Chapter 213

Equivalence Tests for the Difference Between Two Proportions

Introduction

This module provides power analysis and sample size calculation for equivalence tests of the difference in two-sample designs in which the outcome is binary. The equivalence test is usually carried out using the Two One-Sided Tests (TOST) method. This procedure computes power and sample size for the TOST equivalence test method. Users may choose from among eight popular test statistics commonly used for running the hypothesis test.

The power calculations assume that independent, random samples are drawn from two populations.

Example

An equivalence test example will set the stage for the discussion of the terminology that follows. Suppose that the response rate of the standard treatment of a disease is 0.70. Unfortunately, this treatment is expensive and occasionally exhibits serious side-effects. A promising new treatment has been developed to the point where it can be tested. One of the first questions that must be answered is whether the new treatment is therapeutically equivalent to the standard treatment.

Because of the many benefits of the new treatment, clinicians are willing to adopt the new treatment even if its effectiveness is slightly different from the standard. After thoughtful discussion with several clinicians, it is decided that if the response rate of the new treatment is between 0.63 and 0.77, the new treatment would be adopted. The *margin of equivalence* is 0.07.

The developers must design an experiment to test the hypothesis that the response rate of the new treatment does not differ from that of the standard treatment by more than 0.07. The statistical hypothesis to be tested is

$$H_0: |p_1 - p_2| > 0.07 \quad \text{versus} \quad H_1: |p_1 - p_2| \le 0.07$$

Technical Details

The details of sample size calculation for the two-sample design for binary outcomes are presented in the chapter "Tests for Two Proportions," and they will not be duplicated here. Instead, this chapter only discusses those changes necessary for equivalence tests.

Approximate sample size formulas for equivalence tests for the difference between two proportions are presented in Julius and Campbell (2012), section 3.4. Only large sample (normal approximation) results are given there. It is also possible to calculate power based on the enumeration of all possible values in the binomial distribution. Both options are available in this procedure.

Suppose you have two populations from which dichotomous (binary) responses will be recorded. Assume without loss of generality that higher proportions are better. The probability (or risk) of cure in group 1 (the treatment group) is p_1 and in group 2 (the reference group) is p_2. Random samples of n_1 and n_2 individuals are obtained from these two groups. The data from these samples can be displayed in a 2-by-2 contingency table as follows

Group	Success	Failure	Total
Treatment	*a*	*c*	*m*
Control	*b*	*d*	*n*
Totals	*s*	*f*	*N*

The following alternative notation is also used.

Group	Success	Failure	Total
Treatment	x_{11}	x_{12}	n_1
Control	x_{21}	x_{22}	n_2
Totals	m_1	m_2	N

The binomial proportions p_1 and p_2 are estimated from these data using the formulae

$$\hat{p}_1 = \frac{a}{m} = \frac{x_{11}}{n_1} \text{ and } \hat{p}_2 = \frac{b}{n} = \frac{x_{21}}{n_2}$$

Let $p_{1.0}$ represent the group 1 proportion tested by the null hypothesis H_0. The power of a test is computed at a specific value of the proportion which we will call $p_{1.1}$. Let δ represent the smallest difference (margin of equivalence) between the two proportions that still results in the conclusion that the new treatment is equivalent to the current treatment. The set of statistical hypotheses that are tested is

$$H_0: |p_{1.0} - p_2| \geq \delta \quad \text{versus} \quad H_1: |p_{1.0} - p_2| < \delta$$

These hypotheses can be rearranged to give

$$H_0: p_{1.0} - p_2 \leq \delta_L \text{ or } p_{1.0} - p_2 \geq \delta_U \quad \text{versus} \quad H_1: \delta_L \leq p_{1.0} - p_2 \leq \delta_U$$

This composite hypothesis can be reduced to two one-sided hypotheses as follows

$$H_{0L}: p_{1.0} - p_2 \leq \delta_L \quad \text{versus} \quad H_{1L}: \delta_L \leq p_{1.0} - p_2$$

$$H_{0U}: p_{1.0} - p_2 \geq \delta_U \quad \text{versus} \quad H_{1U}: \delta_U \geq p_{1.0} - p_2$$

There are three common methods of specifying the margin of equivalence. The most direct is to simply give values for p_2 and $p_{1.0}$. However, it is often more meaningful to give p_2 and then specify $p_{1.0}$ implicitly by reporting the difference, ratio, or odds ratio. Mathematically, the definitions of these parameterizations are

Parameter	Computation	Alternative Hypotheses
Difference	$\delta = p_{1.0} - p_2$	$H_1: \delta_L \le p_{1.0} - p_2 \le \delta_U$
Ratio	$\phi = p_{1.0} / p_2$	$H_1: \phi_L \le p_{1.0} / p_2 \le \phi_U$
Odds Ratio	$\psi = Odds_{1.0} / Odds_2$	$H_1: \psi_L \le o_{1.0} / o_2 \le \psi_U$

Difference

The difference is perhaps the most direct method of comparison between two proportions. It is easy to interpret and communicate. It gives the absolute impact of the treatment. However, there are subtle difficulties that can arise with its interpretation.

One difficulty arises when the event of interest is rare. If a difference of 0.001 occurs when the baseline probability is 0.40, it would be dismissed as being trivial. However, if the baseline probability of a disease is 0.002, a 0.001 decrease would represent a reduction of 50%. Thus, interpretation of the difference depends on the baseline probability of the event.

Note that $\delta_L < 0$ and $\delta_U < 0$. Usually, $\delta_L = -\delta_U$.

Equivalence using a Difference

The following example might help you understand the concept of an *equivalence* test. Suppose 60% of patients respond to the current treatment method ($p_2 = 0.60$). If the response rate of the new treatment is no less than five percentage points better or worse than the existing treatment, it will be considered to be equivalent. Substituting these figures into the statistical hypotheses gives

$$H_0: p_{1.0} - p_2 \le -0.05 \quad \text{or} \quad p_{1.0} - p_2 \ge 0.05 \quad \text{versus} \quad H_1: -0.05 \le p_{1.0} - p_2 \le 0.05$$

Using the relationship

$$p_{1.0} = p_2 + \delta$$

gives

$$H_0: p_{1.0} \le 0.55 \quad \text{or} \quad p_{1.0} \ge 0.65 \quad \text{versus} \quad H_1: 0.55 \le p_{1.0} \le 0.65$$

In this example, when the null hypothesis is rejected, the concluded alternative is that the response rate is between 0.55 and 0.65.

The equivalence test is usually carried out using the Two One-Sided Tests (TOST) method. This procedure computes power and sample size for the TOST equivalence test method.

Power Calculation

The power for a test statistic that is based on the normal approximation can be computed exactly using two binomial distributions. The following steps are taken to compute the power of these tests.

1. Find the critical values using the standard normal distribution. The critical values z_L and z_U are chosen as that value of *z* that leaves exactly the target value of alpha in the appropriate tail of the normal distribution.

2. Compute the value of the test statistic z_t for every combination of x_{11} and x_{21}. Note that x_{11} ranges from 0 to n_1, and x_{21} ranges from 0 to n_2. A small value (around 0.0001) can be added to the zero-cell counts to avoid numerical problems that occur when the cell value is zero.

3. If $z_t > z_L$ and $z_t < z_U$, the combination is in the rejection region. Call all combinations of x_{11} and x_{21} that lead to a rejection the set *A.*

4. Compute the power for given values of $p_{1.1}$ and p_2 as

$$1-\beta=\sum_{A}\binom{n_1}{x_{11}}p_{1.1}^{x_{11}}q_{1.1}^{n_1-x_{11}}\binom{n_2}{x_{21}}p_2^{x_{21}}q_2^{n_2-x_{21}}$$

5. Compute the actual value of alpha achieved by the design by substituting $p_{1.0L}$ and $p_{1.0U}$ for $p_{1.1}$ to obtain

$$\alpha_L=\sum_{A}\binom{n_1}{x_{11}}p_{1.0L}^{x_{11}}q_{1.0L}^{n_1-x_{11}}\binom{n_2}{x_{21}}p_2^{x_{21}}q_2^{n_2-x_{21}}$$

   and

$$\alpha_U=\sum_{A}\binom{n_1}{x_{11}}p_{1.0U}^{x_{11}}q_{1.0U}^{n_1-x_{11}}\binom{n_2}{x_{21}}p_2^{x_{21}}q_2^{n_2-x_{21}}$$

The value of alpha is then computed as the maximum of α_L and α_U.

Asymptotic Approximations

When the values of n_1 and n_2 are large (say over 200), these formulas take a long time to evaluate. In this case, a large sample approximation can be used. The large sample approximation is made by replacing the values of $\hat{p}_1$ and $\hat{p}_2$ in the *z* statistic with the corresponding values of $p_{1.1}$ and p_2 and then computing the results based on the normal distribution. Note that in large samples, the Farrington and Manning statistic is substituted for the Gart and Nam statistic.

Test Statistics

Several test statistics have been proposed for testing whether the difference, ratio, or odds ratio are different from a specified value. The main difference among the several test statistics is in the formula used to compute the standard error used in the denominator. These tests are based on the following *z*-test

$$z_t = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0 - c}{\hat{\sigma}}$$

The constant, *c*, represents a continuity correction that is applied in some cases. When the continuity correction is not used, *c* is zero. In power calculations, the values of $\hat{p}_1$ and $\hat{p}_2$ are not known. The corresponding values of $p_{1.1}$ and p_2 can be reasonable substitutes.

Following is a list of the test statistics available in **PASS**. The availability of several test statistics begs the question of which test statistic one should use. The answer is simple: one should use the test statistic that will be used to analyze the data. You may choose a method because it is a standard in your industry, because it seems to have better statistical properties, or because your statistical package calculates it. Whatever your reasons for selecting a certain test statistic, you should use the same test statistic when doing the analysis after the data have been collected.

Z Test (Pooled)

This test was first proposed by Karl Pearson in 1900. Although this test is usually expressed directly as a chi-square statistic, it is expressed here as a *z* statistic so that it can be more easily used for one-sided hypothesis testing. The proportions are pooled (averaged) in computing the standard error. The formula for the test statistic is

$$z_t = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\hat{\sigma}_1}$$

where

$$\hat{\sigma}_1 = \sqrt{\bar{p}(1 - \bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

Z Test (Unpooled)

This test statistic does not pool the two proportions in computing the standard error.

$$z_t = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\hat{\sigma}_2}$$

where

$$\hat{\sigma}_2 = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

Z Test with Continuity Correction (Pooled)

This test is the same as Z Test (Pooled), except that a continuity correction is used. Remember that in the null case, the continuity correction makes the results closer to those of Fisher's Exact test.

$$z_t = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0 + \frac{F}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{\hat{\sigma}_1}$$

where

$$\hat{\sigma}_1 = \sqrt{\bar{p}(1-\bar{p})\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}$$

$$\bar{p} = \frac{n_1\hat{p}_1 + n_2\hat{p}_2}{n_1 + n_2}$$

and *F* is -1 for lower-tailed hypotheses and 1 for upper-tailed hypotheses.

Z Test with Continuity Correction (Unpooled)

This test is the same as the Z Test (Unpooled), except that a continuity correction is used. Remember that in the null case, the continuity correction makes the results closer to those of Fisher's Exact test.

$$z_t = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0 - \frac{F}{2}\left(\frac{1}{n_1} + \frac{1}{n_2}\right)}{\hat{\sigma}_2}$$

where

$$\hat{\sigma}_2 = \sqrt{\frac{\hat{p}_1(1-\hat{p}_1)}{n_1} + \frac{\hat{p}_2(1-\hat{p}_2)}{n_2}}$$

and *F* is -1 for lower-tailed hypotheses and 1 for upper-tailed hypotheses.

T-Test

Because of a detailed, comparative study of the behavior of several tests, D'Agostino (1988) and Upton (1982) proposed using the usual two-sample t-test for testing whether the two proportions are equal. One substitutes a '1' for a success and a '0' for a failure in the usual, two-sample *t*-test formula.

Miettinen and Nurminen's Likelihood Score Test

Miettinen and Nurminen (1985) proposed a test statistic for testing whether the difference is equal to a specified, non-zero, value, δ_0. The regular MLE's, $\hat{p}_1$ and $\hat{p}_2$, are used in the numerator of the score statistic while MLE's $\tilde{p}_1$ and $\tilde{p}_2$, constrained so that $\tilde{p}_1 - \tilde{p}_2 = \delta_0$, are used in the denominator. A correction factor of *N*/(*N*-1) is applied to make the variance estimate less biased. The significance level of the test statistic is based on the asymptotic normality of the score statistic.

The formula for computing this test statistic is

$$z_{MND} = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\hat{\sigma}_{MND}}$$

where

$$\hat{\sigma}_{MND} = \sqrt{\left(\frac{\tilde{p}_1\tilde{q}_1}{n_1} + \frac{\tilde{p}_2\tilde{q}_2}{n_2}\right)\left(\frac{N}{N-1}\right)}$$

$$\tilde{p}_1 = \tilde{p}_2 + \delta_0$$

$$\tilde{p}_2 = 2B\cos(A) - \frac{L_2}{3L_3}$$

$$A = \frac{1}{3}\left[\pi + \cos^{-1}\left(\frac{C}{B^3}\right)\right]$$

$$B = \operatorname{sign}(C)\sqrt{\frac{L_2^2}{9L_3^2} - \frac{L_1}{3L_3}}$$

$$C = \frac{L_2^3}{27L_3^3} - \frac{L_1L_2}{6L_3^2} + \frac{L_0}{2L_3}$$

$$L_0 = x_{21}\delta_0(1-\delta_0)$$

$$L_1 = [n_2\delta_0 - N - 2x_{21}]\delta_0 + m_1$$

$$L_2 = (N + n_2)\delta_0 - N - m_1$$

$$L_3 = N$$

Farrington and Manning's Likelihood Score Test

Farrington and Manning (1990) proposed a test statistic for testing whether the difference is equal to a specified value, δ_0. The regular MLE's, $\hat{p}_1$ and $\hat{p}_2$, are used in the numerator of the score statistic while MLE's $\tilde{p}_1$ and $\tilde{p}_2$, constrained so that $\tilde{p}_1 - \tilde{p}_2 = \delta_0$, are used in the denominator. The significance level of the test statistic is based on the asymptotic normality of the score statistic.

The formula for computing the test statistic is

$$z_{FMD} = \frac{\hat{p}_1 - \hat{p}_2 - \delta_0}{\sqrt{\left(\frac{\tilde{p}_1\tilde{q}_1}{n_1} + \frac{\tilde{p}_2\tilde{q}_2}{n_2}\right)}}$$

where the estimates $\tilde{p}_1$ and $\tilde{p}_2$ are computed as in the corresponding test of Miettinen and Nurminen (1985) given above.

Gart and Nam's Likelihood Score Test

Gart and Nam (1990), page 638, proposed a modification to the Farrington and Manning (1988) difference test that corrects for skewness. Let $z_{FMD}(\delta)$ stand for the Farrington and Manning difference test statistic described above. The skewness-corrected test statistic, z_{GND}, is the appropriate solution to the quadratic equation

$$(-\tilde{\gamma})z_{GND}^2 + (-1)z_{GND} + (z_{FMD}(\delta) + \tilde{\gamma}) = 0$$

where

$$\tilde{\gamma} = \frac{\tilde{V}^{3/2}(\delta)}{6}\left(\frac{\tilde{p}_1\tilde{q}_1(\tilde{q}_1 - \tilde{p}_1)}{n_1^2} - \frac{\tilde{p}_2\tilde{q}_2(\tilde{q}_2 - \tilde{p}_2)}{n_2^2}\right)$$

Example 1 – Finding Power

A study is being designed to establish the equivalence of a new treatment compared to the current treatment. Historically, the current treatment has enjoyed a 50% cure rate. The new treatment reduces the seriousness of certain side effects that occur with the current treatment. Thus, the new treatment will be adopted even if it is slightly less effective than the current treatment. The researchers will recommend adoption of the new treatment if its cure rate is within 15% of the standard treatment.

The researchers plan to use the Farrington and Manning likelihood score test statistic to analyze the data. They want to study the power of the Farrington and Manning test at group sample sizes ranging from 50 to 500, for detecting a difference inside 15%, when the actual cure rate of the new treatment ranges from 50% to 60%. The significance level will be 0.05.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 1** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab

Solve For .... **Power**
Power Calculation Method .... **Normal Approximation**
Test Type .... **Likelihood Score (Farr. & Mann.)**
Alpha **0.05**
Group Allocation .... **Equal (N1 = N2)**
Sample Size Per Group .... **50 to 500 by 50**
Input Type .... **Differences**
D0.U (Upper Equivalence Difference) **0.15**
D0.L (Lower Equivalence Difference) .... **-D0.U**
D1 (Actual Difference) **0.00 0.05 0.10**
P2 (Group 2 Proportion) **0.5**

Output

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

Numeric Reports

Numeric Results

Solve For: Power
Groups: 1 = Treatment, 2 = Reference
Test Statistic: Farrington & Manning Likelihood Score Test
Hypotheses: H0: P1 - P2 ≤ D0.L or P1 - P2 ≥ D0.U vs. H1: D0.L < P1 - P2 < D0.U

	Sample Size			Proportions				Difference			
				Equivalence				Equivalence			
Power*	N1	N2	N	Lower P1.0L	Upper P1.0U	Actual P1.1	Reference P2	Lower D0.L	Upper D0.U	Actual D1	Alpha
0.0000	50	50	100	0.35	0.65	0.50	0.5	-0.15	0.15	0.00	0.05
0.3795	100	100	200	0.35	0.65	0.50	0.5	-0.15	0.15	0.00	0.05
0.6689	150	150	300	0.35	0.65	0.50	0.5	-0.15	0.15	0.00	0.05
0.8305	200	200	400	0.35	0.65	0.50	0.5	-0.15	0.15	0.00	0.05
0.9160	250	250	500	0.35	0.65	0.50	0.5	-0.15	0.15	0.00	0.05
0.9594	300	300	600	0.35	0.65	0.50	0.5	-0.15	0.15	0.00	0.05
0.9808	350	350	700	0.35	0.65	0.50	0.5	-0.15	0.15	0.00	0.05
0.9911	400	400	800	0.35	0.65	0.50	0.5	-0.15	0.15	0.00	0.05
0.9959	450	450	900	0.35	0.65	0.50	0.5	-0.15	0.15	0.00	0.05
0.9982	500	500	1000	0.35	0.65	0.50	0.5	-0.15	0.15	0.00	0.05
0.0000	50	50	100	0.35	0.65	0.55	0.5	-0.15	0.15	0.05	0.05
0.3029	100	100	200	0.35	0.65	0.55	0.5	-0.15	0.15	0.05	0.05
0.5104	150	150	300	0.35	0.65	0.55	0.5	-0.15	0.15	0.05	0.05
0.6382	200	200	400	0.35	0.65	0.55	0.5	-0.15	0.15	0.05	0.05
0.7280	250	250	500	0.35	0.65	0.55	0.5	-0.15	0.15	0.05	0.05
0.7954	300	300	600	0.35	0.65	0.55	0.5	-0.15	0.15	0.05	0.05
0.8470	350	350	700	0.35	0.65	0.55	0.5	-0.15	0.15	0.05	0.05
0.8863	400	400	800	0.35	0.65	0.55	0.5	-0.15	0.15	0.05	0.05
0.9161	450	450	900	0.35	0.65	0.55	0.5	-0.15	0.15	0.05	0.05
0.9384	500	500	1000	0.35	0.65	0.55	0.5	-0.15	0.15	0.05	0.05
0.0000	50	50	100	0.35	0.65	0.60	0.5	-0.15	0.15	0.10	0.05
0.1523	100	100	200	0.35	0.65	0.60	0.5	-0.15	0.15	0.10	0.05
0.2206	150	150	300	0.35	0.65	0.60	0.5	-0.15	0.15	0.10	0.05
0.2659	200	200	400	0.35	0.65	0.60	0.5	-0.15	0.15	0.10	0.05
0.3067	250	250	500	0.35	0.65	0.60	0.5	-0.15	0.15	0.10	0.05
0.3455	300	300	600	0.35	0.65	0.60	0.5	-0.15	0.15	0.10	0.05
0.3827	350	350	700	0.35	0.65	0.60	0.5	-0.15	0.15	0.10	0.05
0.4184	400	400	800	0.35	0.65	0.60	0.5	-0.15	0.15	0.10	0.05
0.4525	450	450	900	0.35	0.65	0.60	0.5	-0.15	0.15	0.10	0.05
0.4851	500	500	1000	0.35	0.65	0.60	0.5	-0.15	0.15	0.10	0.05

* Power was computed using the normal approximation method.

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.
P1.0L	The smallest treatment-group response rate that still yields an equivalence conclusion.
P1.0U	The largest treatment-group response rate that still yields an equivalence conclusion.
P1.1	The proportion for group 1 assumed by the alternative hypothesis, H1. Group 1 is the treatment group. P1.1 = P1|H1.
P2	The proportion for group 2. Group 2 is the standard, reference, or control group.
D0.L	The lowest difference that still results in the conclusion of equivalence.
D0.U	The highest difference that still results in the conclusion of equivalence.
D1	The actual difference at which the power is calculated. D1 = P1.1 - P2.
Alpha	The probability of rejecting a true null hypothesis.

Summary Statements

A parallel two-group design will be used to test whether the Group 1 (treatment) proportion (P1) is equivalent to the Group 2 (reference) proportion (P2), with difference equivalence bounds of -0.15 and 0.15 (H0: P1 - P2 ≤ -0.15 or P1 - P2 ≥ 0.15 versus H1: -0.15 < P1 - P2 < 0.15). The comparison will be made using two one-sided, two-sample likelihood score (Farrington & Manning) tests with an overall Type I error rate (α) of 0.05. The reference group proportion is assumed to be 0.5. To detect a proportion difference (P1 - P2) of 0 (or P1 of 0.5) with sample sizes of 50 for Group 1 (treatment) and 50 for Group 2 (reference), the power is 0.

Dropout-Inflated Sample Size

	Sample Size			Dropout-Inflated Enrollment Sample Size			Expected Number of Dropouts		
Dropout Rate	**N1**	**N2**	**N**	**N1'**	**N2'**	**N'**	**D1**	**D2**	**D**
20%	50	50	100	63	63	126	13	13	26
20%	100	100	200	125	125	250	25	25	50
20%	150	150	300	188	188	376	38	38	76
20%	200	200	400	250	250	500	50	50	100
20%	250	250	500	313	313	626	63	63	126
20%	300	300	600	375	375	750	75	75	150
20%	350	350	700	438	438	876	88	88	176
20%	400	400	800	500	500	1000	100	100	200
20%	450	450	900	563	563	1126	113	113	226
20%	500	500	1000	625	625	1250	125	125	250

Dropout Rate The percentage of subjects (or items) that are expected to be lost at random during the course of the study and for whom no response data will be collected (i.e., will be treated as "missing"). Abbreviated as DR.
N1, N2, and N The evaluable sample sizes at which power is computed (as entered by the user). If N1 and N2 subjects are evaluated out of the N1' and N2' subjects that are enrolled in the study, the design will achieve the stated power.
N1', N2', and N' The number of subjects that should be enrolled in the study in order to obtain N1, N2, and N evaluable subjects, based on the assumed dropout rate. N1' and N2' are calculated by inflating N1 and N2 using the formulas N1' = N1 / (1 - DR) and N2' = N2 / (1 - DR), with N1' and N2' always rounded up. (See Julious, S.A. (2010) pages 52-53, or Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. (2018) pages 32-33.)
D1, D2, and D The expected number of dropouts. D1 = N1' - N1, D2 = N2' - N2, and D = D1 + D2.

Dropout Summary Statements

Anticipating a 20% dropout rate, 63 subjects should be enrolled in Group 1, and 63 in Group 2, to obtain final group sample sizes of 50 and 50, respectively.

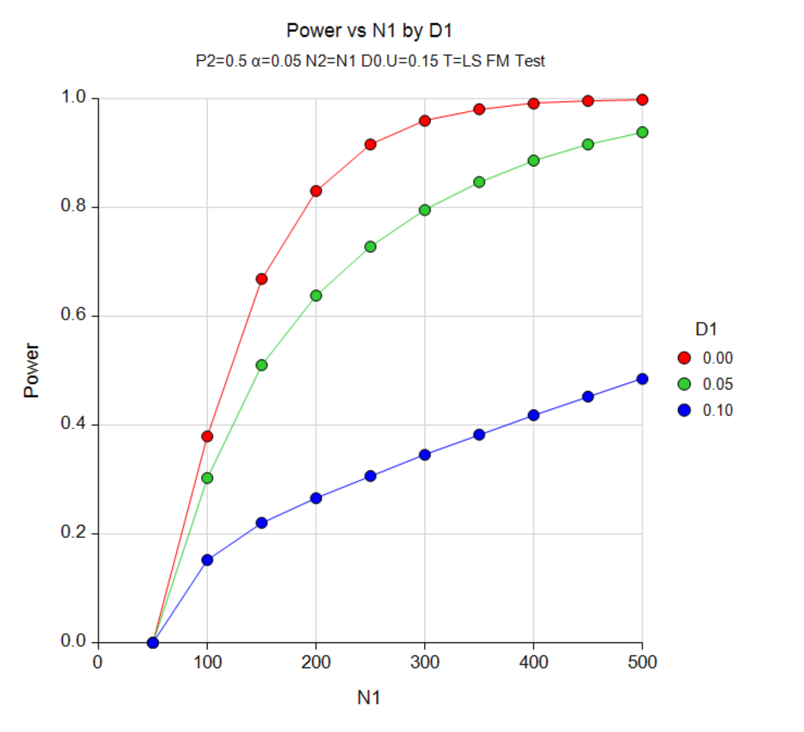
References

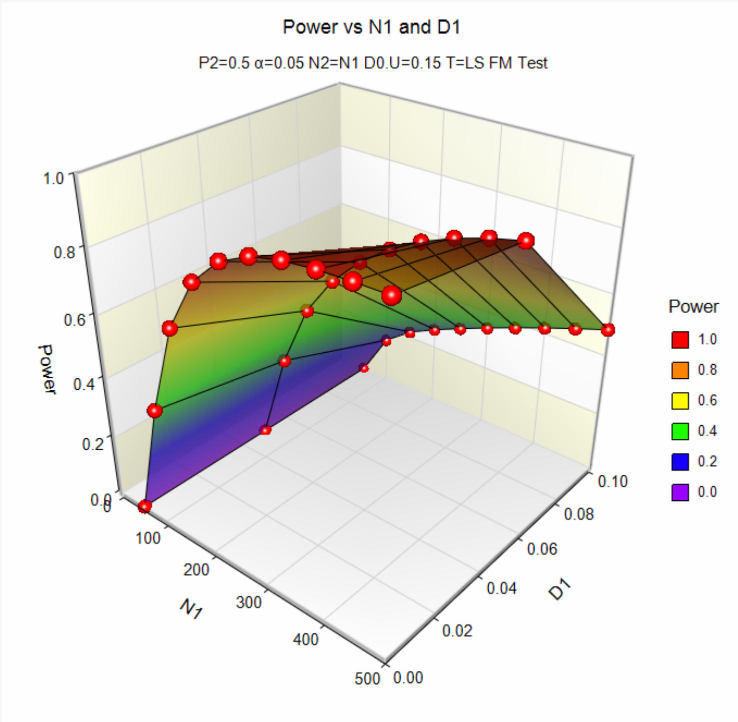
Blackwelder, W.C. 1998. 'Equivalence Trials.' In Encyclopedia of Biostatistics, John Wiley and Sons. New York. Volume 2, 1367-1372.
Chow, S.C. and Liu, J.P. 1999. Design and Analysis of Bioavailability and Bioequivalence Studies. Marcel Dekker. New York.
Chow, S.C., Shao, J., Wang, H., and Lokhnygina, Y. 2018. Sample Size Calculations in Clinical Research, 3rd Edition. Chapman & Hall/CRC. Boca Raton, FL. Pages 86-88.
Farrington, C. P. and Manning, G. 1990. 'Test Statistics and Sample Size Formulae for Comparative Binomial Trials with Null Hypothesis of Non-Zero Risk Difference or Non-Unity Relative Risk.' Statistics in Medicine, Vol. 9, pages 1447-1454.
Fleiss, J. L., Levin, B., Paik, M.C. 2003. Statistical Methods for Rates and Proportions. Third Edition. John Wiley & Sons. New York.
Gart, John J. and Nam, Jun-mo. 1988. 'Approximate Interval Estimation of the Ratio in Binomial Parameters: A Review and Corrections for Skewness.' Biometrics, Volume 44, Issue 2, 323-338.
Gart, John J. and Nam, Jun-mo. 1990. 'Approximate Interval Estimation of the Difference in Binomial Parameters: Correction for Skewness and Extension to Multiple Tables.' Biometrics, Volume 46, Issue 3, 637-643.
Julious, S. A. and Campbell, M. J. 2012. 'Tutorial in biostatistics: sample sizes for parallel group clinical trials with binary data.' Statistics in Medicine, 31:2904-2936.
Lachin, John M. 2000. Biostatistical Methods. John Wiley & Sons. New York.
Machin, D., Campbell, M., Fayers, P., and Pinol, A. 1997. Sample Size Tables for Clinical Studies, 2nd Edition. Blackwell Science. Malden, Mass.
Miettinen, O.S. and Nurminen, M. 1985. 'Comparative analysis of two rates.' Statistics in Medicine 4: 213-226.
Tubert-Bitter, P., Manfredi,R., Lellouch, J., Begaud, B. 2000. 'Sample size calculations for risk equivalence testing in pharmacoepidemiology.' Journal of Clinical Epidemiology 53, 1268-1274.

This report shows the values of each of the parameters, one scenario per row.

Plots Section

Plots





The values from the table are displayed in the above charts. These charts give a quick look at the sample size that will be required for various values of D1.

Example 2 – Finding the Sample Size

Continuing with the scenario given in Example 1, the researchers want to determine the sample size necessary for each value of D1 to achieve a power of 0.80.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 2** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab

Solve For .... **Sample Size**
Power Calculation Method .... **Normal Approximation**
Test Type.. **Likelihood Score (Farr. & Mann.)**
Power... **0.80**
Alpha.. **0.05**
Group Allocation .... **Equal (N1 = N2)**
Input Type... **Differences**
D0.U (Upper Equivalence Difference)............ **0.15**
D0.L (Lower Equivalence Difference) .... **-D0.U**
D1 (Actual Difference)................................... **0.00 0.05 0.10**
P2 (Group 2 Proportion)................................ **0.5**

Output

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

Numeric Results

Solve For: Sample Size
Groups: 1 = Treatment, 2 = Reference
Test Statistic: Farrington & Manning Likelihood Score Test
Hypotheses: H0: P1 - P2 ≤ D0.L or P1 - P2 ≥ D0.U vs. H1: D0.L < P1 - P2 < D0.U

Power		Sample Size			Proportions				Difference			
					Equivalence				Equivalence			
Target	Actual*	N1	N2	N	Lower P1.0L	Upper P1.0U	Actual P1.1	Reference P2	Lower D0.L	Upper D0.U	Actual D1	Alpha
0.8	0.8003	188	188	376	0.35	0.65	0.50	0.5	-0.15	0.15	0.00	0.05
0.8	0.8001	304	304	608	0.35	0.65	0.55	0.5	-0.15	0.15	0.05	0.05
0.8	0.8001	1202	1202	2404	0.35	0.65	0.60	0.5	-0.15	0.15	0.10	0.05

* Power was computed using the normal approximation method.

The required sample size will depend a great deal on the value of D1. Any effort spent determining an accurate value for D1 will be worthwhile.

Example 3 – Comparing the Power of Several Test Statistics

Continuing with Example 2, the researchers want to determine which of the eight possible test statistics to adopt by using the comparative reports and charts that **PASS** produces. They decide to compare the powers and actual alphas for various sample sizes between 50 and 200 when D1 is 0.1.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 3** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab

Solve For .... **Power**
Power Calculation Method .... **Binomial Enumeration**
Maximum N1 or N2 for Binomial Enumeration........... **5000**
Zero Count Adjustment Method .... **Add to zero cells only**
Zero Count Adjustment Value................................... **0.0001**
Test Type.. **Likelihood Score (Farr. & Mann.)**
Alpha.. **0.05**
Group Allocation .... **Equal (N1 = N2)**
Sample Size Per Group .... **50 to 200 by 50**
Input Type... **Differences**
D0.U (Upper Equivalence Difference)........................ **0.15**
D0.L (Lower Equivalence Difference) .... **-D0.U**
D1 (Actual Difference)... **0.10**
P2 (Group 2 Proportion).. **0.5**

Reports Tab

Show Comparative Reports .... **Checked**

Comparative Plots Tab

Show Comparative Plots... **Checked**

Output

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

Power Comparison of Eight Different Tests

Hypotheses: H0: P1 - P2 ≤ D0.L or P1 - P2 ≥ D0.U vs. H1: D0.L < P1 - P2 < D0.U

						Power							
Sample Size N1 | N2 |N	P2	D0.L	D0.U	D1	Target Alpha	Z(P) Test	Z(UnP) CC Test	Z(P) CC Test	Z(UnP) Test	T Score	F.M. Score	M.N. Score	G.N.
50 | 50 | 100	0.5	-0.15	0.15	0.1	0.05	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000	0.0000
100 | 100 | 200	0.5	-0.15	0.15	0.1	0.05	0.1494	0.1494	0.1047	0.1047	0.1493	0.1495	0.1494	0.1494
150 | 150 | 300	0.5	-0.15	0.15	0.1	0.05	0.2208	0.2208	0.1863	0.1863	0.2208	0.2208	0.2208	0.2208
200 | 200 | 400	0.5	-0.15	0.15	0.1	0.05	0.2552	0.2553	0.2238	0.2239	0.2551	0.2566	0.2566	0.2560

Note: Power was computed using binomial enumeration of all possible outcomes.

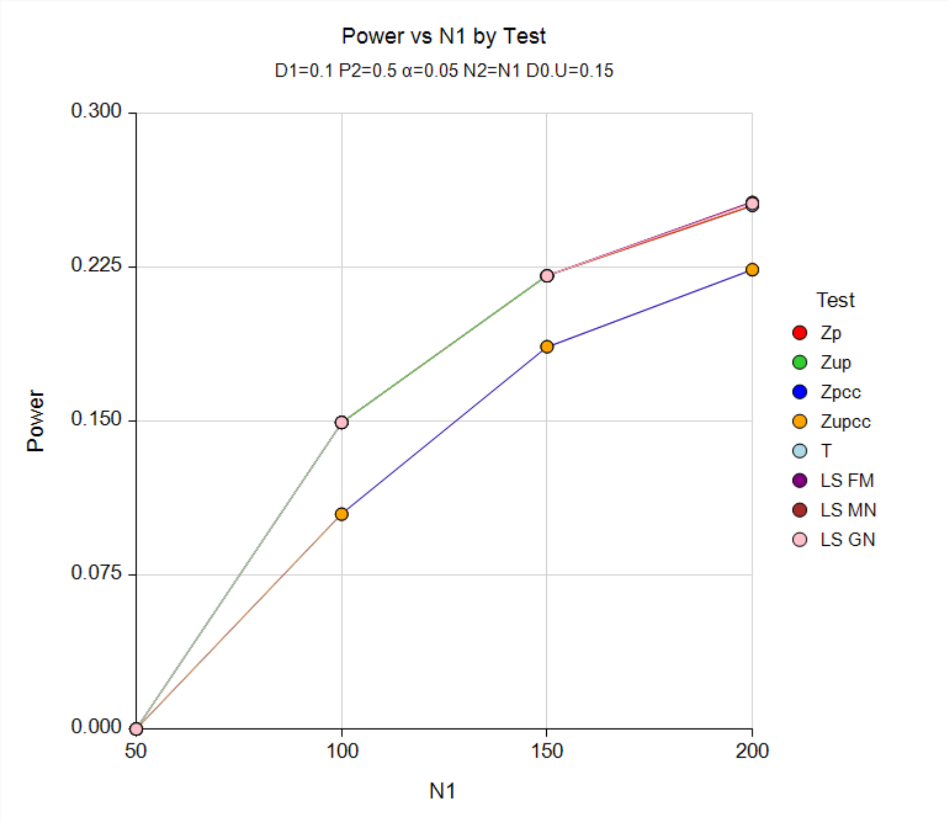
Actual Alpha Comparison of Eight Different Tests

Hypotheses: H0: P1 - P2 ≤ D0.L or P1 - P2 ≥ D0.U vs. H1: D0.L < P1 - P2 < D0.U

						Alpha							
Sample Size N1 | N2 | N	P2	D0.L	D0.U	D1	Target	Z(P) Test	Z(UnP) Test	Z(P) CC Test	Z(UnP) CC Test	T Test	F.M. Score	M.N. Score	G.N. Score
50 | 50 | 100	0.5	-0.15	0.15	0.1	0.05	0.0515	0.0515	0.0334	0.0334	0.0514	0.0515	0.0515	0.0515
100 | 100 | 200	0.5	-0.15	0.15	0.1	0.05	0.0486	0.0486	0.0358	0.0358	0.0485	0.0489	0.0487	0.0487
150 | 150 | 300	0.5	-0.15	0.15	0.1	0.05	0.0495	0.0495	0.0386	0.0386	0.0495	0.0495	0.0495	0.0495
200 | 200 | 400	0.5	-0.15	0.15	0.1	0.05	0.0465	0.0468	0.0376	0.0378	0.0464	0.0488	0.0488	0.0481

Note: Actual alpha was computed using binomial enumeration of all possible outcomes.

Plots



It is interesting to note that the powers of the continuity-corrected test statistics are consistently lower than the other tests. This occurs because the actual alpha achieved by these tests is lower than for the other tests. An interesting finding of this example is that the regular *t*-test performed about as well as the *z*-test.

Example 4 – Comparing Power Calculation Methods

Continuing with Example 3, let's see how the results compare if we were to use approximate power calculations instead of power calculations based on binomial enumeration.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 4** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab

Solve For .... **Power**
Power Calculation Method .... **Normal Approximation**
Test Type .... **Likelihood Score (Farr. & Mann.)**
Alpha... **0.05**
Group Allocation .... **Equal (N1 = N2)**
Sample Size Per Group .... **50 to 200 by 50**
Input Type .... **Differences**
D0.U (Upper Equivalence Difference)............ **0.15**
D0.L (Lower Equivalence Difference) .... **-D0.U**
D1 (Actual Difference)................................... **0.10**
P2 (Group 2 Proportion)................................ **0.5**

Reports Tab

Show Power Detail Report............................ **Checked**

Output

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

Power Detail Report

Test Statistic: Farrington & Manning Likelihood Score Test
Hypotheses: H0: P1 - P2 ≤ D0.L or P1 - P2 ≥ D0.U vs. H1: D0.L < P1 - P2 < D0.U

Sample Size							Normal Approximation		Binomial Enumeration	
N1	**N2**	**N**	**P2**	**D0.L**	**D0.U**	**D1**	**Power**	**Alpha**	**Power**	**Alpha**
50	50	100	0.5	-0.15	0.15	0.1	0.0000	0.05	0.0000	0.0515
100	100	200	0.5	-0.15	0.15	0.1	0.1523	0.05	0.1495	0.0489
150	150	300	0.5	-0.15	0.15	0.1	0.2206	0.05	0.2208	0.0495
200	200	400	0.5	-0.15	0.15	0.1	0.2659	0.05	0.2566	0.0488

Notice that the approximate power values are very close to the binomial enumeration values for all sample sizes.

Example 5 – Computing the Power after Completing an Experiment

Researchers are testing a generic drug to determine if it is equivalent to the name-brand alternative. Equivalence is declared if the success rate of the generic brand is no more than 5% from that of the name-brand drug. In a study with 1000 individuals in each group, they find that 774, or 77.4%, are successfully treated using the name-brand drug, and 700, or 70%, respond to the generic drug. An equivalence test (exact test) with alpha = 0.05 failed to declare that the two drugs are equivalent. The researchers would now like to compute the power for actual differences ranging from 0 to 4%. Suppose that the true value for the response rate for the name-brand drug is 77%.

Note that the power is not calculated at the difference observed in the study, 77.4%. In fact, the difference observed in the study is larger than the proposed equivalence difference, 5%. It would make no sense to perform a power calculation for a difference larger than the equivalence difference. It is more informative to study a range of values smaller than or equal to the equivalence difference.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 5** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab

Solve For .... **Power**
Power Calculation Method .... **Normal Approximation**
Test Type.. **Likelihood Score (Farr. & Mann.)**
Alpha.. **0.05**
Group Allocation .... **Equal (N1 = N2)**
Sample Size Per Group **1000**
Input Type.. **Differences**
D0.U (Upper Equivalence Difference)............ **0.05**
D0.L (Lower Equivalence Difference) .... **-D0.U**
D1 (Actual Difference)................................... **0.00 to 0.04 by 0.01**
P2 (Group 2 Proportion)................................ **0.77**

Output

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

Numeric Results

Solve For: Power
Groups: 1 = Treatment, 2 = Reference
Test Statistic: Farrington & Manning Likelihood Score Test
Hypotheses: H0: P1 - P2 ≤ D0.L or P1 - P2 ≥ D0.U vs. H1: D0.L < P1 - P2 < D0.U

				Proportions				Difference			
	Sample Size			Equivalence				Equivalence			
Power*	N1	N2	N	Lower P1.0L	Upper P1.0U	Actual P1.1	Reference P2	Lower D0.L	Upper D0.U	Actual D1	Alpha
0.6875	1000	1000	2000	0.72	0.82	0.77	0.77	-0.05	0.05	0.00	0.05
0.6313	1000	1000	2000	0.72	0.82	0.78	0.77	-0.05	0.05	0.01	0.05
0.4731	1000	1000	2000	0.72	0.82	0.79	0.77	-0.05	0.05	0.02	0.05
0.2857	1000	1000	2000	0.72	0.82	0.80	0.77	-0.05	0.05	0.03	0.05
0.1362	1000	1000	2000	0.72	0.82	0.81	0.77	-0.05	0.05	0.04	0.05

* Power was computed using the normal approximation method.

The power of the test ranges from 68.75% if the true difference is actually 0.0% to 13.62% if the true difference is 4%.

Example 6 – Finding the Sample Size using Proportions

A study is being designed to prove the equivalence of a new drug to the current standard. The current drug is effective in 85% of cases. The new drug, however, is cheaper to produce. The new drug will be deemed equivalent to the standard if its success rate is between 78% and 92%. What sample sizes are necessary to obtain 80% or 90% power for actual success rates ranging from 80% to 90%? The researchers will test at a significance level of 0.05 using the Farrington and Manning likelihood score test.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 6** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab

Solve For	**Sample Size**
Power Calculation Method	**Normal Approximation**
Test Type	**Likelihood Score (Farr. & Mann.)**
Power	**0.80 0.90**
Alpha	**0.05**
Group Allocation	**Equal (N1 = N2)**
Input Type	**Proportions**
P1.0U (Upper Equivalence Proportion)	**0.92**
P1.0L (Lower Equivalence Proportion)	**0.78**
P1.1 (Actual Proportion)	**0.80 to 0.90 by 0.02**
P2 (Group 2 Proportion)	**0.85**

Output

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

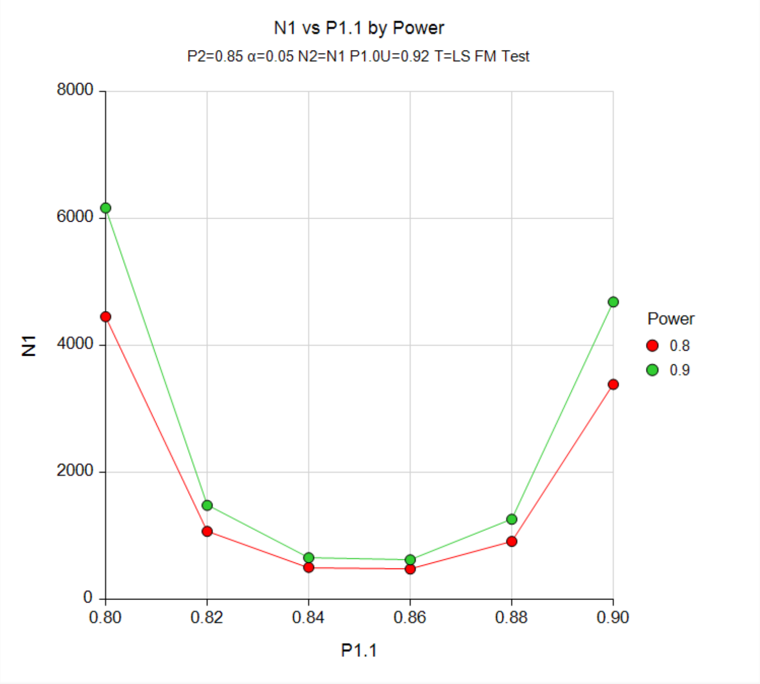
Numeric Results

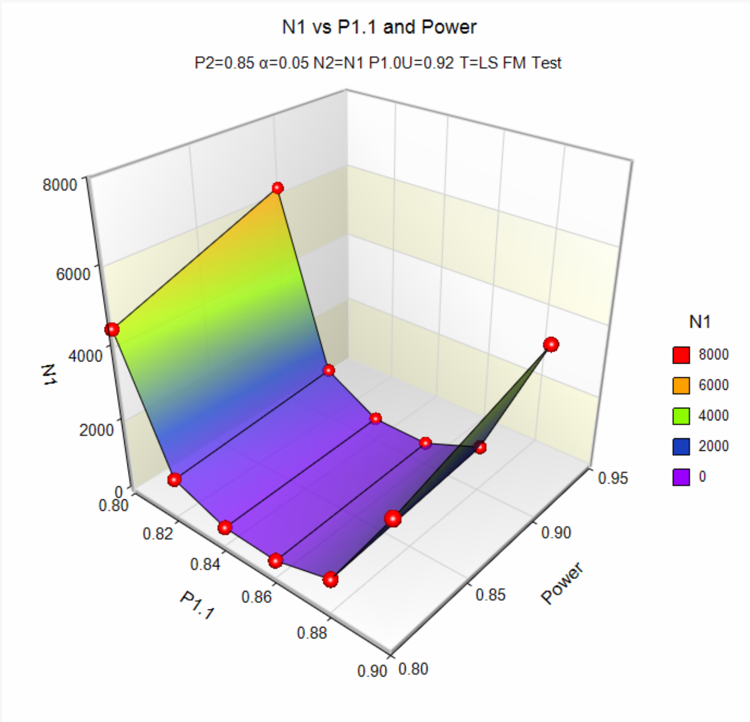
Solve For: Sample Size
Groups: 1 = Treatment, 2 = Reference
Test Statistic: Farrington & Manning Likelihood Score Test
Hypotheses: H0: P1 - P2 ≤ D0.L or P1 - P2 ≥ D0.U vs. H1: D0.L < P1 - P2 < D0.U

Power		Sample Size			Proportions: Equivalence Lower	Proportions: Equivalence Upper	Proportions	Proportions	Difference: Equivalence Lower	Difference: Equivalence Upper	Difference	
Target	Actual*	N1	N2	N	P1.0L	P1.0U	Actual P1.1	Reference P2	D0.L	D0.U	Actual D1	Alpha
0.8	0.8001	4453	4453	8906	0.78	0.92	0.80	0.85	-0.07	0.07	-0.05	0.05
0.9	0.9000	6166	6166	12332	0.78	0.92	0.80	0.85	-0.07	0.07	-0.05	0.05
0.8	0.8002	1070	1070	2140	0.78	0.92	0.82	0.85	-0.07	0.07	-0.03	0.05
0.9	0.9000	1480	1480	2960	0.78	0.92	0.82	0.85	-0.07	0.07	-0.03	0.05
0.8	0.8008	503	503	1006	0.78	0.92	0.84	0.85	-0.07	0.07	-0.01	0.05
0.9	0.9001	655	655	1310	0.78	0.92	0.84	0.85	-0.07	0.07	-0.01	0.05
0.8	0.8004	477	477	954	0.78	0.92	0.86	0.85	-0.07	0.07	0.01	0.05
0.9	0.9004	622	622	1244	0.78	0.92	0.86	0.85	-0.07	0.07	0.01	0.05
0.8	0.8002	912	912	1824	0.78	0.92	0.88	0.85	-0.07	0.07	0.03	0.05
0.9	0.9002	1261	1261	2522	0.78	0.92	0.88	0.85	-0.07	0.07	0.03	0.05
0.8	0.8000	3386	3386	6772	0.78	0.92	0.90	0.85	-0.07	0.07	0.05	0.05
0.9	0.9000	4685	4685	9370	0.78	0.92	0.90	0.85	-0.07	0.07	0.05	0.05

* Power was computed using the normal approximation method.

Plots





It is evident from these results that the sample sizes required to achieve 80% and 90% power depend a great deal on the actual value of the success rate, P1.1.

Example 7 – Validation of Sample Size Calculation for the Unpooled Z-Test using Julius and Campbell (2012)

Julius and Campbell (2012) presents Table XVI that gives the results of sample size calculations for an unpooled Z-test for equivalence with P2 between 0.7 and 0.9, |D0| between 0.05 and 0.20 and D1 between -0.05 and 0.05. Sample sizes are calculated for 90% power and alpha = 0.025. This example will replicate all values of D1 for P2 = 0.70 and |D0| = 0.20 in the table.

The sample sizes reported in the table for D1 between -0.05 and 0.05 are 205, 180, 161, 148, 140, 137, 138, 143, 152, 167, and 186.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 7** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab

Solve For	**Sample Size**
Power Calculation Method	**Normal Approximation**
Test Type	**Z-Test (Unpooled)**
Power	**0.90**
Alpha	**0.025**
Group Allocation	**Equal (N1 = N2)**
Input Type	**Differences**
D0.U (Upper Equivalence Difference)	**0.20**
D0.L (Lower Equivalence Difference)	**-D0.U**
D1 (Actual Difference)	**-0.05 to 0.05 by 0.01**
P2 (Group 2 Proportion)	**0.7**

Output

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

Numeric Results

Solve For: Sample Size
Groups: 1 = Treatment, 2 = Reference
Test Statistic: Z-Test with Unpooled Variance
Hypotheses: H0: P1 - P2 ≤ D0.L or P1 - P2 ≥ D0.U vs. H1: D0.L < P1 - P2 < D0.U

Power		Sample Size			Proportions				Difference			
					Equivalence				Equivalence			
Target	Actual*	N1	N2	N	Lower P1.0L	Upper P1.0U	Actual P1.1	Reference P2	Lower D0.L	Upper D0.U	Actual D1	Alpha
0.9	0.9007	205	205	410	0.5	0.9	0.65	0.7	-0.2	0.2	-0.05	0.025
0.9	0.9010	180	180	360	0.5	0.9	0.66	0.7	-0.2	0.2	-0.04	0.025
0.9	0.9010	161	161	322	0.5	0.9	0.67	0.7	-0.2	0.2	-0.03	0.025
0.9	0.9011	148	148	296	0.5	0.9	0.68	0.7	-0.2	0.2	-0.02	0.025
0.9	0.9006	140	140	280	0.5	0.9	0.69	0.7	-0.2	0.2	-0.01	0.025
0.9	0.9015	137	137	274	0.5	0.9	0.70	0.7	-0.2	0.2	0.00	0.025
0.9	0.9023	138	138	276	0.5	0.9	0.71	0.7	-0.2	0.2	0.01	0.025
0.9	0.9024	143	143	286	0.5	0.9	0.72	0.7	-0.2	0.2	0.02	0.025
0.9	0.9009	152	152	304	0.5	0.9	0.73	0.7	-0.2	0.2	0.03	0.025
0.9	0.9014	167	167	334	0.5	0.9	0.74	0.7	-0.2	0.2	0.04	0.025
0.9	0.9003	186	186	372	0.5	0.9	0.75	0.7	-0.2	0.2	0.05	0.025

* Power was computed using the normal approximation method.

The sample sizes from **PASS** match Table XVI of Julius and Campbell (2012) exactly. You can replicate the other values in the table by changing the values for P2 and D0.U.

Example 8 – Validation of Sample Size Calculation for the Pooled Z-Test using Tuber-Bitter et al. (2000)

Tuber-Bitter et al. (2000), page 1271, present a sample size study in which P2 = 0.1; D0.U = 0.01, 0.02, 0.03; D0.L = -D0.U; D1 = 0.0; alpha = 0.05; and beta = 0.1. Using the pooled Z test statistic, they found the sample sizes to be 19484, 4871, and 2165 in each group.

Setup

If the procedure window is not already open, use the PASS Home window to open it. The parameters for this example are listed below and are stored in the **Example 8** settings file. To load these settings to the procedure window, click **Open Example Settings File** in the Help Center or File menu.

Design Tab

Solve For	**Sample Size**
Power Calculation Method	**Normal Approximation**
Test Type	**Z-Test (Pooled)**
Power	**0.90**
Alpha	**0.05**
Group Allocation	**Equal (N1 = N2)**
Input Type	**Differences**
D0.U (Upper Equivalence Difference)	**0.01 0.02 0.03**
D0.L (Lower Equivalence Difference)	**-D0.U**
D1 (Actual Difference)	**0.0**
P2 (Group 2 Proportion)	**0.1**

Output

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

Numeric Results

Solve For: Sample Size
Groups: 1 = Treatment, 2 = Reference
Test Statistic: Z-Test with Pooled Variance
Hypotheses: H0: P1 - P2 ≤ D0.L or P1 - P2 ≥ D0.U vs. H1: D0.L < P1 - P2 < D0.U

Power		Sample Size			Proportions				Difference			
					Equivalence				Equivalence			
Target	Actual*	N1	N2	N	Lower P1.0L	Upper P1.0U	Actual P1.1	Reference P2	Lower D0.L	Upper D0.U	Actual D1	Alpha
0.9	0.9000	19480	19480	38960	0.09	0.11	0.1	0.1	-0.01	0.01	0	0.05
0.9	0.9000	4870	4870	9740	0.08	0.12	0.1	0.1	-0.02	0.02	0	0.05
0.9	0.9001	2165	2165	4330	0.07	0.13	0.1	0.1	-0.03	0.03	0	0.05

* Power was computed using the normal approximation method.

PASS found the required sample sizes to within rounding error of Tuber-Bitter et al. (2000).