Chapter 274

Tests for Two Independent Specificities

Introduction

This procedure gives power or required sample size for comparing two diagnostic tests when the outcome is specificity. In this design, the outcome of each of two diagnostic screening tests is compared to a gold standard.

Specifically, a set of *N* subjects is randomly divided into two groups. In each group, a portion of the subjects have the disease (condition of interest) and a portion does not. Each subject is given the one of the diagnostic tests. Subsequently, a gold standard test is used to obtain the true presence or absence of the disease. The gold standard may be a more expensive test, difficult to determine, or require the sacrifice of the subject.

The measures of diagnostic accuracy are sensitivity and specificity. *Sensitivity* (*Se*) is the probability that the diagnostic test is positive for the disease, given that the subject actually has the disease. *Specificity* (*Sp*) is the probability that the diagnostic test is negative, given that the subject does not have the disease. Mathematically,

$$\text{Sensitivity } (Se) = \Pr(+\text{Test}|\text{Disease})$$

$$\text{Specificity } (Sp) = \Pr(-\text{Test}|\text{No Disease})$$

Li and Fine (2004) present sample size methodology for testing sensitivity and specificity using a two-group, prospective design. Their methodology is used here. Other useful references are Obuchowski and Zhou (2002), Machin, Campbell, Tan, and Tan (2009), and Zhou, Obuchowski, and McClish (2002).

Prospective Study Design

In a two-group, prospective study, a group of N subjects is split into two groups: those that receive diagnostic test 1 (N_1 subjects) and those that receive diagnostic test 2 (N_2 subjects), such that $N = N_1 + N_2$. Each of these groups is divided further into those with the disease of interest and those without it. Suppose that the k^{th} group (k = 1 or 2) has N_{kD} with the disease and N_{kND} without the disease. A diagnostic test is administered to each subject (usually before the disease status is determined) and its output is recorded. The diagnostic test outcome is either positive or negative for the disease. Suppose that of the N_{kD} subjects with the disease, s_{k+} have a positive test outcome and s_{k-} have a negative outcome. Similarly, of the N_{kND} subjects without the disease, r_{k+} have positive outcomes and r_{k-} have negative outcomes.

The *sensitivity* in each group is estimated by

$$Se = \frac{s_{k+}}{N_{kD}},$$

and the *specificity* in each group is estimated by

$$Sp = \frac{r_{k-}}{N_{kND}}.$$

A useful diagnostic test has high values of both *Se* and *Sp*.

Comparing Two Specificities

When analyzing the data from studies such as this, one usually compares the two binomial specificities, Sp_1 and Sp_2. Note that these values are estimated solely using the $N_{ND} = N_{1ND} + N_{2ND}$ subjects without the disease. The data for the $N_D = N_{1D} + N_{2D}$ subjects with the disease are ignored. The data are displayed in a 2-by-2 contingency table as follows

	Test Outcome		
Group	**Positive**	**Negative**	**Total**
1	r_{1+}	r_{1-}	N_{1ND}
2	r_{2+}	r_{2-}	N_{2ND}

A popular test statistic for comparing the sensitivities is *Fisher's Exact Test* or the *Chi-square Test* with one degree of freedom.

Specificity Hypotheses

Conditional on the values of N_{kND}, r_{k-} is distributed as $\text{Binomial}(N_{kND}, Sp_k)$. Thus, tests of the two-sided statistical hypotheses,

$$H_0: Sp_1 = Sp_2 \quad \text{or} \quad H_0: Sp_1 - Sp_2 = 0$$

$$H_1: Sp_1 \neq Sp_2 \quad \text{or} \quad H_1: Sp_1 - Sp_2 \neq 0$$

can be carried out using any of the two-sample proportion tests (see chapter 200 for more details on two-sample proportion tests).

The upper one-sided null and alternative hypotheses are

$$H_0: Sp_1 \leq Sp_2 \quad \text{or} \quad H_0: Sp_1 - Sp_2 \leq 0$$

$$H_1: Sp_1 > Sp_2 \quad \text{or} \quad H_1: Sp_1 - Sp_2 > 0$$

The lower one-sided null and alternative hypotheses are

$$H_0: Sp_1 \geq Sp_2 \quad \text{or} \quad H_0: Sp_1 - Sp_2 \geq 0$$

$$H_1: Sp_1 < Sp_2 \quad \text{or} \quad H_1: Sp_1 - Sp_2 < 0$$

The power analysis of these tests follows the same pattern as other two-sample proportion tests, except that the disease prevalence in the two groups must be accounted for.

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 such a test.

1. Find the critical value (or values in the case of a two-sided test) using the standard normal distribution. The critical value, $z_{critical}$, is that value of *z* that leaves exactly the target value of alpha in the appropriate tail of the normal distribution. For example, for an upper-tailed test with a target alpha of 0.05, the critical value is 1.645.
2. Compute the value of the test statistic, z_t, for every combination of r_{1-} and r_{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 $z_t > z_{critical}$, the combination is in the rejection region. Call all combinations of r_{1-} and r_{2-} that lead to a rejection the set *A*.
4. Compute the power for given values of Sp_1 and Sp_2 as

$$1-\beta = \sum_{A} \binom{N_{1ND}}{r_{1-}} Sp_1^{r_{1-}}(1-Sp_1)^{N_{1ND}-r_{1-}} \binom{N_{2ND}}{r_{2-}} Sp_2^{r_{2-}}(1-Sp_2)^{N_{2ND}-r_{2-}}$$

5. Compute the actual value of alpha achieved by the design by substituting Sp_1 for Sp_2 in the above formula

$$\alpha^* = \sum_{A} \binom{N_{1ND}}{r_{1-}} Sp_1^{r_{1-}}(1-Sp_1)^{N_{1ND}-r_{1-}} \binom{N_{2ND}}{r_{2-}} Sp_1^{r_{2-}}(1-Sp_1)^{N_{2ND}-r_{2-}}$$

When the sample sizes are large (say over 200), these formulas may take a little time to evaluate. In this case, a large sample approximation may be used.

Calculating Group Sample Sizes Accounting for Disease Prevalence

To obtain the estimate of the group sample sizes, N_k, we inflate the non-diseased group sample sizes, N_{kND}, by one minus the disease prevalence, $1-P$ (the proportion of non-diseased subjects in the population), using the equation

$$N_k = \frac{N_{kND}}{(1-P)},$$

with fractional N_k values always being rounded up to the nearest whole number. This is called Method 0 in the paper by Li and Fine (2004). We can calculate N_{kND} from N_k by rearranging this formula as

$$N_{kND} = N_k \times (1-P),$$

with fractional N_{kND} values always being rounded down to the nearest whole number. The total sample size is calculated as

$$N = N_1 + N_2 \, .$$

Test Statistics

Various test statistics are available. The formulas for their power are given in Chapter 200 and they are not repeated here. The test statistics are

Fisher's Exact Test

The most useful reference we found for power analysis of Fisher's Exact test was in the StatXact 5 (2001) documentation. The material presented here is summarized from Section 26.3 (pages 866 – 870) of the StatXact-5 documentation. In this case, the test statistic is

$$T = -\ln\left[\frac{\binom{n_1}{x_1}\binom{n_2}{x_2}}{\binom{N}{m}}\right]$$

Chi-Square Test (Pooled and Unpooled)

This test statistic 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.

Both *pooled* and *unpooled* versions of this test have been discussed in the statistical literature. The pooling refers to the way in which the standard error is estimated. In the pooled version, the two proportions are averaged, and only one proportion is used to estimate the standard error. In the unpooled version, the two proportions are used separately.

The formula for the test statistic is

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

Pooled Version

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

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

Unpooled Version

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

Chi-Square Test with Continuity Correction

Frank Yates is credited with proposing a correction to the Pearson Chi-Square test for the lack of continuity in the binomial distribution. However, the correction was in common use when he proposed it in 1922.

Both *pooled* and *unpooled* versions of this test have been discussed in the statistical literature. The pooling refers to the way in which the standard error is estimated. In the pooled version, the two proportions are averaged, and only one proportion is used to estimate the standard error. In the unpooled version, the two proportions are used separately.

The continuity corrected z-test is

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

where *F* is -1 for lower-tailed, 1 for upper-tailed, and both -1 and 1 for two-sided hypotheses.

Pooled Version

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

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

Unpooled Version

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

Conditional Mantel Haenszel Test

The conditional Mantel Haenszel test, see Lachin (2000) page 40, is based on the *index frequency*, x_{11}, from the 2x2 table. The formula for the z-statistic is

$$z = \frac{x_{11} - E(x_{11})}{\sqrt{V_c(x_{11})}}$$

where

$$E(x_{11}) = \frac{n_1 m_1}{N}$$

$$V_c(x_{11}) = \frac{n_1 n_2 m_1 m_2}{N^2(N-1)}$$

Likelihood Ratio Test

In 1935, Wilks showed that the following quantity has a chi-square distribution with one degree of freedom. Using this test statistic to compare proportions is presented, among other places, in Upton (1982). The likelihood ratio test statistic is computed as

$$LR = 2\begin{bmatrix} a\ln(a) + b\ln(b) + c\ln(c) + d\ln(d) + \\ N\ln(N) - s\ln(s) - f\ln(f) - m\ln(m) - n\ln(n) \end{bmatrix}$$

Example 1 – Finding the Power

Suppose that diagnosing a certain disease has used a certain diagnostic test which has a specificity of 75%. A new diagnostic test has been developed that is much less expensive and invasive. Researchers want to design a prospective study to compare the old and new tests using a two-sided Z-Test with a significance level of 0.05.

They want to consider changes in specificity of 5% and 10%. These changes translate to specificities of 78.75% and 82.50%. The prevalence of the disease in the population of interest is 20%. The power will be determined for trials with sample sizes between 300 and 3000 incremented by 300.

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**
Alternative Hypothesis	**Two-Sided (H1: Sp1 ≠ Sp2)**
Test Type	**Z Test (Pooled)**
Alpha	**0.05**
Group Allocation	**Equal (N1 = N2)**
Sample Size Per Group	**300 to 3000 by 300**
P (Disease Prevalence)	**0.2**
Sp1 (Specificity of Group 1)	**0.75**
Sp2 (Specificity of Group 2)	**0.7875 0.8250**

Output

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

Numeric Reports

Numeric Results

Solve For: Power
Alternative Hypothesis: Two-Sided (H1: Sp1 ≠ Sp2)
Test Statistic: Z-Test with pooled variance

	Number of Subjects			Number of Non-Diseased Subjects			Disease Prevalence	Specificity			
Power*	N1	N2	N	$N1_{ND}$	$N2_{ND}$	N_{ND}	P	Sp1	Sp2	Sp1 - Sp2	Alpha
0.16356	300	300	600	240	240	480	0.2	0.75	0.7875	-0.0375	0.05
0.28047	600	600	1200	480	480	960	0.2	0.75	0.7875	-0.0375	0.05
0.39267	900	900	1800	720	720	1440	0.2	0.75	0.7875	-0.0375	0.05
0.49550	1200	1200	2400	960	960	1920	0.2	0.75	0.7875	-0.0375	0.05
0.58663	1500	1500	3000	1200	1200	2400	0.2	0.75	0.7875	-0.0375	0.05
0.66531	1800	1800	3600	1440	1440	2880	0.2	0.75	0.7875	-0.0375	0.05
0.73184	2100	2100	4200	1680	1680	3360	0.2	0.75	0.7875	-0.0375	0.05
0.78714	2400	2400	4800	1920	1920	3840	0.2	0.75	0.7875	-0.0375	0.05
0.83244	2700	2700	5400	2160	2160	4320	0.2	0.75	0.7875	-0.0375	0.05
0.86910	3000	3000	6000	2400	2400	4800	0.2	0.75	0.7875	-0.0375	0.05
0.51943	300	300	600	240	240	480	0.2	0.75	0.8250	-0.0750	0.05
0.81166	600	600	1200	480	480	960	0.2	0.75	0.8250	-0.0750	0.05
0.93638	900	900	1800	720	720	1440	0.2	0.75	0.8250	-0.0750	0.05
0.98056	1200	1200	2400	960	960	1920	0.2	0.75	0.8250	-0.0750	0.05
0.99448	1500	1500	3000	1200	1200	2400	0.2	0.75	0.8250	-0.0750	0.05
0.99852	1800	1800	3600	1440	1440	2880	0.2	0.75	0.8250	-0.0750	0.05
0.99962	2100	2100	4200	1680	1680	3360	0.2	0.75	0.8250	-0.0750	0.05
0.99991	2400	2400	4800	1920	1920	3840	0.2	0.75	0.8250	-0.0750	0.05
0.99998	2700	2700	5400	2160	2160	4320	0.2	0.75	0.8250	-0.0750	0.05
0.99999	3000	3000	6000	2400	2400	4800	0.2	0.75	0.8250	-0.0750	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	The number of subjects in Group 1 (Test 1). N1 = $N1_{ND}$ / (1 - P).
N2	The number of subjects in Group 2 (Test 2). N2 = $N2_{ND}$ / (1 - P).
N	The total number of subjects in the study. N = N1 + N2.
$N1_{ND}$	The number of non-diseased subjects in Group 1 (Test 1). $N1_{ND}$ = N1 × (1 - P).
$N2_{ND}$	The number of non-diseased subjects in Group 2 (Test 2). $N2_{ND}$ = N2 × (1 - P).
N_{ND}	The total number of non-diseased subjects in the study. N = $N1_{ND}$ + $N2_{ND}$.
P	Disease Prevalence. The proportion of individuals in the population that have the disease (or condition).
Sp1	The specificity of Group 1 (Test 1).
Sp2	The specificity of Group 2 (Test 2).
Sp1 - Sp2	The specificity difference to detect.
Alpha	The probability of rejecting a true null hypothesis.

Summary Statements

A parallel two-group diagnostic test design will be used to test whether the specificity of diagnostic test 1 is different from the specificity of diagnostic test 2 (H0: Specificity 1 = Specificity 2 versus H1: Specificity 1 ≠ Specificity 2). The comparison will be made using a two-sided Z-Test with pooled variance with a Type I error rate (α) of 0.05. The prevalence of the disease (or condition) in the population is assumed to be 0.2. To detect a specificity difference of -0.0375 (test 1 specificity: 0.75, test 2 specificity: 0.7875) with sample sizes of 300 for group 1 (test 1) and 300 for group 2 (test 2), the power is 0.16356.

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%	300	300	600	375	375	750	75	75	150
20%	600	600	1200	750	750	1500	150	150	300
20%	900	900	1800	1125	1125	2250	225	225	450
20%	1200	1200	2400	1500	1500	3000	300	300	600
20%	1500	1500	3000	1875	1875	3750	375	375	750
20%	1800	1800	3600	2250	2250	4500	450	450	900
20%	2100	2100	4200	2625	2625	5250	525	525	1050
20%	2400	2400	4800	3000	3000	6000	600	600	1200
20%	2700	2700	5400	3375	3375	6750	675	675	1350
20%	3000	3000	6000	3750	3750	7500	750	750	1500

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, 375 subjects should be enrolled in Group 1, and 375 in Group 2, to obtain final group sample sizes of 300 and 300, respectively.

References

Obuchowski, N.A., Zhou, X.H. 2002. 'Prospective studies of diagnostic test accuracy when disease prevalence is low,' Biostatistics, Volume 3, No. 4, pages 477-492.

Li, J., Fine, J. 2004. 'On sample size for sensitivity and specificity in prospective diagnostic accuracy studies,' Statistics in Medicine, Volume 23, pages 2537-2550.

Machin, D., Campbell, M.J., Tan, S.B., Tan, S.H. 2009. Sample Size Tables for Clinical Studies, Third Edition. Wiley-Blackwell, Chichester, United Kingdom.

Zhou, X.H., Obuchowski, N.A., McClish, D.K. 2002. Statistical Methods in Diagnostic Medicine. Wiley-Interscience, New York.

Chow, S.C., Shao, J., Wang, H. 2003. Sample Size Calculations in Clinical Research. Marcel Dekker. New York.

D'Agostino, R.B., Chase, W., Belanger, A. 1988.'The Appropriateness of Some Common Procedures for Testing the Equality of Two Independent Binomial Populations', The American Statistician, August 1988, Volume 42 Number 3, pages 198-202.

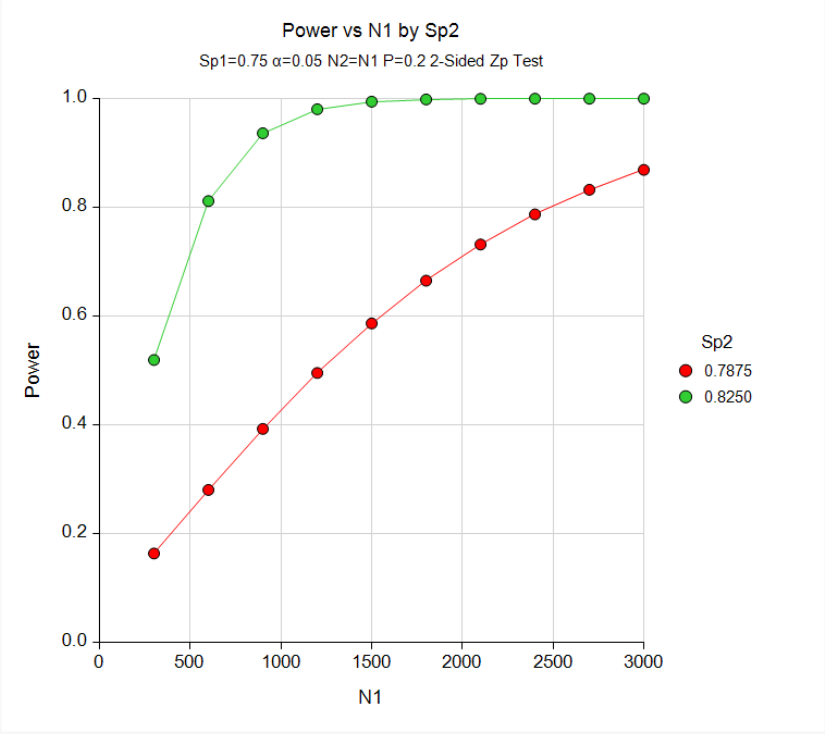
Fleiss, J. L., Levin, B., Paik, M.C. 2003. Statistical Methods for Rates and Proportions. Third Edition. John Wiley & Sons. New York.

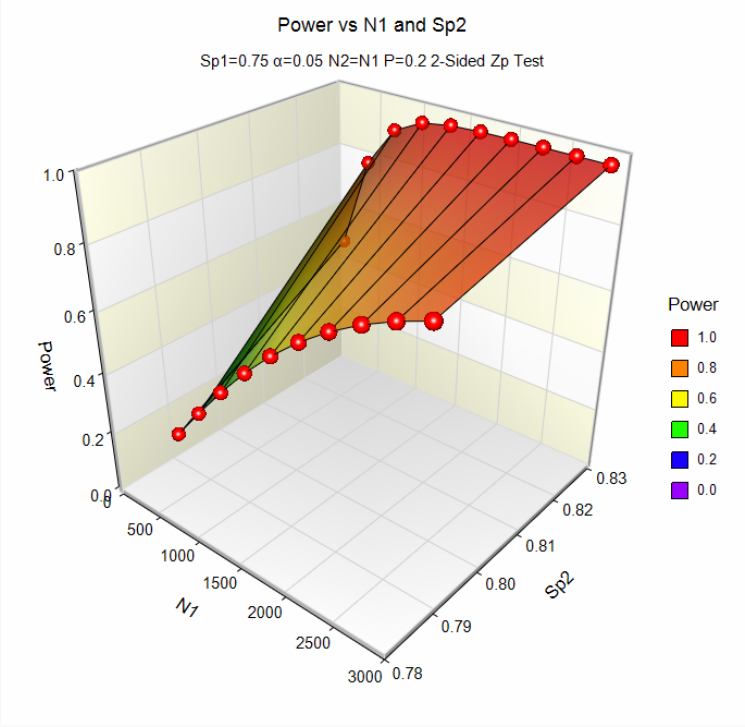
Lachin, John M. 2000. Biostatistical Methods. John Wiley & Sons. New York.

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

Plots Section

Plots





These plots show the relationship between power, sample size, and *Sp2* in this example.

Example 2 – Finding the Sample Size

Continuing with Example 1, suppose you want to study the impact of various choices for Sp2 on sample size. Using a significance level of 0.05 and 90% power, find the sample size when Sp2 is 78.75% and 82.50%. Assume a two-tailed test is used.

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**
Alternative Hypothesis	**Two-Sided (H1: Sp1 ≠ Sp2)**
Test Type	**Z Test (Pooled)**
Power	**0.9**
Alpha	**0.05**
Group Allocation	**Equal (N1 = N2)**
P (Disease Prevalence)	**0.2**
Sp1 (Specificity of Group 1)	**0.75**
Sp2 (Specificity of Group 2)	**0.7875 0.8250**

Output

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

Numeric Results

Solve For: Sample Size
Alternative Hypothesis: Two-Sided (H1: Sp1 ≠ Sp2)
Test Statistic: Z-Test with pooled variance

	Number of Subjects			Number of Non-Diseased Subjects			Disease Prevalence	Specificity			
Power*	N1	N2	N	$N1_{ND}$	$N2_{ND}$	N_{ND}	P	Sp1	Sp2	Sp1 - Sp2	Alpha
0.90005	3319	3319	6638	2655	2655	5310	0.2	0.75	0.7875	-0.0375	0.05
0.90041	780	780	1560	624	624	1248	0.2	0.75	0.8250	-0.0750	0.05

* Power was computed using the normal approximation method.

This report shows the sample size needed to achieve 90% power for each value of *Sp2*.

Example 3 – Validation using Machin et al. (2009)

Machin et al. (2009) page 166-167 give the results of a sample size determination in which Se1 = 0.27, Se2 = 0.66, P = 0.25, alpha = 0.05 (two-sided), and power = 0.80. The resulting sample size is 98 per group or 196 total.

The power calculation formulas for specificity are the same as those for sensitivity, except that 1 – P is used to adjust the sample size instead of P. Therefore, we will match the results in Machin et al. (2009) when entering specificity in place of sensitivity if we enter 1 – P (0.75) for the disease prevalence instead of P (0.25).

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	**Sample Size**
Power Calculation Method	**Binomial Enumeration**
Maximum N1ND or N2ND for Binomial Enumeration	**5000**
Zero Count Adjustment Method	**Add to zero cells only**
Zero Count Adjustment Value	**0.0001**
Alternative Hypothesis	**Two-Sided (H1: Sp1 ≠ Sp2)**
Test Type	**Z Test (Pooled)**
Power	**0.8**
Alpha	**0.05**
Group Allocation	**Equal (N1 = N2)**
P (Disease Prevalence)	**0.75**
Sp1 (Specificity of Group 1)	**0.27**
Sp2 (Specificity of Group 2)	**0.66**

Output

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

Numeric Results

Solve For: Sample Size
Alternative Hypothesis: Two-Sided (H1: Sp1 ≠ Sp2)
Test Statistic: Z-Test with pooled variance

	Number of Subjects			Number of Non-Diseased Subjects			Disease Prevalence	Specificity			Alpha†	
Power*	N1	N2	N	$N1_{ND}$	$N2_{ND}$	N_{ND}	P	Sp1	Sp2	Sp1 - Sp2	Target	Actual*
0.81699	96	96	192	24	24	48	0.75	0.27	0.66	-0.39	0.05	0.05203

* Power and actual alpha were computed using binomial enumeration of all possible outcomes.

† Warning: When solving for sample size with power computed using binomial enumeration, the target alpha level is not guaranteed. Actual alpha may be greater than target alpha in some cases. We suggest that you investigate sample sizes near the solution to find designs with an actual alpha you are willing to tolerate.

PASS has also obtained an *N* of 192 which is slightly different from 196 in the book. This difference is likely due to the exact binomial power calculation used by **PASS**.