would be based on (1 – 0.95)/2 (that is, 0.025).

The one-tailed "interval" yields a lower (or upper) limit closer to the observed value than does the two-tailed interval, but only one of the limits is meaningful. The one-tailed "interval" extends from minus infinity to the upper limit reported, or from the lower limit reported to positive infinity.

Tolerance intervals

The program is able to compute also the tolerance interval for a given confidence interval; that is, to report that "80% of studies will yield a 95% confidence interval no wider than a given value," by taking account of the sampling distribution of the standard deviations.

The sampling distribution of the sample variances follows a chi-squared distribution. For a given tolerance level, the Tolerance Factor in the equation for confidence intervals is modified to reflect the required tolerance level (say, 80% rather than 50%). Specifically, we compute FactorSD = Sqrt(CHI_{Req} / df). This value is given by the function DCHIIN for 1 – Tolerance level, and df.

T-Test for Means (Two-Sample) with Estimated Variance

The width of the confidence interval computed for a completed study depends on the confidence level, the standard deviation, and the sample size. The confidence interval for a planned study depends on these values and also on the sampling distribution of the standard deviation.

The pooled within-group variance SD_p is given by

$$\text{SD}_p = \text{Sqrt}((((N1 - 1) * SD1 * SD1 + (N2 - 1) * SD2 * SD2) / (N1 + N2 - 2)))$$

The standard error of the mean difference is given by

$$\text{SE}_{\text{diff}} = \text{SD}_p * \text{Sqrt}(1 / N1 + 1 / N2)$$

The degree of freedom is given by

$$df = N1 + N2 - 2$$

The t-value corresponding to the confidence level (t_{CI}) is given by the central t-distribution for (1 – CI_{Level}) / Tails and df. The value is computed using the DTIN algorithm.

The lower and upper limits of the confidence interval are given by

LowerLimit = Diff – t_{CI} * SE * 50% Tolerance Factor
UpperLimit = Diff + t_{CI} * SE * 50% Tolerance Factor

where

the 50% Tolerance Factor = Sqrt(CHI_{Req} / df) for df=N1+N2-2. This value is given by the function DCHIIN for 1–Tolerance level, and df. This takes account of the sampling distribution of the standard deviation.

The value computed in this way is the median value. Half of the studies will yield a confidence interval wider than the expected value, and roughly half will yield a narrower interval.

One-Tailed versus Two-Tailed Tests

Assuming the 95% confidence level, the t-value computed for a one-tailed test would be based on 1–0.95 (that is, 0.05). The t-value corresponding to a two-tailed test would be based on (1-0.95)/2 (that is, 0.025).

The one-tailed "interval" yields a lower (or upper) limit closer to the observed value than does the two-tailed interval, but only one of the limits is meaningful. The one-tailed "interval" extends from minus infinity to the upper limit reported, or from the lower limit reported to positive infinity.

Tolerance Intervals

The program is also able to compute the tolerance interval for a given confidence interval—that is, to report that 80% of studies will yield a 95% confidence interval no wider than a given value—by taking account of the sampling distribution of the standard deviations.

The sampling distribution of the sample variances follows a chi-squared distribution. For a given tolerance level, the Tolerance Factor in the equation for confidence intervals is modified to reflect the required tolerance level (say, 80% rather than 50%). Specifically, we compute FactorSD = Sqrt(CHI_{Req} / df). This value is given by the function DCHIIN for 1–Tolerance level, and df.

Z-Test for Means (One-Sample)

The width of the confidence interval computed for a completed study depends on the confidence level, the standard deviation, and the sample size. The confidence interval width for a planned study depends on these values as well. Since the (known) standard deviation will be used in computing confidence intervals, the standard deviation observed in the sample has no impact on the width of the interval.

The z-value corresponding to the confidence level (Z_{CI}) is given by the normal distribution for (1 – CI_{Level}) / Tails. The value is computed using the DNORIN algorithm.

The standard error of the mean is given by SE = SD1 / Sqrt(N).

The expected value of the lower and upper limits of the confidence interval are given by

$$\text{LowerLimit} = \text{Mean1} - Z_{CI} * \text{SE}$$
$$\text{UpperLimit} = \text{Mean1} + Z_{CI} * \text{SE}$$

Since the standard deviation is known, rather than estimated, the width of the confidence interval is completely determined by the sample size and will not vary from study to study.

One-Tailed versus Two-Tailed Tests

Assuming the 95% confidence level, the z-value computed for a one-tailed test would be based on 1–0.95 (that is, 0.05). The z-value corresponding to a two-tailed test would be based on (1–0.95)/2 (that is, 0.025).

The one-tailed “interval” yields a lower (or upper) limit closer to the observed value than does the two-tailed interval, but only one of the limits is meaningful. The one-tailed “interval” extends from minus infinity to the upper limit reported, or from the lower limit reported to positive infinity.

Tolerance intervals

When the standard deviation is known, the width of the confidence interval will not vary from study to study.

Z-Test for Means (Two-Sample)

The width of the confidence interval computed for a completed study depends on the confidence level, the standard deviation, and the sample size. The confidence interval width for a planned study depends on these values as well. Since the (known) standard deviation will be used in computing confidence intervals, the standard deviation observed in the sample has no impact on the width of the interval.

The z-value corresponding to the confidence level (Z_{CI}) is given by the normal distribution for (1 – CI_{Level}) / Tails. The value is computed using the DNORIN algorithm.

The pooled within-group variance SDp is given by

SDp = Sqrt((((N1 - 1) * SD1 * SD1 + (N2 - 1) * SD2 * SD2) / (N1 + N2 – 2)))

The standard error of the mean difference is given by

SE_{Diff} = SDp * Sqrt(1 / N1 + 1 / N2)

The degrees of freedom is given by

df = N1 + N2 – 2

The z-value corresponding to the confidence level (Z_{CI}) is given by the normal distribution for (1 – CI_{Level}) / Tails. The value is computed using the DNORIN algorithm.

The expected value of the lower and upper limits of the confidence interval are given by

LowerLimit = Mean1 – Z_{CI} * SE_{Diff}
UpperLimit = Mean1 + Z_{CI} * SE_{Diff}

Since the standard deviation is known, rather than estimated, the width of the confidence interval is completely determined by the sample size and will not vary from study to study.

One-Tailed versus Two-Tailed Tests

Assuming the 95% confidence level, the z-value computed for a one-tailed test would be based on 1 – 0.95 (that is, 0.05). The z-value corresponding to a two-tailed test would be based on (1 – 0.95)/2 (that is, 0.025).

The one-tailed "interval" yields a lower (or upper) limit closer to the observed value than does the two-tailed interval, but only one of the limits is meaningful. The one-tailed "interval" extends from minus infinity to the upper limit reported, or from the lower limit reported to positive infinity.

Tolerance Intervals

When the standard deviation is known, the width of the confidence interval will not vary from study to study.

One Proportion—Normal Approximation

The lower limit is computed as follows:

If P=0, the lower limit is 0
If P=1, the lower limit is $((1 - CI_{Level}) / \text{Tails})^{(1 / n)}$

Otherwise, the lower limit is

Lower = (P + (a / 2) - z * Sqrt(((P * Q) / n) + (a / (4 * n)))) / (1 + a)

where

Q=1 – P
Z is the z-value corresponding to $(1 - CI_{Level})$/Tails
$a = z^2 / n$
N is N of cases

The upper limit is computed as follows:

If P = 1, the upper limit is 1
If P = 0, the upper limit is $1 - (((1 - CI_{Level}) / \text{Tails})^{(1 / n)})$

Otherwise, the upper limit is

Upper = (P + (a / 2) + z * Sqrt(((P * Q) / n) + (a / (4 * n)))) / (1 + a)

where

Q = 1 – P
Z is the z-value corresponding to $(1 - CI_{Level})$/Tails
$a = z^2 / n$
N is the N of cases

The confidence interval width computed in this manner approximates the value expected over an infinite number of trials.

One Proportion—Exact (Binomial) Formula

The confidence interval for a single proportion depends on the confidence level and the sample size. It also depends on the proportion of successes observed in the study, since the standard error of a proportion varies as a function of the proportion.

To compute the expected width of the confidence interval, we need to take account of the likelihood of all possible outcomes, and the confidence interval associated with each of these.

For a study with $N = 10$, the possible outcomes are 0 successes, 1 success, …. 10 successes, and the likelihood of each is given by the binomial distribution. This value is given by the DBINPR function for N cases, k successes, and true proportion as specified by the alternate hypothesis.

For each possible outcome, the lower limit (and the upper limit) of the confidence interval that would be reported may be computed using the binomial distribution. These values are computed by the DBINES function for N cases, k successes, and the specified confidence level.

The program iterates through every possible outcome, multiplies the lower limit of the confidence interval by the likelihood that the given outcome will actually be observed in the study, and then sums these products over all possible outcomes to yield the expected lower limit. The process is repeated for the upper limit.

The confidence interval width computed in this manner gives the expected width over an infinite number of trials.

One-Tailed versus Two-Tailed Intervals

For a two-tailed confidence interval, the confidence level is used as given (for example, 95 for 95%). For a one-tailed confidence "interval," the value used for 95% is 100 – 2*(100-95). The one-tailed "interval" yields a lower (or upper) limit closer to the observed value than does the two-tailed interval, but only one of the limits is meaningful. The one-tailed "interval" extends from minus infinity to the upper limit reported, or from the lower limit reported to positive infinity.

Two Independent Proportions—Normal Approximations

Rate Difference

With Z = the z-value corresponding to $(1 - CI_{Level})$/Tails, the lower and upper limits of the rate difference are given by

Lower = Diff – Z * SE_{Diff}
Upper = Diff + Z * SE_{Diff}

where

SE_{Diff} = Sqrt(((RowPct(1, 1) * RowPct(1, 2)) / Cell(1, 3))
+ ((RowPct(2, 1) * RowPct(2, 2)) / Cell(2, 3)))

and RowPct(Row,Col) refers to the percent of the row.

Odds Ratio

The lower and upper limits of the log odds are computed as

Lower = (LOGODDS – z * $SE_{LOGODDS}$)
Upper = (LOGODDS + z * $SE_{LOGODDS}$)

where

$SE_{LOGODDS}$ = Sqrt(1 / (Cell(1, 1)) + 1 / (Cell(1, 2)) + 1 / (Cell(2, 1)) + 1 / (Cell(2, 2)))

The lower and upper limit of the odds ratio are then given by

Lower = Exp(Lower Limit Log Odds)
Upper = Exp(Upper Limit Log Odds)

Relative Risk

The lower and upper limits of the log of the relative risk are computed as

Lower= (LOGRR – z * SE_{LOGRR})
Upper= (LOGRR + z * SE_{LOGRR})

where

$$SE_{LOGRR} = \text{Sqrt}((\text{Cell}(1, 2) / \text{Cell}(1, 1)) / (\text{Cell}(1, 1) + \text{Cell}(1, 2)) + (\text{Cell}(2, 2) / \text{Cell}(2, 1)) / (\text{Cell}(2, 1) + \text{Cell}(2, 2)))$$

The lower and upper limit of the relative risk are then given by

Lower = Exp(Lower Limit Log RR)
Upper = Exp(Upper Limit Log RR)

Two Independent Proportions—Cornfield Method

The logic of the Cornfield method is as follows. The chi-square test is used to test the null hypothesis that the observed sample was drawn from a specific population. Typically, this is a population in which the proportion of positive cases is identical in the two rows (that is, the test of independence). However, this is not a requirement of the test.

The logic of the Cornfield method is to identify the first population that will yield a significant effect (with alpha = 1 – CI_{Level}). Since a 2×2 table has only one degree of freedom, all four cells are completely determined by any single cell. In this example, we use the upper left cell.

Beginning with the observed sample, we subtract one case from the upper left cell (and adjust the other three cells as required by the marginals) to yield a “population,” and test the observed study against this population. If the p-value exceeds 0.05 (assuming a 95% confidence level), we subtract an additional case. The process is repeated until we find a population that yields a p-value of exactly 0.05 for the test (the number of cases in the upper left cell can be a fractional value, so that a p-value of exactly 0.05 can be obtained).

The population identified in this way defines the lower limit. The rate difference, the odds ratio, and the relative risk corresponding to this population are reported as the lower limit for the population. To find the upper limit for each statistic, the process is repeated.

The computational methods used for confidence intervals for a 2 × 2 test of proportions are typically used to analyze a completed study, where the sample values are known. Since the sample values (which affect the variance) will vary from one sample to the next and cannot be known when the study is being planned, the confidence interval width reported here is only an estimate of the width that will be obtained in any given study.

Correlations (One-Sample)

The expected confidence interval for a correlation (r) is based on the Fisher-Z transformation of r (denoted Z_r).

The width of the confidence interval computed for a completed study depends on the confidence level, the observed correlation, and the sample size. The expected value of the confidence interval for a planned study depends on these values and also on the sampling distribution of the correlation coefficient.

The z-value corresponding to the confidence level (Z_{CI}) is given by the normal distribution for (1 – CI_{Level}) / Tails. The value is computed using the DNORN algorithm.

The confidence interval for r is computed using the Fisher-Z transformation of r,

$$Z_r = 0.5 * \text{Log}((1 + r) / (1 - r))$$

The standard error of Zr is given by

$$SE = \text{Sqrt}(1 / (N1 - 3))$$

The expected value of the lower and upper limits of the confidence interval in Zr units are given by

$$Z_{LOWER} = Z_r - Z_{CI} * SE$$
$$Z_{UPPER} = Z_r + Z_{CI} * SE$$

The expected values in r units are then computed by transforming the z-values into r values by

$$r_Z = (c - 1) / (c + 1), \text{ where } c = \text{Exp}(2 * Z)$$

The computational methods used for confidence intervals for a single correlation is typically used to analyze a completed study, where the sample values are known. Since the sample values (which affect the variance) will vary from one sample to the next and cannot be known when the study is being planned, the confidence interval width reported here is only an estimate of the width that will be obtained in any given study.

One-Tailed versus Two-Tailed Tests

Assuming the 95% confidence level, the t-value computed for a one-tailed test would be based on 1 – 0.95 (that is, 0.05). The t-value corresponding to a two-tailed test would be based on (1 – 0.95)/2,(that is, 0.025).

The one-tailed interval yields a lower (or upper) limit closer to the observed value than does the two-tailed interval, but only one of the limits is meaningful. The one-tailed interval extends from minus infinity to the upper limit reported, or from the lower limit reported to positive infinity.

Appendix E
Computational Algorithms for Clustered Trials

Two-level hierarchical designs with covariates

Suppose we have a two-level design (individuals within clusters such as schools) in which treatments are assigned at the highest level, the cluster level (level 2). Suppose that there are m clusters per treatment and n individuals per cluster. Suppose that there are covariates at both levels of the design. In particular suppose that there are q_2 covariates at the cluster level (level 2) that explain the proportion R_2^2 of the variance at the cluster level (so the cluster level multiple correlation is R_2), and q_1 covariates at the individual level (level 1) that explain the proportion R_1^2 of the variance at the individual within-cluster level (so the individual level multiple correlation is R_1).

Suppose that, before adjustment for any covariates, the between-cluster and between individual-within-cluster variance components are σ_2^2 and σ_1^2, respectively. Denote the covariate adjusted the between-cluster and between-individual-within-cluster variance components by σ_{A2}^2 and σ_{A1}^2, respectively. Thus we can define R_2^2 and R_1^2 as

$$R_2^2 = 1 - \sigma_{A2}^2 / \sigma_2^2$$

and

$$R_1^2 = 1 - \sigma_{A1}^2 / \sigma_1^2.$$

Define the intraclass correlation ρ (at the cluster level) by

$$\rho = \frac{\sigma_2^2}{\sigma_2^2 + \sigma_1^2} \quad . \tag{1.1}$$

Note that ρ is the proportion of total variance at the cluster level. The quantity $\bar{\rho}$ = $1-\rho$ is analogous to it in representing the proportion of the total variance that is at the individual level.

When cluster sizes n are equal, the test for treatment effects in two-level cluster randomized experiments is an exact t-test. The test statistic has a non-central t-distribution with $2m - 2 - q_2$ degrees of freedom and covariate adjusted noncentrality parameter

$$\lambda = d_T \sqrt{\frac{nm}{2}} \sqrt{\frac{1}{1+(n-1)\rho - \left[R_1^2 + \left(nR_2^2 - R_1^2\right)\rho\right]}} \tag{1.2}$$

where there are m clusters in each of the control and treatment groups, q_2 is the number of covariates at the cluster level (level 2), d_T is the effect size (the difference between the treatment and control group means divided by the total within-group standard deviation), and ρ is the (unadjusted) intraclass correlation. Note that the fact that we sample from a finite population of clusters has no impact on the power of the test for treatment effects. The reason is that, although the variance of the school-effects components in the treatment and control groups means is smaller when a finite population of schools is assigned to treatments, these school effects components are negatively correlated. Because the variance of the mean difference is the sum of the variance minus twice the covariance, the negative covariance term increases the variance of the difference and exactly cancels the reduction in variance due to finite population sampling.

Note that the maximum value of the noncentrality parameter as the cluster size $n \to \infty$ with fixed m is

$$\lambda_{Max} = \delta\sqrt{\frac{m}{2}}\sqrt{\frac{1}{\left(1-R_2^2\right)\rho}} \quad (1.3)$$

.

Of course, the maximum value of the noncentrality parameter tends to infinity as $m \to \infty$. These maxima are useful for computing the maximum power that can be obtained by increasing n with other design parameters fixed (for example in determining whether any value of n can attain a desired power).

The power of a level α one-tailed test for the treatment effect is therefore

$$p_1 = 1 - \mathrm{f}(c_\alpha, 2m - 2 - q_2\,, \lambda),$$

where $f(x, v, \lambda)$ is the cumulative distribution function of the noncentral t-distribution with v degrees of freedom, and noncentrality parameter λ and c_α is the $100(1 - \alpha)$ percentile of the central t-distribution with $2m - 2 - q_2$ degrees of freedom. The power of a level α two-tailed test for the treatment effect is therefore

$$p_2 = 1 - \mathrm{f}(c_{\alpha/2}, 2m - 2 - q_2\,, \lambda) + \mathrm{f}(-c_{\alpha/2}, 2m - 2 - q_2\,, \lambda).$$

Unequal sample sizes

When the numbers of observations in each cluster are not equal within each treatment group, the test is no longer exact. However if there are m^T clusters in the treatment group and m^C clusters in the control group, the test statistic has approximately a noncentral t-distribution, with $m^C + m^T - 2 - q_2$ degrees of freedom and covariate adjusted noncentrality parameter

$$\lambda = d_T\sqrt{\frac{N^C N^T}{N^C + N^T}}\sqrt{\frac{1}{1+\left(\tilde{n}_U - 1\right)\rho - \left[R_1^2 + \left(\tilde{n}_U R_2^2 - R_1^2\right)\rho\right]}} \quad (1.4)$$

where N^C and N^T are the total number of observations in the control and treatment groups, respectively, and

$$\tilde{n}_U = \frac{N^C \sum_{i=1}^{m^T} \left(n_i^T\right)^2}{N^T N} + \frac{N^T \sum_{i=1}^{m^C} \left(n_i^C\right)^2}{N^C N} \qquad (1.5)$$

where n_i^T is the number of observations in the ith cluster of the treatment group and n_i^C is the number of observations in the i^{th} cluster of the control group. Note that, when cluster sizes are equal within treatment groups so that $n_i^C = n^C$, i = 1, …, m^C and $n_i^T = n^T$, i = 1, …, m^T, $\tilde{n}_U$ in () reduces to

$$\tilde{n}_U = \frac{n^C n^T \left(m^C + m^T\right)}{m^C n^C + m^T n^T} . \qquad (1.6)$$

Note that, when $n^C = n^T = n$, $\tilde{n}_U$ in () reduces to n and () reduces to (1).

The power of a level α one-tailed test for the treatment effect is therefore

$$p_1 = 1 - \mathrm{f}(c_\alpha, m^C + m^T - 2 - q_2, \lambda),$$

where $f(x, v, \lambda)$ is the cumulative distribution function of the noncentral t-distribution with ν degrees of freedom and noncentrality parameter λ, and c_α is the 100(1 – α) percentile of the central t-distribution with $m^C + m^T - 2 - q_2$ degrees of freedom. The power of a level α two-tailed test for the treatment effect is therefore

$$p_2 = 1 - \mathrm{f}(c_{\alpha/2}, m^C + m^T - 2 - q_2 , \lambda) + \mathrm{f}(-c_{\alpha/2}, m^C + m^T - 2 - q_2 , \lambda).$$

Glossary

alpha

Alpha is the criterion required to establish statistical significance. Assuming that the null hypothesis is true, alpha is also the proportion of studies expected to result in a type 1 error. See “tails.”

beta

The proportion of studies that will yield a type 2 error. Equal to one minus power.

confidence interval

Interval that will include the population parameter in a known proportion of all possible studies.

confidence level

A level of certainty for the confidence interval; the proportion of studies in which the confidence interval is expected to include the population parameter.

effect size

The magnitude of the effect—for example, the standard difference in means (for a t-test). Power analysis works with the effect size, which is independent of sample size. Significance tests combine the observed effect with the sample size to yield a combined test statistic.

paired proportions

A paired analysis is used when cases in one group are somehow linked with cases in the other group (for example, the researcher recruits pairs of siblings and assigns one to either treatment, or patients are matched for severity of disease). When the outcome is a dichotomous classification (proportion), the study is described as a paired proportion (also known as McNemar’s test).

paired t-test

A paired analysis is used when cases in one group are somehow linked with cases in the other group (for example, the researcher recruits pairs of siblings and assigns one to either treatment, or patients are matched for severity of disease). When the study employs two groups and the outcome is continuous, the appropriate analysis is a paired (dependent) t-test.

power

Power is the proportion of studies that will yield a statistically significant effect (assuming the effect size, sample size, and criterion alpha specified in the study design).

precision

Used in this program to refer to the width of the confidence interval.

p-value

Computed for a significance test, the p-value gives the proportion of cases (for a population in which the effect is null) that will yield a sample in which the effect is as large (or larger) than the observed effect.

tails

A one-tailed test is used when an effect in one direction would be meaningful, but an effect in the opposite direction would have the same practical impact as no effect. Only an effect in the expected direction is interpreted. A two-tailed test is used when an effect in either direction would be meaningful (even if the researcher expects the effect to fall in a specified direction).

tolerance interval

Proportion of studies in which the width of the confidence interval is no greater than a given value.

type 1 error

The error committed when the true effect is null but the study yields a significant p-value and leads the researcher (in error) to reject the null hypothesis. With alpha set at 0.05, a type 1 error would be expected in 5% of trials in which the null hypothesis is true. By definition, a type 1 error will occur in 0% of trials in which the null hypothesis is false.

type 2 error

The error committed when the true effect is not null but the study fails to yield a significant p-value and the researcher (in error) fails to reject the null hypothesis. If power is 80%, a type 2 error would be expected in 20% of trials (assuming that the null hypothesis is false, which is an assumption of the power analysis). The type 2 error rate is also known as beta.

Bibliography

Altman, D. G. (1980). Statistics and ethics in medical research: How large a sample? *British Medical Journal* 281: 1336–1338.

Altman, D. G., Gore, S. M., Gardner, M. J., and Pocock, S. J. (1983). Statistical guidelines for contributors to medical journals. *British Medical Journal* 286: 1489–1493.

Bailar, J. C., and Mosteller, F. (1988). Guidelines for statistical reporting in articles for medical journals: Amplifications and explanations. *Annals of Internal Medicine* 108: 266–273.

Bakan, D. (1966). The effect of significance in psychological research. *Psychological Bulletin* 66: 423–437.

Berry, G. (1986). Statistical significance and confidence intervals. *Medical Journal of Australia* 144: 618–619.

Birnbaum, A. (1961). Confidence curves: An omnibus technique for estimation and testing statistical hypotheses. *Journal of the American Statistical Association* 56: 246–249.

Blackwelder, W. C. (1982). “Proving the null hypothesis” in clinical trials. *Controlled Clinical Trials* 3: 345–353.

Borenstein, M., Cohen, J., Rothstein, H. R., Pollack, S., et al. (1990). Statistical power analysis for one-way analysis of variance: A computer program. *Behavior Research Methods, Instruments, and Computers.*

Borenstein, M., Cohen, J., Rothstein, H. R., Pollack, S., and Kane, J. M. (1992). A visual approach to statistical power analysis on the microcomputer. *Behavior Research Methods, Instruments and Computers* 24: 565–572.

Borenstein, M. (1994a). Planning for precision in survival studies. *Journal of Clinical Epidemiology*.

_____. (1994b). The case for confidence intervals in controlled clinical trials. *Controlled Clinical Trials* 15: 411–428.

_____. (1994c). A note on confidence intervals in medical research. *Psychopharmacology Bulletin*.

_____. (1997). Hypothesis testing and effect size estimation in clinical trials. *Annals of Allergy, Asthma, and Immunology* 78: 5–16.

Braitman, L. E. (1988). Confidence intervals extract clinically useful information from data (editorial). *Annals of Internal Medicine* 108: 296–298.

Brewer, J. K. (1972). On the power of statistical tests in the American Educational Research Journal. *American Educational Research Journal* 9: 391–401.

Brewer, J. K., and Sindelar, P. T. (1988). Adequate sample size: A priori and post hoc considerations. *The Journal of Special Education* 21: 74–84.

Bristol, D. R. (1989). Sample sizes for constructing confidence intervals and testing hypotheses. *Statistics in Medicine* 8: 803–811.

Brown, J., and Hale, M. S. (1992). The power of statistical studies in consultation-liaison psychiatry. *Psychosomatics* 33: 437–443.

Bulpitt, C. J. (1987). Confidence intervals. *Lancet* 1: 494–497.

Chase, L. J., and Chase, R. B. (1976). A statistical power analysis of applied psychological research. *Journal of Applied Psychology* 61: 234–237.

Cobb, E. B. (1985). Planning research studies: An alternative to power analysis. *Nursing Research* 34: 386–388.

Cohen, J. (1962). The statistical power of abnormal-social psychological research: A review. *Journal of Abnormal and Social Psychology* 65(3): 145–153.

_____. (1965). Some statistical issues in psychological research. In *Handbook of Clinical Psychology*, ed. B. B. Wolman, 95–121. New York: McGraw-Hill.

_____. (1977). *Statistical power analysis for the behavioral sciences—Revised edition*. Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc.

_____. (1988). *Statistical power analysis for the behavioral sciences*. 2nd ed. Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc.

_____. (1990). Things I have learned (so far). *American Psychologist* 45: 1304–1312.

_____. (1992) A power primer. *Psychological Bulletin* 112: 155–159.

_____. (1994). The earth is round (p<.05). *American Psychologist* 49: 997–1003.

Cohen, J., and Cohen, P. (1983). *Applied multiple regression/correlation analysis for the behavioral sciences*. 2nd ed. Hillsdale, N.J.: Lawrence Erlbaum Associates, Inc.

Detsky, A. S., and Sackett, D. L. (1985). When was a negative clinical trial big enough? *Archives of Internal Medicine* 145: 709–712.

Dunlap, W. P. (1981). An interactive FORTRAN IV program for calculating power, sample size, or detectable differences in means. *Behavior Research Methods and Instrumentation* 13: 757–759.

Dupont, W. D., and Plummer, W. D. (1990). Power and sample size calculations: A review and computer program. *Controlled Clinical Trials* 11: 116–128.

Feinstein, A. R. (1975). The other side of "statistical significance": Alpha, beta, delta, and the calculation of sample size. *Clinical Pharmacology and Therapeutics* 18: 491–505.

_____. (1976). Clinical biostatistics XXXVII. Demeaned errors, confidence games, nonplused minuses, inefficient coefficients, and other statistical disruptions of scientific communication. *Clinical Pharmacology and Therapeutics* 20: 617–631.

Fisher, R. A. (1955). Statistical methods and scientific induction. *Journal of the Royal Statistical Society, Series B* 17: 69–78.

Fleiss, J. (1979). Confidence intervals for the odds ratio in case-control studies: The state of the art. *Journal of Chronic Diseases* 32: 69–77.

_____. (1981). *Statistical methods in rates and proportions*. 2nd ed. New York: John Wiley and Sons, Inc.

Fleiss, J. L. (1986a). Confidence intervals vs. significance tests: Quantitative interpretation (letter). *American Journal of Public Health* 76: 587–588.

_____. (1986b). Significance tests have a role in epidemiologic research: Reactions to A. M. Walker. *American Journal of Public Health* 76: 559–560.

Foster, D. A., and Sullivan, K. M. (1987). Computer program produces p-value graphics. *American Journal of Public Health* 77: 880–881.

Freiman, J. A., Chalmers, T. C., Smith, H. J., and Kuebler, R. R. (1978). The importance of beta, the type II error and sample size in the design and interpretation of the randomized control trial: Survey of 71 "negative" trials. *New England Journal of Medicine* 299: 690–694.

Gardner, M. J., and Altman, D. G. (1986). Confidence intervals rather than P values: Estimation rather than hypothesis testing. *British Medical Journal (Clinical Research Edition)* 292: 746–750.

Gardner, M. J., and Altman, D. G. (1989a). *Statistics with confidence—Confidence intervals and statistical guidelines*. London: BMJ.

_____. (1989b). *Statistics with confidence*. Belfast: The Universities Press.

Gart, J. J. (1962). Approximate confidence intervals for the relative risk. *Journal of the Royal Statistical Society* 24: 454–463.

Gart, J. J., and Thomas, D. G. (1972). Numerical results on approximate confidence limits for the odds ratio. *Journal of the Royal Statistical Society* 34: 441–447.

Gordon, I. (1987). Sample size estimation in occupational mortality studies with use of confidence interval theory. *American Journal of Epidemiology* 125: 158–162.

Gore, S. M. (1981). Assessing methods—Confidence intervals. *British Medical Journal* 283: 660–662.

Greenland, S. (1984). A counterexample to the test-based principle of setting confidence limits. *American Journal of Epidemiology* 120: 4–7.

_____. (1988). On sample-size and power calculations for studies using confidence intervals. *American Journal of Epidemiology* 128: 231–237.

Hahn, G. J., and Meeker, W. Q. (1991). *Statistical intervals*. New York: John Wiley and Sons, Inc.

Hanley, J., and Lippman-Hand, A. (1983). "If nothing goes wrong, is everything all right?": Interpreting zero numerators. *Journal of the American Medical Association* 249: 1743–1745.

Hartung, J., Cottrell, J. E., and Giffen, J. P. (1983). Absence of evidence is not evidence of absence. *Anesthesiology* 58: 298–300.

Ingelfinger, F. J. (1975). The confusions and consolations of uncertainty (letter). *New England Journal of Medicine* 292: 1402–1403.

International Committee of Medical Journal Editors. (1991). Uniform requirements for manuscripts submitted to biomedical journals. *New England Journal of Medicine* 324: 424–428.

Kahn, H. A., and Sempos, C. T. (1989). *Statistical methods in epidemiology*. New York: Oxford University Press.

Kleinbaum, D. G., Kupper, L. L., and Morgenstern, H. (1982). *Epidemiologic research*. Belmont, Calif.: Lifetime Learning Publications.

Kraemer, H. C., and Thiemann, S. (1987). *How many subjects? Statistical power analysis in research*. Newbury Park: Sage.

Lachin, J. M. (1981). Introduction to sample size determination and power analysis for clinical trials. *Controlled Clinical Trials* 2: 93–113.

Machin, D., and Gardner, M. J. (1988). Reply to confidence intervals (letter). *British Medical Journal* 296: 1372.

Mainland, D. (1982). Medical statistics—Thinking vs. arithmetic. *Journal of Chronic Diseases* 35: 413–417.

Makuch, R. W., and Johnson, M. F. (1986). Some issues in the design and interpretation of "negative" clinical studies. *Archives of Internal Medicine* 146: 986–989.

McHugh, R. B., and Le, C. T. (1984). Confidence estimation and the size of a clinical trial. *Controlled Clinical Trials* 5: 157–163.

Morgan, P. P. (1989). Confidence intervals: From statistical significance to clinical significance (editorial). *Canadian Medical Association Journal* 141: 881–883.

Murphy, K. R. (1990). If the null hypothesis is impossible, why test it? *American Psychologist* 45: 403–404.

Nelson, N., Rosenthal, R., and Rosnow, R. L. (1986). Interpretation of significance levels and effect sizes by psychological researchers. *American Psychologist* 1299–1301.

Phillips, W. C., Scott, J. A., and Blasczcynski, G. (1983). The significance of "no significance": What a negative statistical test really means. *American Journal of Radiology* 141: 203–206.

Poole, C. (1987a). Mr. Poole's response (letter). *American Journal of Public Health* 77: 880.

_____. (1987b). Beyond the confidence interval. *American Journal of Public Health* 77: 195–199.

_____. (1987c). Confidence intervals exclude nothing. *American Journal of Public Health* 77: 492–493.

Reed, J. F., and Slaichert, W. (1981). Statistical proof in inconclusive "negative" trials. *Archives of Internal Medicine* 141: 1307–1310.

Reynolds, T. B. (1980). Type II error in clinical trials (editor's reply to letter). *Gastroenterology* 79: 180

Rosenthal, R., and Rubin, D. (1985). Statistical analysis: Summarizing evidence versus establishing facts.

Rosner, B. (1990). *Fundamentals of biostatistics*. Boston: PWS-Kent Publishing Co.

Rosnow, R. L., and Rosenthal, R. (1988). Focused tests of significance and effect size estimation in counseling psychology. *Journal of Counseling Psychology* 35: 203–208.

Rothman, K. J. (1978). A show of confidence (letter). *New England Journal of Medicine* 299: 1362–1363.

_____. (1986a). Significance questing (editorial). *Annals of Internal Medicine* 105: 445–447.

_____. (1986b). *Modern epidemiology*. Boston: Little, Brown and Company.

Rothman, K. J., and Yankauer, A. (1986). Editors' note. *American Journal of Public Health* 76: 587–588.

Rothstein, H. R., Borenstein, M., Cohen, J., and Pollack, S. (1990). Statistical power analysis for multiple regression/correlation: A computer program. *Educational and Psychological Measurement*.

Sedlmeyer, P., and Gigerenzer, G. (1989). Do studies of statistical power have an effect on the power of studies? *Psychological Bulletin* 105: 309–316.

Sheehe, P. R. (1993). A variation on a confidence interval theme (letter). *Epidemiology*, 4: 185–187.

Simon, R. (1986). Confidence intervals for reporting results of clinical trials. *Annals of Internal Medicine* 105: 429–435.

Smith, A. H., and Bates, M. N. (1992). Confidence limit analyses should replace power calculations in the interpretation of epidemiologic studies. *Epidemiology* 3: 449–452.

_____. (1993). A variation on a confidence interval theme (reply to letter). *Epidemiology* 4: 186–187.

Sullivan, K. M., and Foster, D. A. (1990). Use of the confidence interval function. *Epidemiology* 1: 39–42.

Thompson, W. D. (1987a). Exclusion and uncertainty (letter). *American Journal of Public Health* 77: 879–880.

_____. (1987b). Statistical criteria in the interpretation of epidemiologic data [published erratum appears in *American Journal of Public Health*, April 1987, 77(4): 515]. *American Journal of Public Health* 77: 191–194.

_____. (1993). Policy and p-values (letter) [published erratum appears in *American Journal of Public Health*, April 1987, 77(4): 515]. *American Journal of Public Health.*

Tiku, M. L. (1967). Tables of the power of the F-test. *Journal of the American Statistical Association* 62: 525–539.

Tversky, A., and Kahneman, D. (1971). Belief in the law of small numbers. *Psychological Bulletin* 76: 105–110.

Walker, A. M. (1986a). Significance tests represent consensus and standard practice (letter). *American Journal of Public Health* 76: 1033–1034.

_____. (1986b). Reporting the results of epidemiologic studies. *American Journal of Public Health* 76: 556–558.

Wonnacott, T. (1985). Statistically significant. *Canadian Medical Association Journal* 133: 843.

Zelen, M., and Severo, N. C. (1964). Probability functions. In *Handbook of Mathematical Functions*, ed. M. Abromowitz, I. Stegan, et al., National Bureau of Standards. Washington, D.C.: U. S. Government Printing Office.

Index