

Chapter 220

Group-Sequential Tests for Two Proportions (Legacy)

This procedure is the original two-proportions group-sequential procedure in PASS. Power calculations and boundaries are generated from analytic calculations (simulation is not used). This procedure does not give any options for futility boundaries.

Introduction

Clinical trials are longitudinal. They accumulate data sequentially through time. The participants cannot be enrolled and randomized on the same day. Instead, they are enrolled as they enter the study. It may take several years to enroll enough patients to meet sample size requirements. Because clinical trials are long term studies, it is in the interest of both the participants and the researchers to monitor the accumulating information for early convincing evidence of either harm or benefit. This permits early termination of the trial.

Group sequential methods allow statistical tests to be performed on accumulating data while a phase III clinical trial is ongoing. Statistical theory and practical experience with these designs have shown that making four or five *interim analyses* is almost as effective in detecting large differences between treatment groups as performing a new analysis after each new data value. Besides saving time and resources, such a strategy can reduce the experimental subject's exposure to an inferior treatment and make superior treatments available sooner.

When repeated significance testing occurs on the same data, adjustments have to be made to the hypothesis testing procedure to maintain overall significance and power levels. The landmark paper of Lan & DeMets (1983) provided the theory behind the *alpha spending function* approach to group sequential testing. This paper built upon the earlier work of Armitage, McPherson, & Rowe (1969), Pocock (1977), and O'Brien & Fleming (1979). **PASS** implements the methods given in Reboussin, DeMets, Kim, & Lan (1992) to calculate the power and sample sizes of various group sequential designs.

This module calculates sample size and power for group sequential designs used to compare two group proportions. Other modules perform similar analyses for the comparison of means and survival functions. The program allows you to vary the number and times of interim tests, the type of alpha spending function, and the test boundaries. It also gives you complete flexibility in solving for power, significance level, sample size, or effect size. The results are displayed in both numeric reports and informative graphics.

Technical Details

Suppose the means of two samples of $N1$ and $N2$ individuals will be compared at various stages of a trial using the z_k statistic:

$$z_k = \frac{\hat{p}_{1k} - \hat{p}_{2k}}{\sqrt{\hat{p}_{1k}(1 - \hat{p}_{1k}) + \hat{p}_{2k}(1 - \hat{p}_{2k})}}$$

The subscript k indicates that the computations use all data that are available at the time of the k^{th} interim analysis or k^{th} look (k goes from 1 to K). This formula computes the standard z-test that is assumed to be normally distributed.

Spending Functions

Lan and DeMets (1983) introduced alpha spending functions, $\alpha(\tau)$, that determine a set of boundaries b_1, b_2, \dots, b_K for the sequence of test statistics z_1, z_2, \dots, z_K . These boundaries are the critical values of the sequential hypothesis tests. That is, after each interim test, the trial is continued as long as $|z_k| < b_k$. When $|z_k| \geq b_k$, the hypothesis of equal means is rejected, and the trial is stopped early.

The time argument τ either represents the proportion of elapsed time to the maximum duration of the trial or the proportion of the sample that has been collected. When elapsed time is being used it is referred to as *calendar time*. When time is measured in terms of the sample, it is referred to as *information time*. Since it is a proportion, τ can only vary between zero and one.

Alpha spending functions have the characteristics:

$$\alpha(0) = 0$$

$$\alpha(1) = \alpha$$

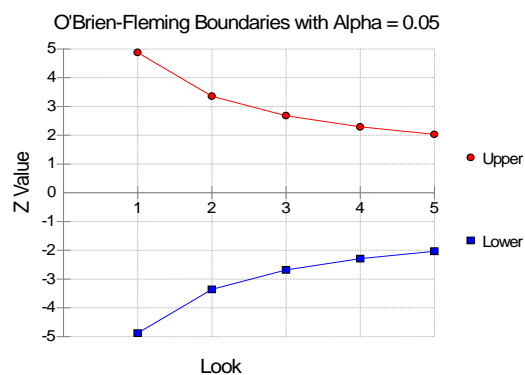
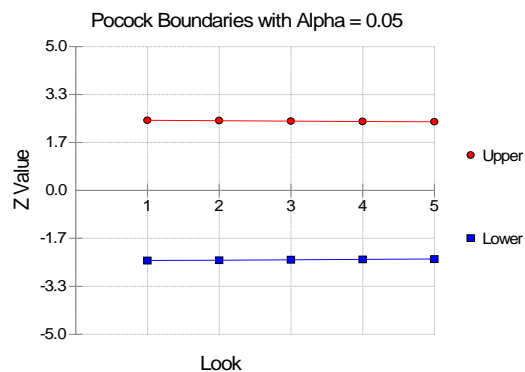
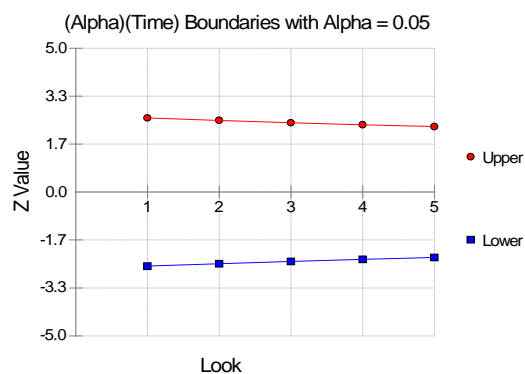
The last characteristic guarantees a fixed α level when the trial is complete. That is,

$$\Pr(|z_1| \geq b_1 \text{ or } |z_2| \geq b_2 \text{ or } \dots \text{ or } |z_K| \geq b_K) = \alpha(\tau)$$

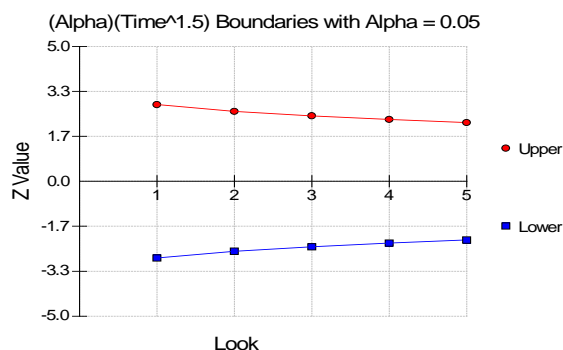
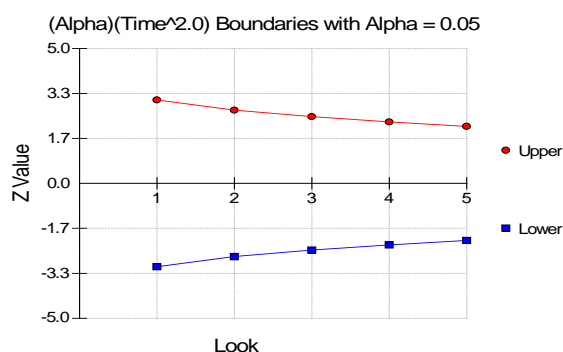
This methodology is very flexible since neither the times nor the number of analyses must be specified in advance. Only the functional form of $\alpha(\tau)$ must be specified.

PASS provides five popular spending functions plus the ability to enter and analyze your own boundaries. These are calculated as follows:

Group-Sequential Tests for Two Proportions (Legacy)

1. O'Brien-Fleming $2 - 2\Phi\left(\frac{Z_{\alpha/2}}{\sqrt{t}}\right)$ 2. Pocock $\alpha \cdot \ln(1 + (e - 1)t)$ 3. Alpha * time $\alpha \cdot t$ 

Group-Sequential Tests for Two Proportions (Legacy)

4. Alpha * time^{1.5} $\alpha \cdot t^{3/2}$ 5. Alpha * time² $\alpha \cdot t^2$ 

6. User-Supplied

A custom set of boundaries may be entered.

The O'Brien-Fleming boundaries are commonly used because they do not significantly increase the overall sample size and because they are conservative early in the trial. Conservative in the sense that the proportions must be extremely different before statistical significance is indicated. The Pocock boundaries are nearly equal for all times. The Alpha*t boundaries use equal amounts of alpha when the looks are equally spaced. You can enter your own set of boundaries using the User Supplied option.

Theory

A detailed account of the methodology is contained in Lan & DeMets (1983), DeMets & Lan (1984), Lan & Zucker (1993), and DeMets & Lan (1994). The theoretical basis of the method will be presented here.

Group sequential procedures for interim analysis are based on their equivalence to discrete boundary crossing of a Brownian motion process with drift parameter θ . The test statistics z_k follow the multivariate normal distribution with means $\theta\sqrt{\tau_k}$ and, for $j \leq k$, covariances $\sqrt{\tau_k/\tau_j}$. The drift parameter is related to the parameters of the z-test through the equation

$$\theta = \frac{p_1 - p_2}{\sqrt{\frac{\bar{p}(1-\bar{p})}{N_1} + \frac{\bar{p}(1-\bar{p})}{N_2}}}$$

where

$$\bar{p} = \frac{N_1 p_1 + N_2 p_2}{N_1 + N_2}$$

Hence, the algorithm is as follows:

1. Compute boundary values based on a specified spending function and alpha value.
2. Calculate the drift parameter based on those boundary values and a specified power value.
3. Use the drift parameter and the above equation to calculate the appropriate sample size.

Continuity Correction

The Continuity Correction option applies an adjustment to the sample sizes that is recommend by Fleiss(1981) page 45 to make the alpha and beta values more accurate. The formula for the adjustment is

$$N_{1new} = \frac{N_{1old}}{4} \left(1 + \sqrt{1 + \frac{2(R+1)}{R(N_{1old})|P1-P2|}} \right)^2$$

Example 1 – Finding the Sample Size

A clinical trial is to be conducted over a two-year period to compare the proportion response of a new treatment to that of the current treatment. The current response proportion is 0.53. Although the researchers do not know the true proportion of patients that will survive with the new treatment, they would like to examine the power that is achieved if the proportion under the new treatment is 0.63. In order to compare the sample size requirements for different effect sizes, it is also of interest to compute the sample size at response rates of 0.60, 0.65, 0.70, and 0.75.

Testing will be done at the 0.05 significance level and the power should be set to 0.90. A total of four tests are going to be performed on the data as they are obtained. The O'Brien-Fleming boundaries will be used.

Find the necessary sample sizes and test boundaries assuming equal sample sizes per arm and two-sided hypothesis tests.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load this procedure window. You may then make the appropriate entries as listed below, or open **Example 1** by going to the **File** menu and choosing **Open Example Template**.

Design Tab

Solve For	Sample Size
Alternative Hypothesis	Two-Sided
Continuity Correction	Checked
Power.....	0.90
Alpha.....	0.05
Group Allocation	Equal (N1 = N2)
P1 (Proportion in Group 1)	0.53
P2 (Proportion in Group 2)	0.60 0.63 0.65 0.70 0.75
Number of Looks	4
Spending Function	O'Brien-Fleming
Boundary Truncation.....	None
Max Time	2
Times.....	Equally Spaced
Informations	<Empty>

Group-Sequential Tests for Two Proportions (Legacy)

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Reports

Numeric Results for Two-Sided Test of Proportions. Continuity Correction Applied.

Solve For:

Target Power	Actual Power	N1	N2	N	P1	P2	Alpha
0.9	0.900168	1102	1102	2204	0.53	0.60	0.05
0.9	0.900930	542	542	1084	0.53	0.63	0.05
0.9	0.900408	376	376	752	0.53	0.65	0.05
0.9	0.901093	187	187	374	0.53	0.70	0.05
0.9	0.903126	111	111	222	0.53	0.75	0.05

Target Power	The desired power value (or values) entered in the procedure. Power is the probability of rejecting a false null hypothesis.
Actual Power	The power obtained in this scenario. Because N1 and N2 are discrete, this value is often (slightly) larger than the target power.
N1 and N2	The number of items sampled from each population.
N	The total sample size. $N = N1 + N2$.
P1	The proportion of populations 1 and 2 under the null hypothesis of equality.
P2	The proportion of population 2 at which power and sample size calculations are made.
Alpha	The probability of rejecting a true null hypothesis.

Summary Statements

Sample sizes of 1102 and 1102 achieve 90% power to detect a proportion difference of 0.07 ($P1 = 0.53$ and $P2 = 0.6$) with an overall Type I error rate (α) of 0.05 using a two-sided z-test with continuity correction. These results assume 4 sequential tests are made if the final stage is reached. The O'Brien-Fleming spending function was used to determine the test boundaries.

References

- Chow, S.C., Shao, J., and Wang, H. 2003. Sample Size Calculations in Clinical Research. Marcel Dekker. New York.
- Lan, K.K.G. and DeMets, D.L. 1983. 'Discrete sequential boundaries for clinical trials.' Biometrika, 70, pages 659-663.
- O'Brien, P.C. and Fleming, T.R. 1979. 'A multiple testing procedure for clinical trials.' Biometrics, 35, pages 549-556.
- Pocock, S.J. 1977. 'Group sequential methods in the design and analysis of clinical trials.' Biometrika, 64, pages 191-199.
- Reboussin, D.M., DeMets, D.L., Kim, K., and Lan, K.K.G. 1992. 'Programs for computing group sequential boundaries using the Lan-DeMets Method.' Technical Report 60, Department of Biostatistics, University of Wisconsin-Madison.

This report shows the values of each of the parameters, one scenario per row. Note that 542 participants in each arm of the study are required to meet the 90% power requirement when the proportion is 0.63.

The values from this table are in the chart below. Note that this plot actually occurs further down in the report.

Group-Sequential Tests for Two Proportions (Legacy)

Details Section

Details when Spending = O'Brien-Fleming, N1 = 1102, N2 = 1102, P1 = 0.53, P2 = 0.6, Continuity Correction.

Look	Time	Lower Bndry	Upper Bndry	Nominal Alpha	Inc Alpha	Total Alpha	Inc Power	Total Power
1	0.5	-4.33263	4.33263	0.000	0.000	0.000	0.003502	0.004
2	1.0	-2.96311	2.96311	0.003	0.003	0.003	0.254594	0.258
3	1.5	-2.35902	2.35902	0.018	0.016	0.019	0.427460	0.686
4	2.0	-2.01406	2.01406	0.044	0.031	0.050	0.214611	0.900

Drift = 3.27202

This report shows information about the individual interim tests. One report is generated for each scenario.

Look

These are the sequence numbers of the interim tests.

Time

These are the time points at which the interim tests are conducted. Since the Max Time was set to 2 (for two years), these time values are in years. Hence, the first interim test is at half a year, the second at one year, and so on.

We could have set Max Time to 24 so that the time scale was in months.

Lower and Upper Boundary

These are the test boundaries. If the computed value of the test statistic z is between these values, the trial should continue. Otherwise, the trial can be stopped.

Nominal Alpha

This is the value of alpha for these boundaries if they were used for a single, standalone, test. Hence, this is the significance level that must be found for this look in a standard statistical package that does not adjust for multiple looks.

Inc Alpha

This is the amount of alpha that is *spent* by this interim test. It is close to, but not equal to, the value of alpha that would be achieved if only a single test was conducted. For example, if we lookup the third value, 2.35902, in normal probability tables, we find that this corresponds to a (two-sided) alpha of 0.018323. However, the entry is 0.016248. The difference is due to the correction that must be made for multiple tests.

Total Alpha

This is the total amount of alpha that is used up to and including the current test.

Inc Power

These are the amounts that are added to the total power at each interim test. They are often called the exit probabilities because they give the probability that significance is found and the trial is stopped, given the alternative hypothesis.

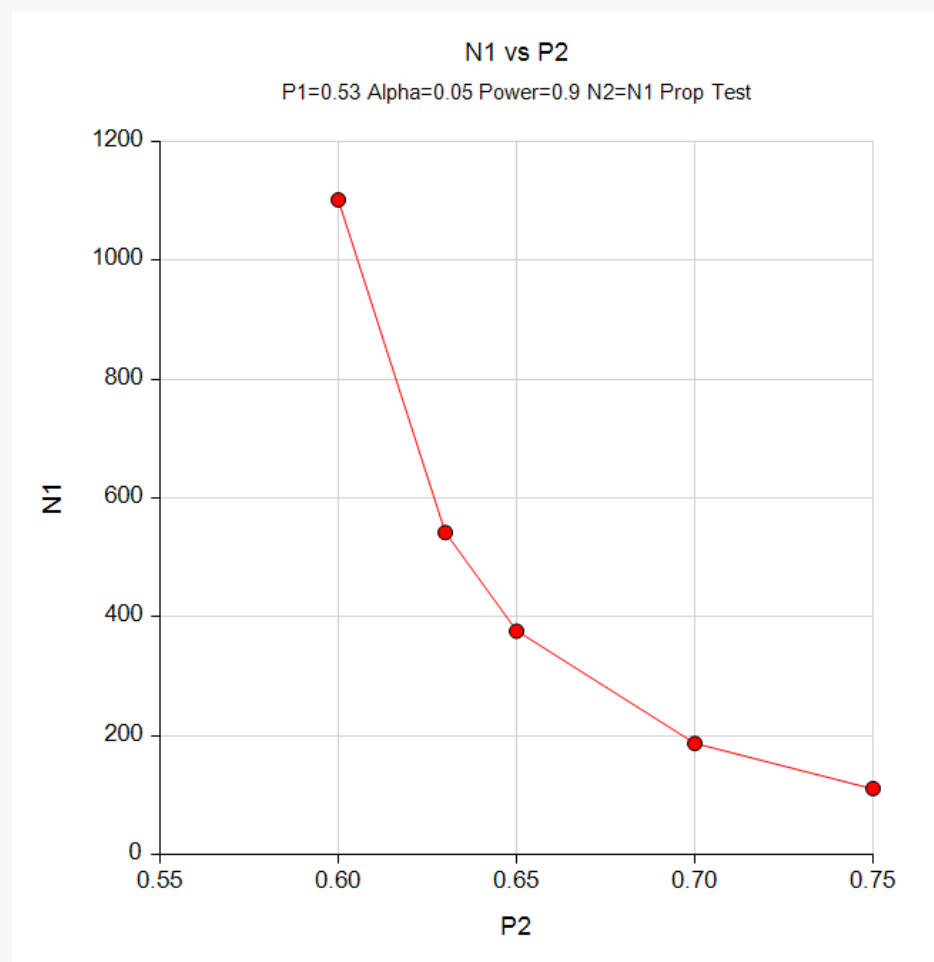
Group-Sequential Tests for Two Proportions (Legacy)

Total Power

These are the cumulative power values. They are also the cumulative exit probabilities. That is, they are the probability that the trial is stopped at or before the corresponding time.

Drift

This is the value of the Brownian motion drift parameter.

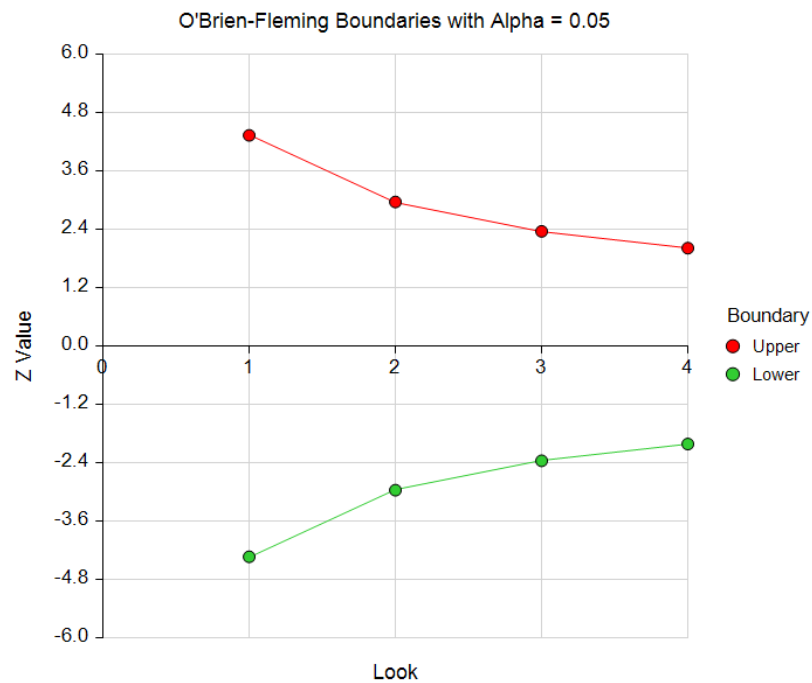
Plots Section**Plots**

This plot shows that a large increase in sample size is necessary when the detectable proportion in group two is less than 0.63.

Group-Sequential Tests for Two Proportions (Legacy)

Boundary Plots

Plots



This plot shows the interim boundaries for each look. This plot shows very dramatically that the results must be extremely significant at early looks, but that they are near the single test boundary (1.96 and -1.96) at the last look.

Example 2 – Finding the Power

Continuing the scenario began in Example1, the researcher wishes to calculate the power of the design at sample sizes 200, 400, 600, 800, 1000. Testing will be done at the 0.01, 0.05, 0.10 significance levels and the overall power will be set to 0.10. Find the power of these sample sizes and test boundaries assuming equal sample sizes per arm and two-sided hypothesis tests.

Proceeding as in Example1, we decide to translate the mean and standard deviation into a percent of mean scale.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load this procedure window. You may then make the appropriate entries as listed below, or open **Example 2** by going to the **File** menu and choosing **Open Example Template**.

Design Tab

Solve For	Power
Alternative Hypothesis	Two-Sided
Continuity Correction	Checked
Alpha.....	0.01 0.05 0.10
Group Allocation	Equal (N1 = N2)
Sample Size Per Group	200 to 1000 by 200
P1 (Proportion in Group 1).....	0.53
P2 (Proportion in Group 2).....	0.63
Number of Looks	4
Spending Function	O'Brien-Fleming
Boundary Truncation.....	None
Max Time	2
Times.....	Equally Spaced
Informations	<Empty>

Group-Sequential Tests for Two Proportions (Legacy)

Output

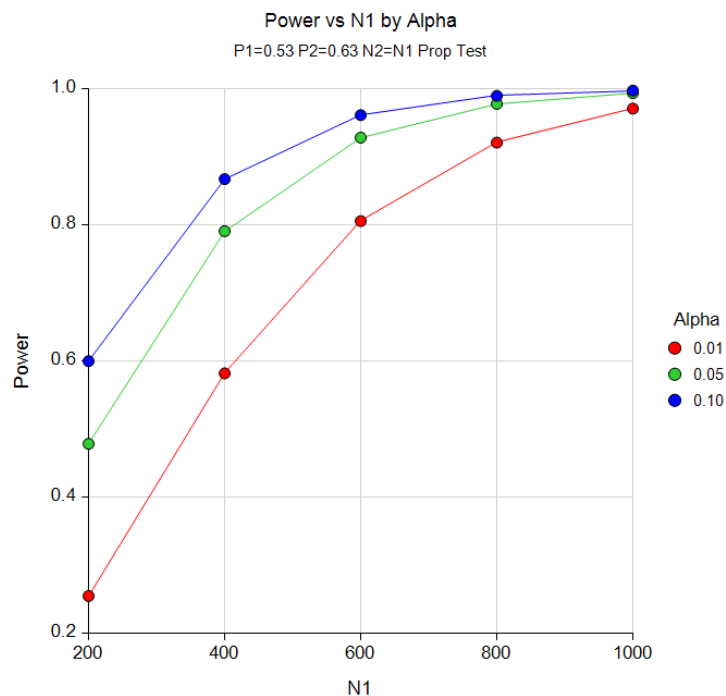
Click the Calculate button to perform the calculations and generate the following output.

Numeric Results for Two-Sided Test of Proportions. Continuity Correction Applied.

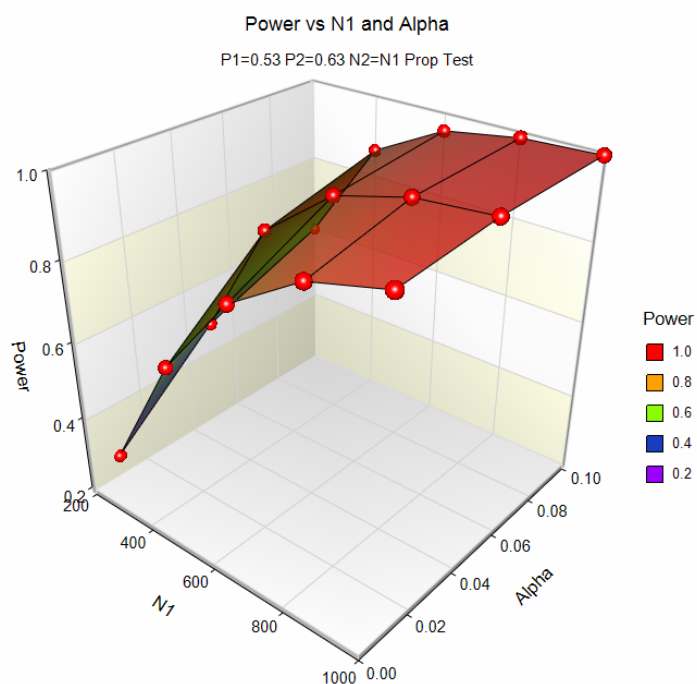
Solve For: **Power**

Power	N1	N2	N	P1	P2	Alpha
0.254797	200	200	400	0.53	0.63	0.01
0.581920	400	400	800	0.53	0.63	0.01
0.805679	600	600	1200	0.53	0.63	0.01
0.920951	800	800	1600	0.53	0.63	0.01
0.970898	1000	1000	2000	0.53	0.63	0.01
0.478378	200	200	400	0.53	0.63	0.05
0.790815	400	400	800	0.53	0.63	0.05
0.928267	600	600	1200	0.53	0.63	0.05
0.977859	800	800	1600	0.53	0.63	0.05
0.993673	1000	1000	2000	0.53	0.63	0.05
0.599827	200	200	400	0.53	0.63	0.10
0.867330	400	400	800	0.53	0.63	0.10
0.961331	600	600	1200	0.53	0.63	0.10
0.989659	800	800	1600	0.53	0.63	0.10
0.997396	1000	1000	2000	0.53	0.63	0.10

Plots



Group-Sequential Tests for Two Proportions (Legacy)



These data show the power for various sample sizes and alphas. It is interesting to note that once the sample size is greater than 700, the value of alpha makes little difference on the value of power.

Example 3 – Effect of Number of Looks

Continuing with examples one and two, it is interesting to determine the impact of the number of looks on power. **PASS** allows only one value for the Number of Looks parameter per run, so it will be necessary to run several analyses. To conduct this study, set alpha to 0.05, N1 to 500, and leave the other parameters as before. Run the analysis with Number of Looks equal to 1, 2, 3, 4, 6, 8, 10, and 20. Record the power for each run.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load this procedure window. You may then make the appropriate entries as listed below, or open **Example 3** by going to the **File** menu and choosing **Open Example Template**.

Design Tab

Solve For	Power
Alternative Hypothesis	Two-Sided
Continuity Correction	Not checked
Alpha.....	0.05
Group Allocation	Equal (N1 = N2)
Sample Size Per Group	500
P1 (Proportion in Group 1).....	0.53
P2 (Proportion in Group 2).....	0.63
Number of Looks	1 (Also run with 2, 3, 4, 6, 8, 10, and 20)
Spending Function	O'Brien-Fleming
Boundary Truncation.....	None
Max Time	2
Times	Equally Spaced
Informations	<Empty>

Group-Sequential Tests for Two Proportions (Legacy)

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results for Two-Sided Test of Proportions

Solve For: [Power](#)

Power	N1	N2	N	P1	P2	Alpha	Looks
0.893174	500	500	1000	0.53	0.63	0.05	1
0.892118	500	500	1000	0.53	0.63	0.050	2
0.889620	500	500	1000	0.53	0.63	0.050	3
0.887691	500	500	1000	0.53	0.63	0.050	4
0.885125	500	500	1000	0.53	0.63	0.050	6
0.883535	500	500	1000	0.53	0.63	0.050	8
0.882456	500	500	1000	0.53	0.63	0.050	10
0.879929	500	500	1000	0.53	0.63	0.050	20

This analysis shows how little the number of looks impacts the power of the design. The power of a study with no interim looks is 0.893174. When twenty interim looks are made, the power falls to 0.879929—a very small change.

Example 4 – Studying a Boundary Set

Continuing with the previous examples, suppose that you are presented with a set of boundaries and want to find the quality of the design (as measured by alpha and power). This is easy to do with **PASS**. Suppose that the analysis is to be run with five interim looks at equally spaced time points. The upper boundaries to be studied are 3.5, 3.5, 3.0, 2.5, 2.0. The lower boundaries are symmetric. The analysis would be run as follows.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load this procedure window. You may then make the appropriate entries as listed below, or open **Example 4** by going to the **File** menu and choosing **Open Example Template**.

Design Tab

Solve For	Power
Alternative Hypothesis	Two-Sided
Continuity Correction	Not checked
Alpha.....	0.05 (will be calculated from boundaries)
Group Allocation	Equal (N1 = N2)
Sample Size Per Group	500
P1 (Proportion in Group 1).....	0.53
P2 (Proportion in Group 2).....	0.63
Number of Looks	5
Spending Function.....	User Supplied
Boundary Truncation.....	None
Max Time	2
Times	Equally Spaced
Informations	<Empty>

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results for Two-Sided Test of Proportions

Solve For: **Power**

Power	N1	N2	N	P1	P2	Alpha
0.887792	500	500	1000	0.53	0.63	0.048157

Power	The probability of rejecting a false null hypothesis when the alternative hypothesis is true.
N1 and N2	The number of items sampled from each population.
N	The total sample size. $N = N1 + N2$.
P1	The proportion of populations 1 and 2 under the null hypothesis of equality.
P2	The proportion of population 2 at which power and sample size calculations are made.
Alpha	The probability of rejecting a true null hypothesis.

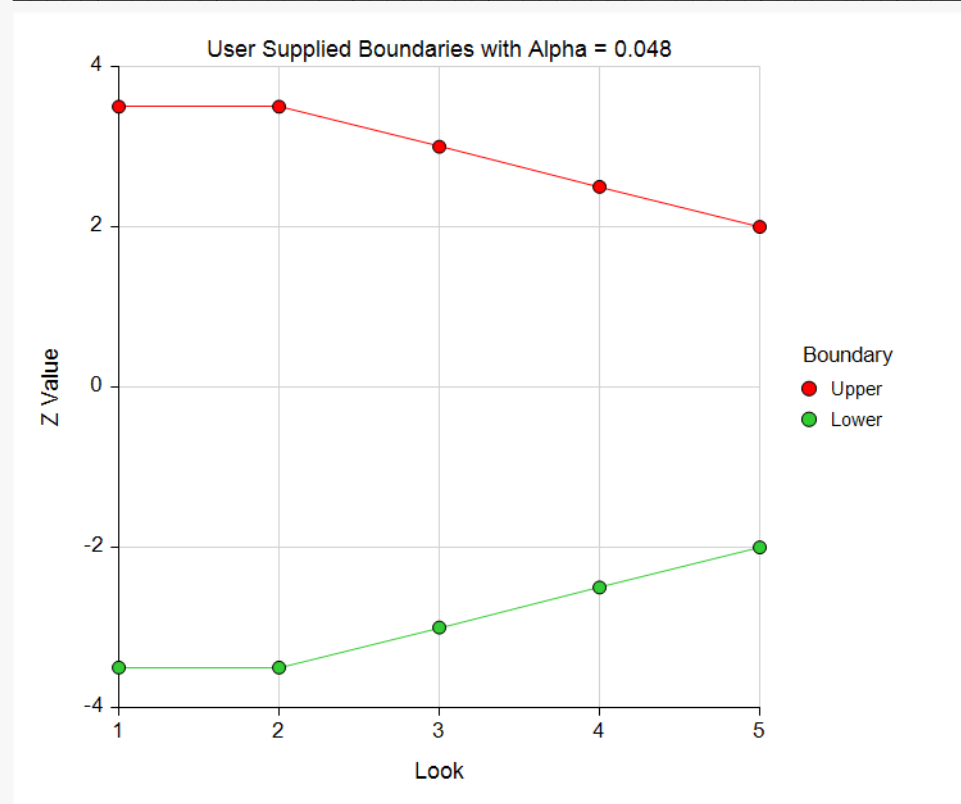
Group-Sequential Tests for Two Proportions (Legacy)

Details when Spending = User Supplied, $N1 = 500$, $N2 = 500$, $P1 = 0.53$, $P2 = 0.63$

Look	Time	Lower Bndry	Upper Bndry	Nominal Alpha	Inc Alpha	Total Alpha	Inc Power	Total Power
1	0.4	-3.5	3.5	0.000465	0.000465	0.000465	0.019352	0.019352
2	0.8	-3.5	3.5	0.000465	0.000408	0.000874	0.058108	0.077460
3	1.2	-3.0	3.0	0.002700	0.002410	0.003284	0.230567	0.308027
4	1.6	-2.5	2.5	0.012419	0.010331	0.013615	0.339341	0.647367
5	2.0	-2.0	2.0	0.045500	0.034542	0.048157	0.240425	0.887792

Drift = 3.20355

Plots



The power for this design is about 0.89. This value depends on both the boundaries and the sample size. The alpha level is about 0.048. This value only depends on the boundaries.

Example 5 – Validation using O’Brien-Fleming Boundaries

Reboussin (1992) presents an example for binomial distributed data for a design with two-sided O’Brien-Fleming boundaries, looks = 5, $\alpha = 0.05$, $\beta = 0.10$, $P_1 = 0.1100$, $P_2 = 0.0825$. They compute a drift of 3.28 and a sample size of 2381.78 per group. The upper boundaries are: 4.8769, 3.3569, 2.6803, 2.2898, 2.0310.

To test that PASS provides the same result, enter the following.

Setup

This section presents the values of each of the parameters needed to run this example. First, from the PASS Home window, load this procedure window. You may then make the appropriate entries as listed below, or open **Example 5** by going to the **File** menu and choosing **Open Example Template**.

Design Tab

Solve For	Sample Size
Alternative Hypothesis	Two-Sided
Continuity Correction	Not checked
Power.....	0.90
Alpha.....	0.05
Group Allocation	Equal (N1 = N2)
P1 (Proportion in Group 1).....	0.1100
P2 (Proportion in Group 2).....	0.0825
Number of Looks	5
Spending Function.....	O’Brien-Fleming
Boundary Truncation.....	None
Max Time	1
Times.....	Equally Spaced
Informations.....	<Empty>

Output

Click the Calculate button to perform the calculations and generate the following output.

Numeric Results for Two-Sided Test of Proportions

Solve For: [Sample Size](#)

Target Power	Actual Power	N1	N2	N	P1	P2	Alpha
0.9	0.90011	2474	2474	4948	0.11	0.0825	0.05

Group-Sequential Tests for Two Proportions (Legacy)

Details when Spending = O'Brien-Fleming, N1 = 2474, N2 = 2474, P1 = 0.11, P2 = 0.0825

Look	Time	Lower Bndry	Upper Bndry	Nominal Alpha	Inc Alpha	Total Alpha	Inc Power	Total Power
1	0.2	-4.87688	4.87688	0.000001	0.000001	0.000001	0.00032	0.000324
2	0.4	-3.35695	3.35695	0.000788	0.000787	0.000788	0.09945	0.099778
3	0.6	-2.68026	2.68026	0.007357	0.006828	0.007616	0.34670	0.446477
4	0.8	-2.28979	2.28979	0.022034	0.016807	0.024424	0.29964	0.746120
5	1.0	-2.03100	2.03100	0.042255	0.025576	0.050000	0.15399	0.900105

Drift = 3.27939

The difference in the sample sizes (2474 versus 2382) is due to rounding errors in the Reboussin article. Reboussin rounds from four-digits to three-digits, which caused a large difference. **PASS** uses more accurate routines.

To see that the results are equal to within rounding error, we will compute the sample size using Reboussin's results, but with more decimal places in the intermediate steps. They had

$$n_K = \frac{2(0.096)(0.904)(3.28)^2}{(0.028)^2} = 2381.78$$

When we compute this without rounding, we get

$$n_K = \frac{2(0.09625)(0.90375)(3.27939)^2}{(0.0275)^2} = 2474.00$$

A sample size of 2474 is the result obtained in PASS.