Paper SAS1911-2015

Equivalence and Noninferiority Testing Using SAS/STAT® Software

John Castelloe and Donna Watts, SAS Institute Inc.

ABSTRACT

Proving difference is the point of most statistical testing. In contrast, the point of equivalence and noninferiority tests is to prove that results are substantially the same, or at least not appreciably worse. An equivalence test can show that a new treatment, one that is less expensive or causes fewer side effects, can replace a standard treatment. A noninferiority test can show that a faster manufacturing process creates no more product defects or industrial waste than the standard process. This paper reviews familiar and new methods for planning and analyzing equivalence and noninferiority studies in the POWER, TTEST, and FREQ procedures in SAS/STAT® software.

Techniques that are discussed range from Schuirmann's classic method of two one-sided tests (TOST) for demonstrating similar normal or lognormal means in bioequivalence studies, to Farrington and Manning's noninferiority score test for showing that an incidence rate (such as a rate of mortality, side effects, or product defects) is no worse. Real-world examples from clinical trials, marketing, and industrial process design are included.

PROLOGUE

You are a consulting statistician at a pharmaceutical company, charged with designing a study of your company's new arthritis drug, SASGoBowlFor'Em (abbreviated as "Bowl"). Your boss realizes that Bowl is unlikely to demonstrate better efficacy than the gold standard, Armanaleg, but its lower cost will make it an attractive alternative for consumers as long as you can show that the efficacy is about the same.

Your boss communicates the following study plans to you:

- The outcome to be measured is a "relief score," which ranges from 0 to 20 and is assumed to be approximately normally distributed.
- Subjects are to be allocated to Armanaleg and Bowl at a ratio of 2 to 3, respectively.
- The relief score is to be assessed after four weeks on the treatment.
- Bowl is expected to be slightly less effective than Armanaleg, with a mean relief score of 9.5 compared to 10 for Armanaleg.
- The minimally acceptable decrease in relief score is considered to be 2 units, corresponding to a 20% decrease, assuming a mean relief score of 10 for Armanaleg.
- The standard deviation of the relief score is expected to be approximately 2.25 for each treatment. Common standard deviation will be assumed in the data analysis.
- The sample size should be sufficient to produce an 85% chance of a significant result—that is, a power of 0.85—at a 0.05 significance level.

You recognize that a typical hypothesis test is inappropriate here because you are trying to demonstrate similarity rather than difference. A *noninferiority* test or an *equivalence* test is the way to go, but which is the better choice? You realize that because you're interested in only one direction—Bowl scoring better than some "not substantially worse" threshold compared to Armanaleg—a noninferiority test will be both more aligned with the study goals and more powerful.

Because of the normality and equal-variance assumptions, the classic pooled t test is a natural choice for the data analysis. But it won't be classic in terms of the hypotheses: you will need to incorporate the aforementioned "not substantially worse" threshold, also called the *noninferiority margin*. This margin is 2 units, the minimally acceptable decrease in relief score, because your boss wants to be able to announce with confidence at the conclusion of the study that the efficacy of Bowl is within 20% (2 units, given the mean assumptions) of Armanaleg's. In particular, he wants an 85% chance (the power) of being able to make this announcement with 95% confidence (one minus the significance level)—in other words, asserting a mere 5% chance that he's wrong.

So your hypotheses are

$$H_0{:}\mu_B - \mu_A \le -2$$
$$H_1{:}\mu_B - \mu_A > -2$$

where μ_B and μ_A are the mean relief scores for Bowl and Armanaleg, respectively.

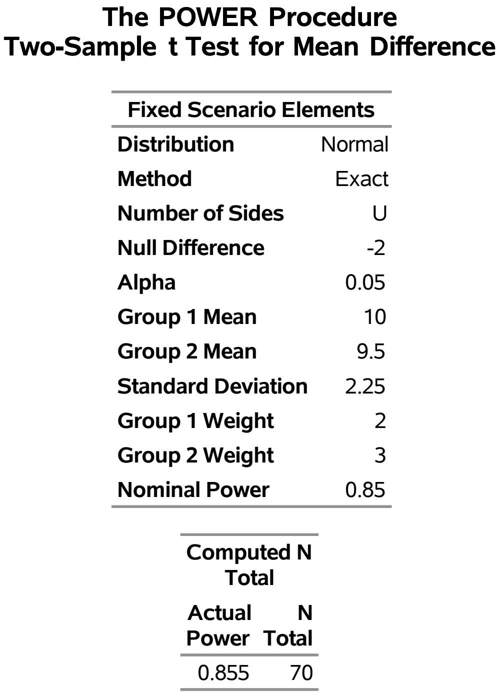
You use the following statements to determine the required sample size:

```
proc power;
   twosamplemeans
      sides = u
      groupweights = 2 | 3
      groupmeans = 10 | 9.5
      nulldiff = -2
      stddev = 2.25
      power = 0.85
      alpha = 0.05
      ntotal = .
   ;
run;
```

The TWOSAMPLEMEANS statement in PROC POWER doesn't have an explicit option to represent the noninferiority margin, but you can use the NULLDIFF= option. (For more information about using null value options to represent noninferiority margins, see the section "Data Analysis for Normal and Lognormal Means" on page 5.)

The results in Figure 1 suggest a sample size of 70.

Figure 1 Sample Size Determination for Arthritis Study



```
                 The POWER Procedure
        Two-Sample t Test for Mean Difference

            Fixed Scenario Elements
            Distribution              Normal
            Method                     Exact
            Number of Sides                U
            Null Difference               -2
            Alpha                       0.05
            Group 1 Mean                  10
            Group 2 Mean                 9.5
            Standard Deviation          2.25
            Group 1 Weight                 2
            Group 2 Weight                 3
            Nominal Power               0.85

                  Computed N
                     Total

                 Actual      N
                 Power   Total
                 0.855      70
```

Your boss is able to get funding for a study with 70 patients. After the study ends, he hands you the data and asks you to perform the noninferiority test. You use the following DATA step to create a SAS® data set:

```
data ArthritisRelief;
   Treatment = "Armanaleg";
   do i = 1 to 28; input Relief @@; output; end;
   Treatment = "Bowl     ";
   do i = 1 to 42; input Relief @@; output; end;
   drop i;
datalines;
 9 14 13  8 10  5 11  9 12 10  9 11  8 11
 4  8 11 16 12 10  9 10 13 12 11 13  9  4
 7 14  8  4 10 11  7  7 13  8  8 13 10  9
12  9 11 10 12  7  8  5 10  7 13 12 13 11
 7 12 10 11 10  8  6  9 11  8  5 11 10  8
;
```

You use the following statements to perform the noninferiority test:

```
proc ttest data=ArthritisRelief sides=l h0=2;
   class Treatment;
   var Relief;
run;
```

Like the TWOSAMPLEMEANS statement in PROC POWER, the TTEST procedure doesn't have an explicit option to represent the noninferiority margin, but you can use the H0= option in the PROC TTEST statement instead.

The results in Figure 2 show a significant result, $p = 0.0192$, for the pooled *t* test. This suggests, as you'd hoped, that the efficacy of Bowl is not appreciably worse than that of Armanaleg—that is, the mean relief score for Bowl is at most 2 units less than that for Armanaleg.

Figure 2 Noninferiority Test for Arthritis Study

The TTEST Procedure

Variable: Relief

Treatment	N	Mean	Std Dev	Std Err	Minimum	Maximum
Armanaleg	28	10.0714	2.7879	0.5269	4.0000	16.0000
Bowl	42	9.4048	2.4501	0.3781	4.0000	14.0000
Diff (1-2)		0.6667	2.5895	0.6318		

						95%	
Treatment	**Method**	**Mean**	**95% CL Mean**		**Std Dev**	**CL Std Dev**	
Armanaleg		10.0714	8.9904	11.1525	2.7879	2.2042	3.7947
Bowl		9.4048	8.6413	10.1683	2.4501	2.0159	3.1243
Diff (1-2)	**Pooled**	0.6667	-Infty	1.7202	2.5895	2.2180	3.1117
Diff (1-2)	**Satterthwaite**	0.6667	-Infty	1.7524			

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	68	-2.11	0.0192
Satterthwaite	Unequal	52.752	-2.06	0.0224

INTRODUCTION

Equivalence and noninferiority tests are useful in many different industries. In drug testing, for example, you can show that a generic alternative—one that is less expensive or causes fewer side effects than a popular name-brand drug—is similar in efficacy or mortality to the better-known drug. As a consulting statistician in the Prologue, you designed and analyzed such a study. In manufacturing, you can show that a faster manufacturing process creates no more product defects or industrial waste than the standard process.

The usual scenario in hypothesis testing is demonstration of a *difference* (between two treatments or processes, or between a treatment or process and a benchmark). For example, if you are estimating a parameter θ (such as a mean or proportion difference or ratio), the hypotheses for a typical two-sided test are

$$H_0\colon \theta = \theta_0$$
$$H_1\colon \theta \neq \theta_0$$

where θ_0 is the null value.

The hypotheses for a typical "upper" one-sided test are

$$H_0\colon \theta \leq \theta_0$$
$$H_1\colon \theta > \theta_0$$

and for a typical "lower" two-sided test are

$$H_0\colon \theta \geq \theta_0$$
$$H_1\colon \theta < \theta_0$$

One alternative testing scenario is the *equivalence test*, which aims to demonstrate *similar* results (efficacy, mortality rate, yield, and so on) when an advantage lies elsewhere, such as lower cost, fewer side effects, or a faster process. For an equivalence test, you specify "equivalence limits" (θ_L, θ_U) to characterize a range of values for θ that you consider to be acceptable. In other words, you would consider an observed difference at one of the equivalence limits to be minimally important.

The hypotheses for an equivalence test have the form

$$H_0\colon \theta \leq \theta_L \quad \text{or} \quad \theta \geq \theta_U$$
$$H_1\colon \theta_L < \theta < \theta_U$$

where θ_L and θ_U are the equivalence limits. If the equivalence limits are symmetric with respect to a particular value (for example, 0 for a difference parameter or 1 for a ratio parameter), then you can express the equivalence limits in terms of a "margin" (δ). If θ represents a difference parameter, then the equivalence limits in terms of a margin are $(-\delta, \delta)$. If θ represents a ratio parameter, then the equivalence limits in terms of a margin are $(1/\delta, \delta)$.

Three main varieties of equivalence are discussed in the application area of bioequivalence: average, population, and individual. The scope of this paper is limited to average bioequivalence.

Another alternative testing scenario is the *noninferiority* test, which aims to demonstrate that results are *not appreciably worse*. For a noninferiority test, you specify a noninferiority margin (δ) to characterize the largest absolute difference that you consider to be dismissible. If larger values of θ are better, then you construct the hypotheses for a noninferiority test as

$$H_0\colon \theta \leq \theta_0 - \delta$$
$$H_1\colon \theta > \theta_0 - \delta$$

where δ is a positive-valued margin. If smaller values of θ are better, then you use the following hypotheses:

$$H_0\colon \theta \geq \theta_0 + \delta$$
$$H_1\colon \theta < \theta_0 + \delta$$

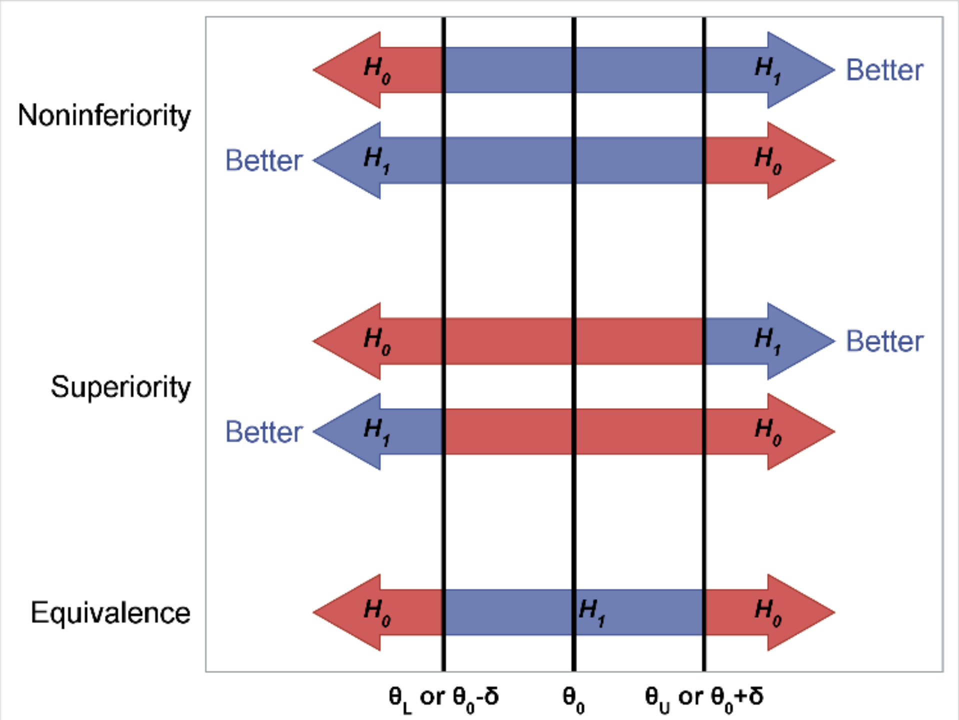
There is often confusion about the roles of the null value and noninferiority margin. The null value usually represents an important threshold, such as a minimally clinically meaningful difference or a cost-benefit breakpoint, the main focus of the hypothesis test. The noninferiority margin is more of a "fuzz factor," a bit of wiggle room to allow for a trivial difference in the wrong direction. However, in some cases the null value is set to zero difference by convention, and the margin then subsumes the usual role of the null value.

For equivalence analyses, you can construct an "equivalence interval" that contains all possible values of the parameter of interest that would result in rejecting the null hypothesis in favor of equivalence. Likewise, you can conduct a noninferiority test by checking whether a one-sided confidence interval for the parameter of interest lies entirely above the noninferiority limit of $\theta_0 - \delta$ (if bigger is better) or below the noninferiority limit of $\theta_0 + \delta$ (if smaller is better).

There is also a testing scenario called *superiority testing* that is similar to noninferiority testing except that the goal is to demonstrate that results are *appreciably better*. Superiority tests are not discussed in this paper because their mechanics are so similar to those of noninferiority tests; the only difference is that you add the margin to the null value if bigger is better and subtract it if smaller is better.

Figure 3 shows a visual summary of the hypotheses involved in noninferiority, superiority, and equivalence tests.

Figure 3 Hypotheses in Noninferiority, Superiority, and Equivalence Tests



POPULAR METHODS

Here are some popular methods of testing equivalence and noninferiority that are discussed in this paper:

- Schuirmann's two one-sided tests (TOST) method (Schuirmann 1987) for equivalence of normal or lognormal means based on *t* tests
- classic one-sided *t* test with a margin added to the null value, for noninferiority of normal or lognormal means
- Farrington-Manning score tests (Farrington and Manning 1990, Miettinen and Nurminen 1985) for equivalence or noninferiority of risk difference or relative risk for two independent samples
- assessment of noninferiority by comparing confidence limits for a risk difference or relative risk to a noninferiority margin (Schuirmann 1999, Dann and Koch 2008)

NORMAL AND LOGNORMAL MEANS

Data Analysis for Normal and Lognormal Means

For equivalence and noninferiority analyses that involve normal or lognormal data, you can use PROC TTEST to compute *p*-values and confidence intervals for a variety of designs and criteria for means, mean differences, and mean ratios.

Table 1 shows the statements and options in PROC TTEST that correspond to the supported designs for hypothesis testing and confidence intervals.

Table 1 Designs for Normal and Lognormal Means in PROC TTEST

Design	**Syntax**
One-sample	VAR *variables*
Two-sample	CLASS *variable*, VAR *variables*
Paired-sample	PAIRED *pair-lists*
AB/BA crossover	VAR *variables* / CROSSOVER=(*variable1 variable2*)

For a two-sample analysis or an AB/BA crossover analysis that includes a period effect (the default, unless you specify the IGNOREPERIOD option in the VAR statement), the results include both pooled and Satterthwaite versions of the tests and confidence intervals.

Table 2 shows the options in the PROC TTEST statement that correspond to the supported distributional assumptions and parameters of interest.

Table 2 Parameters of Interest for Normal and Lognormal Means

Parameter	**PROC TTEST Statement Options**
Normal mean or mean difference	(default)
Lognormal mean or mean ratio	DIST=LOGNORMAL
Normal mean ratio	TEST=RATIO DIST=NORMAL

The means in a DIST=LOGNORMAL analysis are geometric means rather than arithmetic means.

Table 3 shows the options that you can use in the PROC TTEST statement to specify equivalence or noninferiority criteria.

Table 3 Criteria for Normal and Lognormal Means

Criterion	**PROC TTEST Statement Options**
Equivalence	TOST(<*lower*,> *upper*)
Noninferiority	SIDES=L|U H0=*m*

PROC TTEST has no explicit options to specify a noninferiority test or margin (such as the NONINF and MARGIN= options in PROC FREQ). Instead, you should specify the noninferiority margin by using the H0= option. If you have both a custom null value *and* a noninferiority margin, then you need to either add them together (for DIST=NORMAL analyses) or multiply them together (for DIST=LOGNORMAL analyses) to yield the correct value for the H0= option.

Tests for all possible combinations of the rows of Table 1 through Table 3 are supported. Confidence limits are available for all possible combinations except the normal mean ratio either for paired-sample designs or for AB/BA crossover designs where the period effect is ignored.

For an equivalence analysis, a confidence interval that corresponds to the TOST equivalence test is usually called an "equivalence interval." Its confidence level turns out to be $100(1-2\alpha)$% instead of the $100(1-\alpha)$% that you might expect, because it's constructed as the overlap of two one-sided $100(1-\alpha)$% confidence intervals, one corresponding to each of the "two one-sided tests." (If each of the two tests is performed with significance level α, then the overall significance level is also α.) The equivalence interval is the same as the rejection region for the level-α TOST significance test.

For a noninferiority analysis, you can compare the usual $100(1-\alpha)$% confidence interval to the noninferiority limit. This interval falls entirely above the limit when bigger is better, or entirely below the limit when smaller is better, if and only if the null hypothesis is rejected in favor of noninferiority.

Table 4 shows hypothesis tests and corresponding confidence interval interpretations for some typical examples of equivalence and noninferiority analyses that involve normal or lognormal means.

Table 4 Examples of Hypotheses and Confidence Limits for Normal and Lognormal Means in PROC TTEST

Testing Scenario	Hypotheses	Confidence Limits	Syntax
Noninferiority of product yield in paired design with margin of 15 g, assuming normality	$H_0\colon \mu_1 - \mu_2 \le -15$ g, $H_1\colon \mu_1 - \mu_2 > -15$ g	Reject H_0 if lower $100(1-\alpha)$% confidence limit is at least –15 g	PROC TTEST SIDES=U H0=-15; PAIRED Yield1*Yield2;
Noninferiority of industrial waste in two-sample design with margin of 4.5 kg, assuming normality	$H_0\colon \mu_1 - \mu_2 \ge 4.5$ kg, $H_1\colon \mu_1 - \mu_2 < 4.5$ kg	Reject H_0 if upper $100(1-\alpha)$% confidence limit is at most 4.5 kg	PROC TTEST SIDES=L H0=4.5; CLASS Catalyst; VAR Waste;
(4/5, 5/4) lognormal ratio equivalence in AB/BA crossover design including period effect	$H_0\colon \gamma_1/\gamma_2 \le 0.8$ or $\gamma_1/\gamma_2 \ge 1.25$ $H_1\colon 0.8 < \gamma_1/\gamma_2 < 1.25$	Reject H_0 if $100(1-2\alpha)$% confidence interval for γ_1/γ_2 falls entirely within $[0.8, 1.25]$	PROC TTEST DIST=LOGNORMAL TOST(0.8, 1.25); VAR AUC1 AUC2; CROSSOVER=(Trt1 Trt2)

Power and Sample Size Analysis for Normal and Lognormal Means

You can use PROC POWER to compute power or sample size when planning studies to be analyzed using the methods in Table 1 through Table 3. PROC POWER supports all combinations of the rows in those tables except the following:

- tests that involve the normal mean ratio in paired-sample designs
- Satterthwaite versions of equivalence tests for two-sample designs and for the AB/BA crossover design that includes a period effect

Table 5 shows the statements in PROC POWER that correspond to the different designs.

Table 5 Designs in Power for Normal and Lognormal Means in PROC POWER

Design	Analysis Statement
One-sample	ONESAMPLEMEANS
Two-sample	TWOSAMPLEMEANS
Paired-sample	PAIREDMEANS
AB/BA crossover, including period effect	TWOSAMPLEMEANS
AB/BA crossover, ignoring period effect	PAIREDMEANS

There is no explicit statement for the AB/BA crossover design in PROC POWER. But because the underlying analysis for this design is either a two-sample or paired *t* test (depending on whether or not you include a period effect), you can use either the TWOSAMPLEMEANS or PAIREDMEANS statement. (For more information about power analyses for AB/BA crossover designs, see the section "Power and Sample Size Analysis for the AB/BA Crossover Design" on page 8.)

Table 6 shows the options that you can use in any of the statements in Table 5 to specify the distributional assumption and parameter of interest.

Table 6 Parameters of Interest in Power for Normal and Lognormal Means in PROC POWER

Parameter	Analysis Statement Option
Normal mean or mean difference	<default>
Lognormal mean or mean ratio	DIST=LOGNORMAL

Table 7 shows the options that you can use in the relevant statement in Table 5 to specify equivalence or noninferiority criteria.

Table 7 Criteria in Power for Normal and Lognormal Means in PROC POWER

Criterion	Analysis Statement Options

Equivalence	TEST=EQUIV|EQUIV_DIFF|EQUIV_RATIO
	LOWER=*number* UPPER=*number*

Noninferiority	TEST=DIFF|DIFF_SATT|RATIO SIDES=1|U|L NULLMEAN|NULLDIFF|NULLRATIO=*number*

You use the TEST=EQUIV and NULLMEAN=*number* options with a one-sample design; the TEST=EQUIV_DIFF and NULLDIFF=*number* options for a test of normal difference with a two-sample, paired, or AB/BA crossover design; and the TEST=EQUIV_RATIO and NULLRATIO=*number* options for a test of lognormal ratio with a two-sample, paired, or AB/BA crossover design.

As with the noninferiority tests in PROC TTEST, there are no explicit options to specify a noninferiority test or margin in PROC POWER. You should specify the noninferiority margin by using the appropriate null option for the design and parameter of interest.

Note that PROC POWER, compared to PROC TTEST, swaps “group 1” and “group 2” in the definitions of mean difference and ratio. In PROC POWER, a difference is for group 2 minus group 1, and a ratio is for group 2 over group 1.

Power and Sample Size Analysis for the AB/BA Crossover Design

If you ignore the period effect in an AB/BA crossover design, the power analysis for an equivalence or noninferiority test of the treatment effect is comparatively simple. Such a test is merely a paired *t* test on all the (treatment A, treatment B) response value pairs, regardless of treatment sequence. Thus, you can simply ignore the treatment sequence and proceed as if it’s a paired design.

However, if you include a period effect, the power analysis is more complicated. The treatment effect test in this case is a two-sample *t* test on either the halved period differences (for a test of normal mean difference) or the square roots of the period ratios (for a test of lognormal mean ratio), where the two “groups” are the two treatment sequences (AB and BA). The period difference or ratio is the difference or ratio, respectively, between the period 1 and period 2 response values. The normal mean difference that is estimated in such a test is simply the difference of treatment means,

$$\mu_{\mathrm{diff}} = \mu_A - \mu_B$$

But the standard deviation (for a test of normal mean difference) or coefficient of variation (for a test of lognormal mean ratio) that is estimated by such a test is more complicated. The “group” standard deviations or coefficients of variation are assumed to be equal because of the symmetry of the period differences or ratios. For a test of normal mean difference, the common standard deviation σ_c involves both treatment means and the correlation between the observations for a given subject:

$$\sigma_c = \frac{1}{2}\left(\sigma_A^2 + \sigma_B^2 - 2\sigma_A\sigma_B\rho_{AB}\right)^{\frac{1}{2}}$$

As a special case, if you assume that the treatment standard deviations are equal (to σ', for example) and that observations on the same subject are uncorrelated, then

$$\sigma_c = \sigma'/\sqrt{2}$$

Thus, if you are doing a power analysis for a test of normal mean treatment difference in an AB/BA crossover design that includes a period effect, then specify μ_{diff} for the MEANDIFF= option and σ_c for the STDDEV= option in the TWOSAMPLEMEANS statement in PROC POWER.

When the period effect is included, the lognormal mean ratio that is estimated in the treatment effect test is again the ratio of geometric treatment means,

$$\gamma_{\mathrm{ratio}} = \gamma_A/\gamma_B$$

and the common coefficient of variation CV_C is

$$\mathrm{CV}_C = \left[\frac{\left(\mathrm{CV}_A^2 + 1\right)^{\frac{1}{4}} \left(\mathrm{CV}_B^2 + 1\right)^{\frac{1}{4}}}{\left(\rho_{AB}\mathrm{CV}_A\mathrm{CV}_B + 1\right)^{\frac{1}{2}}} - 1 \right]^{\frac{1}{2}}$$

As a special case, if you assume that the treatment coefficients of variation are equal (to CV', for example) and that observations on the same subject are uncorrelated, then

$$\mathrm{CV}_C = \left[\left((\mathrm{CV}')^2 + 1\right)^{\frac{1}{2}} - 1 \right]^{\frac{1}{2}}$$

Thus, if you are doing a power analysis for a test of lognormal mean treatment ratio in an AB/BA crossover design that includes a period effect, then specify γ_{ratio} for the MEANRATIO= option and CV_C for the CV= option in the TWOSAMPLEMEANS statement in PROC POWER.

PROPORTIONS, RISK DIFFERENCES, AND RELATIVE RISKS

Data Analysis for Proportions, Risk Differences, and Relative Risks

For equivalence and noninferiority analyses that involve simple categorical data analyses, you can use PROC FREQ to compute *p*-values and confidence intervals for a variety of designs and criteria for binomial proportions, risk differences, and relative risks.

Table 8 shows the options in the TABLES statement in PROC FREQ that correspond to the supported parameters for hypothesis testing and confidence intervals.

Table 8 Parameters for Tests and Confidence Intervals for Proportions in PROC FREQ

Parameter	TABLES Statement Option
Binomial proportion (one-way table)	BINOMIAL
Risk difference (2 × 2 table)	RISKDIFF
Relative risk (2 × 2 table)	RELRISK

Table 9 shows the options that you can specify in parentheses after a relevant option in Table 8 for equivalence and noninferiority analyses.

Table 9 Equivalence and Noninferiority Options for Proportions in PROC FREQ

Criterion	TABLES Statement Statistic Options
Equivalence	EQUIV MARGIN= *value* | (*lower*,*upper*)
Noninferiority	NONINF MARGIN= *value*

All combinations of the rows of Table 8 and Table 9 are supported.

PROC FREQ has several options beyond the options in Table 8 and Table 9 that you can use to request specific hypothesis tests and confidence limits for equivalence or noninferiority:

- The METHOD= option in parentheses after the RISKDIFF or RELRISK option in the TABLES statement specifies the test method.
- The CL= option in parentheses after the BINOMIAL, RISKDIFF, or RELRISK option in the TABLES statement requests specific types of confidence limits.
- The BINOMIAL, RISKDIFF, or RELRISK option in the EXACT statement requests exact versions of tests and confidence limits.

All the equivalence and noninferiority confidence limits are $100(1-2\alpha)\%$ limits based on the approach of Schuirmann (1999). You can compare the confidence limits to either the equivalence limits (θ_L, θ_U) or the noninferiority limit (either $\theta_0 - \delta$ or $\theta_0 + \delta$).

Table 10 through Table 12 show all the tests and confidence limits that are applicable for both equivalence and noninferiority analyses.

Table 10 shows the tests and confidence limits for a binomial proportion in a one-way table.

Table 10 BINOMIAL Options in the TABLES Statement in PROC FREQ for Tests and Confidence Limits for a Binomial Proportion

Test or Confidence Limits	**Options**
Exact (Clopper-Pearson) test[1]	(default)[1]
Wald test with sample variance	(default)
Wald test with null variance	VAR=NULL
Wald test with continuity correction and sample variance	CORRECT
Wald test with continuity correction and null variance	CORRECT VAR=NULL
Agresti-Coull confidence limits	CL=AGRESTICOULL
Blaker confidence limits	CL=BLAKER
Exact (Clopper-Pearson) confidence limits	CL=EXACT
Jeffreys confidence limits	CL=JEFFREYS
Logit confidence limits	CL=LOGIT
Likelihood ratio confidence limits	CL=LR
Mid-*p* (exact) confidence limits	CL=MIDP
Wald confidence limits	CL=WALD
Wald confidence limits with continuity correction	CL=WALD(CORRECT)
Wilson (score) confidence limits	CL=WILSON
Wilson (score) confidence limits with continuity correction	CL=WILSON(CORRECT)

[1] Also specify the BINOMIAL option in the EXACT statement.

Table 11 shows the tests and confidence limits for a risk difference in a 2 × 2 table.

Table 11 RISKDIFF Options in the TABLES Statement in PROC FREQ for Tests and Confidence Limits for a Risk Difference

Test or Confidence Limits	**Options**
Farrington-Manning (score) test	METHOD=FM
Hauck-Anderson test	METHOD=HA
Newcombe (hybrid-score) test	METHOD=NEWCOMBE
Wald test with sample variance	(default)
Wald test with null variance	VAR=NULL
Wald test with continuity correction and sample variance	CORRECT
Wald test with continuity correction and null variance	CORRECT VAR=NULL
Agresti-Caffo confidence limits	CL=AC
Exact unconditional confidence limits[1]	CL=EXACT[1]
Exact unconditional confidence limits based on score statistic[2]	CL=EXACT[2]
Hauck-Anderson confidence limits	CL=HA
Miettinen-Nurminen (score) confidence limits	CL=MN
Miettinen-Nurminen-Mee (uncorrected score) confidence limits	CL=MN(CORRECT=NO)
Newcombe confidence limits	CL=NEWCOMBE
Newcombe confidence limits with continuity correction	CL=NEWCOMBE(CORRECT)
Wald confidence limits	CL=WALD
Wald confidence limits with continuity correction	CL=WALD(CORRECT)

[1] Also specify the RISKDIFF option in the EXACT statement.

[2] Also specify the RISKDIFF(METHOD=SCORE) option in the EXACT statement.

Table 12 shows the tests and confidence limits for a relative risk in a 2 × 2 table. All these options are new for equivalence and noninferiority in SAS/STAT 14.1, even though some of the confidence limits are supported in earlier releases.

Table 12 RELRISK Options in the TABLES Statement in PROC FREQ for Tests and Confidence Limits for a Relative Risk

Test or Confidence Limits	**Options**
Farrington-Manning (score) test	METHOD=FM
Wald test	(default)
Wald modified test	METHOD=WALDMODIFIED
Likelihood ratio test	METHOD=LR
Exact unconditional confidence limits[1]	CL=EXACT[1]
Exact unconditional confidence limits based on score statistic[2]	CL=EXACT[2]
Likelihood ratio confidence limits	CL=LR
Score confidence limits	CL=SCORE
Score confidence limits (uncorrected)	CL=SCORE(CORRECT=NO)
Wald confidence limits	CL=WALD
Wald modified confidence limits	CL=WALDMODIFIED

[1] Also specify the RELRISK option in the EXACT statement.

[2] Also specify the RELRISK(METHOD=SCORE) option in the EXACT statement.

PROC FREQ provides McNemar's test for the analysis of dependent proportions (where the data consist of paired responses). In SAS/STAT 14.1 you can specify a custom null value for the ratio of discordant pairs. This doesn't support a full-fledged equivalence or noninferiority analysis for dependent proportions because PROC FREQ doesn't provide the one-sided tests or confidence limits, but you can produce approximate noninferiority and equivalence results by doubling the significance level α and ignoring the minor tail.

Power and Sample Size Analysis for Proportions, Risk Differences, and Relative Risks

You can use PROC POWER to compute power or sample size when planning studies to be analyzed using the equivalence and noninferiority tests discussed in the section "Data Analysis for Proportions, Risk Differences, and Relative Risks" on page 9. Table 13 and Table 14 show the PROC POWER syntax that corresponds to each supported equivalence or noninferiority test in Table 10 through Table 12.

Table 13 Power Analyses for Exact and Wald Tests for a Binomial Proportion in PROC POWER

Test	**ONESAMPLEFREQ Statement Syntax**
Exact equivalence test	TEST=EQUIV_EXACT LOWER= UPPER=
Exact noninferiority test	TEST=EXACT SIDES=1|U|L MARGIN=
Wald equivalence test with sample variance	TEST=EQUIV_Z VAREST=SAMPLE LOWER= UPPER=
Wald noninferiority test with sample variance	TEST=Z VAREST=SAMPLE SIDES=1|U|L MARGIN=
Wald equivalence test with null variance	TEST=EQUIV_Z LOWER= UPPER=
Wald noninferiority test with null variance	TEST=Z SIDES=1|U|L MARGIN=
Wald equivalence test with continuity correction and sample variance	TEST=EQUIV_ADJZ VAREST=SAMPLE LOWER= UPPER=
Wald noninferiority test with continuity correction and sample variance	TEST=ADJZ VAREST=SAMPLE SIDES=1|U|L MARGIN=
Wald equivalence test with continuity correction and null variance	TEST=EQUIV_ADJZ LOWER= UPPER=
Wald noninferiority test with continuity correction and null variance	TEST=ADJZ SIDES=1|U|L MARGIN=

Table 14 Power Analyses for Farrington-Manning Score Tests for Two Independent Proportions in PROC POWER

Test	TWOSAMPLEFREQ Statement Syntax
Farrington-Manning (score) noninferiority test of risk difference	TWOSAMPLEFREQ TEST=FM SIDES=1|U|L NULLPDIF=
Farrington-Manning (score) noninferiority test of relative risk	TWOSAMPLEFREQ TEST=FM_RR SIDES=1|U|L NULLRR=

The power analyses in Table 14 are new in SAS/STAT 13.2 (for risk difference) and SAS/STAT 14.1 (for relative risk).

The TEST=PCHI option in the TWOSAMPLEFREQ statement matches the Wald tests for *equality* in PROC FREQ for risk differences. But power analysis is not supported for equivalence tests for 2 × 2 tables, and the power analyses for noninferiority tests based on Wald statistics use different forms of the Wald statistics than PROC FREQ for nonzero null plus margin. Consequently, in order to properly align the power analysis and data analysis, you should use the Farrington-Manning score statistics for each.

You can also compute power or sample size for noninferiority tests based on the same McNemar statistics as supported in PROC FREQ (TABLES AGREE(MNULLRATIO=*value*)) with the SIDES=1|U|L and NULLDISCPROPRATIO= options in the PAIREDFREQ statement in PROC POWER.

EXAMPLES

Noninferiority in Manufacturing: Comparing Normal Means

This example from industrial manufacturing shows how to design an experiment that compares two normal means, taking both power and noninferiority considerations into account. You will see how to use the TWOSAMPLEMEANS statement in PROC POWER to compute an appropriate sample size and how to use PROC TTEST to test for noninferiority.

You are an industrial engineer who has invented a new process for manufacturing your company's product. Results from pilot tests are encouraging but not conclusive: the new process makes significantly more product of significantly better quality, but it also seems to produce more waste. Is it *too much* more waste?

This situation calls for a noninferiority test, and you want to design the experiment carefully because each observation requires an expensive run of your process. Managers say that they can deal with a waste increase of as much as 3.7 units. Pilot data indicate that the actual waste increase is probably around 1.5 units, with a standard deviation of about 2 units. How many runs will it take to get a significant noninferiority test for this difference margin of 3.7 units, with reasonable power?

Managers want the experiment to be large enough to leave only a 1% chance of an erroneously significant result (which translates to a power of 0.99). They will tolerate a higher chance (5%) of an erroneously *insignificant* result (which translates to a significance level α of 0.05).

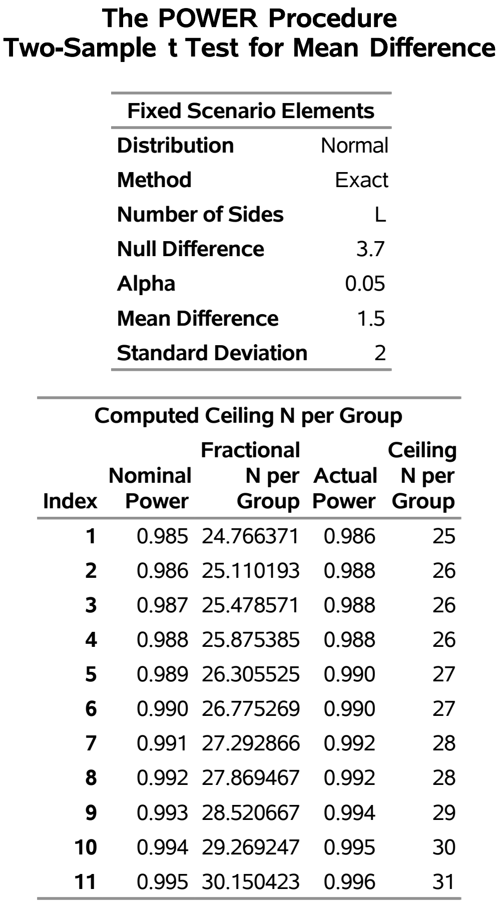
You use the following SAS statements to determine the required number of runs for each process by using a balanced design. The analysis includes several target powers in a small interval around 0.99 to explore the sensitivity of sample size to power.

```
proc power;
   twosamplemeans test=diff nfrac
      nulldiff = 3.7
      meandiff = 1.5
      sides = L
      alpha = 0.05
      stddev = 2
      npergroup = .
      power = 0.985 to 0.995 by 0.001;
run;
```

Recall that the NULLDIFF= option here represents the sum of the null value (here 0) and the margin.

The computed sample sizes are shown in Figure 4. They indicate that, with a true difference of 1.5 and a significance level of 0.05, it will take about 27 runs of each process to give you a solid 99% chance of a significant noninferiority test.

Figure 4 Sample Size Determination for Industrial Waste Study



The POWER Procedure
Two-Sample t Test for Mean Difference

Fixed Scenario Elements	
Distribution	Normal
Method	Exact
Number of Sides	L
Null Difference	3.7
Alpha	0.05
Mean Difference	1.5
Standard Deviation	2

Computed Ceiling N per Group

Index	Nominal Power	Fractional N per Group	Actual Power	Ceiling N per Group
1	0.985	24.766371	0.986	25
2	0.986	25.110193	0.988	26
3	0.987	25.478571	0.988	26
4	0.988	25.875385	0.988	26
5	0.989	26.305525	0.990	27
6	0.990	26.775269	0.990	27
7	0.991	27.292866	0.992	28
8	0.992	27.869467	0.992	28
9	0.993	28.520667	0.994	29
10	0.994	29.269247	0.995	30
11	0.995	30.150423	0.996	31

Now suppose you have gathered the results from the two processes, as shown in Table 15.

Table 15 Mean (Standard Deviation) for 27 Runs

Process	Quality	Yield	Waste
Old	19.9530(0.8077)	40.1481(0.7698)	10.3556(2.1445)
New	34.6667(1.8605)	50.0000(0.8321)	11.9889(2.2548)

As expected, quality, yield, and waste are all elevated in the new process. You enter the numbers from Table 15 into a special input data set for PROC TTEST:

```
data Process;
   input _STAT_ $4. @6 Process $3. @10 Quality Yield Waste;
cards;
N    New 27      27      27
MEAN New 34.6667 50.0000 11.9889
STD  New  1.8605  0.8321  2.2548
N    Old 27      27      27
MEAN Old 19.9530 40.1481 10.3556
STD  Old  0.8077  0.7698  2.1445
;
```

You use the following PROC TTEST statements to perform simple *t* tests for improvements in quality and yield:

```
proc ttest data=Process;
   class Process;
   var Quality Yield;
run;
```

The results (not shown here) are indeed highly significant. How about the waste? Is the increase from the new process too much? The following PROC TTEST code performs a noninferiority test to answer that question:

```
proc ttest data=Process sides=l h0=3.7;
   class Process;
   var Waste;
run;
```

The results are shown in Figure 5.

Figure 5 Noninferiority Test for Industrial Waste Study, Tabular Results

The TTEST Procedure

Variable: Waste

Process	N	Mean	Std Dev	Std Err	Minimum	Maximum
New	27	11.9889	2.2548	0.4339	.	.
Old	27	10.3556	2.1445	0.4127	.	.
Diff (1-2)		1.6333	2.2003	0.5989		

Process	Method	Mean	95% CL Mean		Std Dev	95% CL Std Dev	
New		11.9889	11.0969	12.8809	2.2548	1.7757	3.0900
Old		10.3556	9.5073	11.2039	2.1445	1.6888	2.9389
Diff (1-2)	**Pooled**	1.6333	-Infty	2.6362	2.2003	1.8469	2.7224
Diff (1-2)	**Satterthwaite**	1.6333	-Infty	2.6362			

Method	Variances	DF	t Value	Pr < t
Pooled	Equal	52	-3.45	0.0006
Satterthwaite	Unequal	51.87	-3.45	0.0006

Good news! The mean amount of waste from the new process is found to be significantly noninferior to that from the old process (p = 0.0006).

Noninferiority in Marketing: Comparing Response Rates

This marketing example shows how to design a study to compare two response rates in terms of risk (proportion) difference in a noninferiority setting. You will see how to use the TWOSAMPLEFREQ statement in PROC POWER to compute power and sample size and how to use PROC FREQ to test for noninferiority.

Your company is planning its next big marketing campaign, and you've submitted a proposal to the executive board that advocates the use of recycled paper in company mailings. Printing options are limited compared to those for standard paper, possibly lowering the customer response rate, but public relations priorities favor using recycled paper. At a board meeting, the executives decide that if you can demonstrate that the response rate for mailings with recycled paper isn't appreciably worse than for mailings with standard paper, they'll approve your proposal. They give you permission to send out 4,000 mailings with each type of paper.

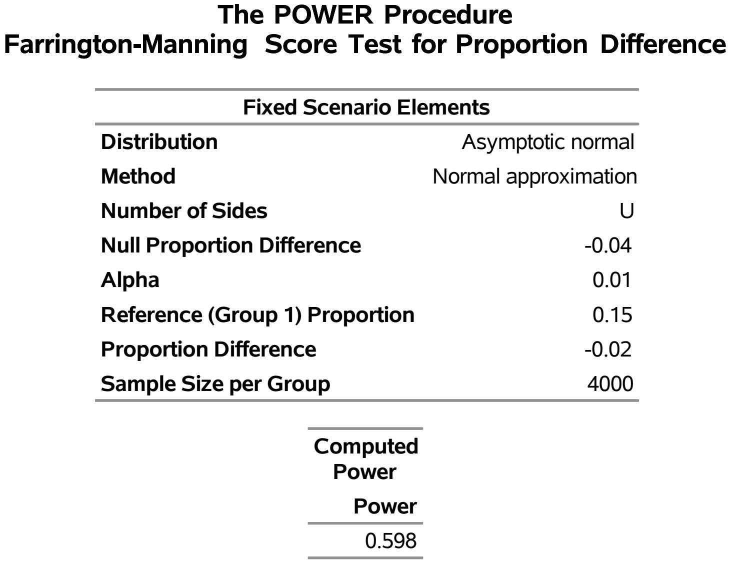
First you need to clarify the executives' definition of "appreciably worse" and the chances of an erroneously significant result—that is, a false positive—that they will tolerate. They inform you that they'd be willing to ignore a response rate difference of 4% or less, and they'll allow for a 1% chance of a false positive.

You decide that you'd better figure out your chances of a significant result before you commit the company's effort and resources to this comparison study. You talk to some colleagues and come up with an educated guess of response rates: 15% for standard paper and 13% for recycled. The recommended choice of statistical test is the Farrington-Manning score test for the difference between proportions (risks) in the two groups, standard paper and recycled paper.

You run the following statements to calculate the power of your planned study. The results are shown in Figure 6.

```
proc power;
   twosamplefreq test=fm
      nullproportiondiff = -0.04
      refproportion = 0.15
      proportiondiff = -0.02
      sides = U
      alpha = 0.01
      npergroup = 4000
      power = .;
run;
```

Figure 6 Power Calculation for Paper Comparison



The POWER Procedure
Farrington-Manning Score Test for Proportion Difference

Fixed Scenario Elements	
Distribution	Asymptotic normal
Method	Normal approximation
Number of Sides	U
Null Proportion Difference	-0.04
Alpha	0.01
Reference (Group 1) Proportion	0.15
Proportion Difference	-0.02
Sample Size per Group	4000

Computed Power
Power
0.598

The power is less than 60%; you really don't want to proceed with the study as planned if your chance of a significant result is that low. You run the following statements to check the required sample size per group for powers between 0.8 and 0.95:

```
proc power;
   twosamplefreq test=fm
      nullproportiondiff = -0.04
      refproportion = 0.15
      proportiondiff = -0.02
      sides = U
      alpha = 0.01
      npergroup = .
      power = 0.8 0.85 0.9 0.95;
run;
```

Figure 7 shows that you'd need over 6,000 mailings per group just to get a power of 80%.

Figure 7 Sample Size Determination for Paper Comparison

The POWER Procedure
Farrington-Manning Score Test for Proportion Difference

Fixed Scenario Elements	
Distribution	Asymptotic normal
Method	Normal approximation
Number of Sides	U
Null Proportion Difference	-0.04
Alpha	0.01
Reference (Group 1) Proportion	0.15
Proportion Difference	-0.02

Computed N per Group

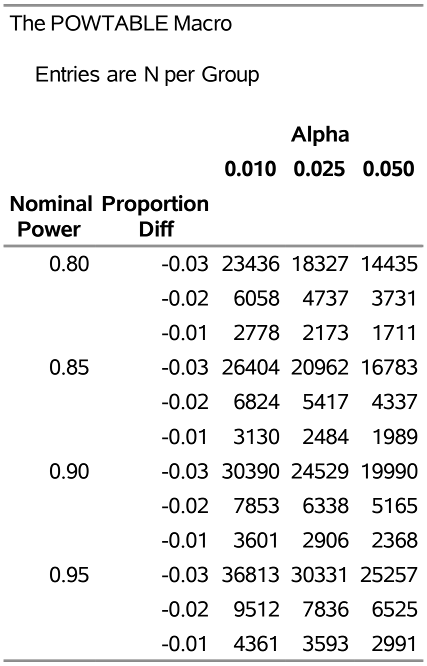
Index	Nominal Power	Actual Power	N per Group
1	0.80	0.800	6058
2	0.85	0.850	6824
3	0.90	0.900	7853
4	0.95	0.950	9512

You wonder how sensitive the required sample size is to choice of significance level and variations in the actual proportion difference. You run the following statements, using the %POWTABLE autocall macro to produce the compact sample size table in Figure 8:

```
proc power;
   ods output output=powdata;
   twosamplefreq test=fm
      nullproportiondiff = -0.04
      refproportion = 0.15
      proportiondiff = -0.03 -0.02 -0.01
      sides = U
      alpha = 0.01 0.025 0.05
      npergroup = .
      power = 0.8 0.85 0.9 0.95;
run;

%powtable (
   Data = powdata,
   Entries = npergroup,
   Panels = power,
   Cols = alpha,
   Rows = proportiondiff
)
```

Figure 8 Sensitivity Analysis for Paper Comparison



```
The POWTABLE Macro

   Entries are N per Group

                                  Alpha
                          0.010  0.025  0.050
Nominal  Proportion
 Power      Diff
  0.80         -0.03     23436  18327  14435
               -0.02      6058   4737   3731
               -0.01      2778   2173   1711
  0.85         -0.03     26404  20962  16783
               -0.02      6824   5417   4337
               -0.01      3130   2484   1989
  0.90         -0.03     30390  24529  19990
               -0.02      7853   6338   5165
               -0.01      3601   2906   2368
  0.95         -0.03     36813  30331  25257
               -0.02      9512   7836   6525
               -0.01      4361   3593   2991
```

The first thing you notice is the dramatic increase in required sample size for the scenario of a 3% lower response rate for recycled paper. You're not too surprised, because that's awfully close to the margin of 4%, and tiny proportion differences are very difficult to detect. If the response rate for recycled paper were 1% lower than for standard paper, your required sample size would be more than cut in half.

You check again with your colleagues about their conjecture of a 2% lower response rate, but they stand by it firmly. Taking another look at Figure 8, you notice that increasing the number of mailings per group to 6,500 or relaxing the significance level requirement from $\alpha = 0.01$ to $\alpha = 0.05$ would increase the power to a level between 80% and 85%. You explain this to the board.

The executives won't budge on the number of mailings, but they are willing to allow for a 5% false positive chance. So you proceed with the original study plan, grudgingly accepting the 80%–85% risk of being foiled by Lady Luck. After a few months you tally the results of the mailings to produce Table 16.

Table 16 Results of Paper Comparison Study

	Response	
Paper	Yes	No
Recycled	507	3,493
Standard	622	3,378

You write the following SAS statements to perform the noninferiority test:

```
data PaperComp;
   input Paper $ Response $ Count;
   datalines;
   Recycled Yes  507
   Recycled  No 3493
   Standard Yes  622
   Standard  No 3378
;
```

```
proc freq data=PaperComp order=data;
   tables Paper*Response /
      riskdiff(noninf margin=.04 method=fm norisks);
   weight Count;
run;
```

The results in Table 9 show a *p*-value of 0.0743, insignificant even at the board's relaxed criterion of $\alpha = 0.05$.

Figure 9 Noninferiority Analysis for Paper Comparison

```
                     The FREQ Procedure

          Statistics for Table of Paper by Response

 ------------------------------------------------------------
    Noninferiority Analysis for the Proportion (Risk)
                        Difference

     H0: P1 - P2 <= -Margin    Ha: P1 - P2 > -Margin

        Margin = 0.04    Score (Farrington-Manning)
                             Method

 Proportion Difference  ASE (F-M)          Z     Pr > Z
 ------------------------------------------------------------
               -0.0288     0.0078     1.4447     0.0743
 ------------------------------------------------------------

      ---------------------------------------------------
      Noninferiority Limit   90% Confidence Limits
      ---------------------------------------------------
                   -0.0400      -0.0416        -0.0159
      ---------------------------------------------------
```

The observed proportion difference of –0.0288 is more than a standard error below your conjectured difference of –0.02. This leads you to suspect that your calculated power (which assumed the difference of –0.02, among other things) might very well have been overly optimistic. So it often goes with power analysis: the power calculation is only as accurate as the conjectures that drive it.

Equivalence in Clinical Drug Trials: Comparing Bioavailability in Terms of Lognormal Mean Ratio

This pharmacokinetics example shows how to plan and analyze a clinical trial to establish bioequivalence between a generic drug and a name-brand drug by using an AB/BA crossover design with lognormal data. You will see how to use both the PAIREDMEANS and TWOSAMPLEMEANS statements in PROC POWER (depending on whether you include a period effect in the crossover design) to compute an appropriate sample size and how to use PROC TTEST to test for bioequivalence.

As the principal statistician for a new generic "drug A" developed by your pharmaceutical company, you're planning a clinical trial to demonstrate bioavailability similar to that of a rival, "drug B."

The data analysis plan is to compare the area under the serum-concentration curve (AUC) for the two drugs by using the two one-sided test (TOST) approach for a lognormal mean ratio with the classic 80%–125% average bioequivalence limits, assuming equal coefficient of variation (CV) for the two drugs.

The design is the AB/BA crossover design, and supply chain limitations mean that you need to plan for twice as many subjects getting the rival drug first (sequence BA). You want to determine the number of subjects to recruit in order to achieve a power of 0.9 for the equivalence test with a significance level of $\alpha = 0.05$.

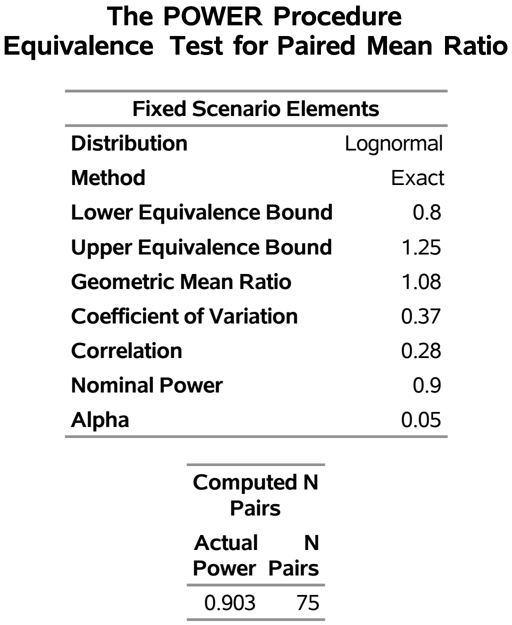
Based on results of previous studies, you conjecture a geometric mean AUC ratio (A to B) of 1.08, a correlation of 0.28 between AUC measurements on the same subject (one for each drug), and a common CV of 0.37.

Your company hasn't decided yet whether to include a period effect in the crossover analysis, so you perform two power analyses to cover both cases. PROC POWER doesn't handle crossover designs directly, but you know that if the period effect is ignored in the data analysis, then the statistical test boils down to a paired *t* test on the square roots of the period ratios, ignoring treatment sequence. The parameters that are estimated in this paired *t* test are the same as the ones you have already conjectured values for. Thus you already have all the information you need to perform this power analysis by using the PAIREDMEANS statement in the POWER procedure, as follows:

```
proc power;
   pairedmeans test=equiv_ratio
      lower = 0.8
      upper = 1.25
      meanratio = 1.08
      corr = 0.28
      cv = 0.37
      npairs = .
      power = 0.9;
run;
```

The results in Figure 10 show that 75 subjects are needed to achieve a power of 0.9 if the data analysis ignores the crossover period effect.

Figure 10 Sample Size Determination for Bioequivalence Study Assuming Period Effect Will Be Ignored



The POWER Procedure
Equivalence Test for Paired Mean Ratio

Fixed Scenario Elements	
Distribution	Lognormal
Method	Exact
Lower Equivalence Bound	0.8
Upper Equivalence Bound	1.25
Geometric Mean Ratio	1.08
Coefficient of Variation	0.37
Correlation	0.28
Nominal Power	0.9
Alpha	0.05

Computed N Pairs	
Actual Power	**N Pairs**
0.903	75

You also know that if the data analysis includes a period effect, then the statistical test is a two-sample *t* test on the square roots of the period ratios, where the "groups" being compared are the two treatment sequences. So you can treat the crossover analysis instead as a two-sample *t* test for purposes of power analysis.

The geometric mean ratio parameter in this two-sample *t* test is the same as the one you already conjectured a value for. But the CV is different, as discussed in the section "Power and Sample Size Analysis for Normal and Lognormal Means" on page 7. You calculate the CV for use in PROC POWER as

$$\begin{aligned}
\mathrm{CV}_C &= \left[\frac{\left(\mathrm{CV}_A^2 + 1\right)^{\frac{1}{4}} \left(\mathrm{CV}_B^2 + 1\right)^{\frac{1}{4}}}{\left(\rho_{AB}\mathrm{CV}_A\mathrm{CV}_B + 1\right)^{\frac{1}{2}}} - 1 \right]^{\frac{1}{2}} \\
&= \left[\frac{\left(0.37^2 + 1\right)^{\frac{1}{4}} \left(0.37^2 + 1\right)^{\frac{1}{4}}}{\left(0.28(0.37)(0.37) + 1\right)^{\frac{1}{2}}} - 1 \right]^{\frac{1}{2}} \\
&= 0.21538
\end{aligned}$$

You use the following statements to compute the number of subjects to recruit if the data analysis will include a crossover period effect:

```
proc power;
   twosamplemeans test=equiv_ratio
      lower = 0.8
      upper = 1.25
      meanratio = 1.08
      cv = 0.21538
      groupweights = (1 2)
      ntotal = .
      power = 0.9;
run;
```

The results in Figure 11 show that you'll need 84 subjects if the data analysis includes a period effect, compared to 75 if the period effect is ignored.

Figure 11 Sample Size Determination for Bioequivalence Study Including Period Effect

The POWER Procedure
Equivalence Test for Mean Ratio

Fixed Scenario Elements	
Distribution	Lognormal
Method	Exact
Lower Equivalence Bound	0.8
Upper Equivalence Bound	1.25
Mean Ratio	1.08
Coefficient of Variation	0.21538
Group 1 Weight	1
Group 2 Weight	2
Nominal Power	0.9
Alpha	0.05

Computed N Total

Actual Power	N Total
0.903	84

Your company decides to include a period effect, and a clinical trial is conducted with 84 subjects, 28 getting your generic drug A first and the other 56 getting the rival drug B first. With the AUC measurements from the study in hand, you run the following statements to perform the bioequivalence test:

```
data auc;
   input Trt1 $ Trt2 $ AUC1 AUC2 @@;
   datalines;
A B 336 339 A B 325 335 A B 310 217 A B 263 128 A B 244 305 A B 517 268
A B 226 163 A B 230 210 A B 300 364 A B 309 234 A B 259 349 A B 251 223
A B 359 288 A B 192 174 A B 396 313 A B 170 187 A B 526 360 A B 170 127
A B 200 130 A B 324 186 A B 352 281 A B 179 355 A B 387 332 A B 278 287
A B 464 244 A B 216 385 A B 587 434 A B 419 458
B A 201 273 B A 142 194 B A 214 374 B A 548 280 B A 220 417 B A 345 178
B A 321 654 B A 254 503 B A 268 354 B A 217 172 B A 228 158 B A 230 108
B A 262 280 B A 189 277 B A 184 283 B A 123 123 B A 338 455 B A 392 359
B A 215 377 B A 249 410 B A 123 243 B A 354 279 B A 205 172 B A 303 255
B A 249 255 B A 280 398 B A 237 292 B A 301 239 B A 278 344 B A 241 192
B A 151 213 B A 154 500 B A 417 394 B A 245 400 B A 198 245 B A 466 337
B A 275 262 B A 109 147 B A 205 198 B A 268 163 B A 213 334 B A 195 202
B A 213 297 B A 195 120 B A 212 401 B A 238 177 B A 235 171 B A 224 379
B A 392 154 B A 248 299 B A 159 250 B A 381 386 B A 370 230 B A 127 195
B A 279 335 B A 308 282
;
```

```
proc ttest data=auc dist=lognormal tost(0.8, 1.25) plots;
   var AUC1 AUC2 / crossover=(Trt1 Trt2);
run;
```

The first several tables in the output (not shown) display information about the crossover variables, basic summary statistics, and the usual confidence limits relevant to a typical (as opposed to equivalence) data analysis. Figure 12 shows $100(1-2\alpha)\%$ = 90% confidence limits, which is relevant here because these confidence limits are contained completely within the equivalence limits of [0.8, 1.25] if and only if the level-α TOST equivalence test is rejected. The confidence interval for the treatment geometric mean ratio does in fact lie completely within [0.8, 1.25], resulting in an assessment of "Equivalent."

Figure 12 Equivalence Intervals for Analysis Including Period Effect in Crossover Design

The TTEST Procedure

Response Variables: AUC1, AUC2

TOST Level 0.05 Equivalence Analysis

Treatment	Period	Method	Geometric Mean	Lower Bound		90% CL Mean			Upper Bound	Assessment
Ratio (1/2)		**Pooled**	1.1260	0.8	<	1.0419	1.2169	<	1.25	Equivalent
Ratio (1/2)		**Satterthwaite**	1.1260	0.8	<	1.0476	1.2103	<	1.25	Equivalent
	Ratio (1/2)	**Pooled**	1.0190	0.8	<	0.9428	1.1012	<	1.25	Equivalent
	Ratio (1/2)	**Satterthwaite**	1.0190	0.8	<	0.9480	1.0952	<	1.25	Equivalent

Figure 13 shows a significant *p*-value of 0.014 for the pooled test. Your company celebrates the successful demonstration of pharmacokinetic equivalence between your generic drug and the name-brand rival.

Figure 13 Equivalence Tests Including Period Effect in Crossover Design

Treatment	Period	Method	Coefficients of Variation	Test	Null	DF	t Value	P-Value
Ratio (1/2)		**Pooled**	Equal	**Upper**	0.8	82	7.32	<.0001
Ratio (1/2)		**Pooled**	Equal	**Lower**	1.25	82	-2.24	0.0140
Ratio (1/2)		**Pooled**	Equal	**Overall**				0.0140
Ratio (1/2)		**Satterthwaite**	Unequal	**Upper**	0.8	66.163	7.90	<.0001
Ratio (1/2)		**Satterthwaite**	Unequal	**Lower**	1.25	66.163	-2.41	0.0093
Ratio (1/2)		**Satterthwaite**	Unequal	**Overall**				0.0093
	Ratio (1/2)	**Pooled**	Equal	**Upper**	0.8	82	5.18	<.0001
	Ratio (1/2)	**Pooled**	Equal	**Lower**	1.25	82	-4.38	<.0001
	Ratio (1/2)	**Pooled**	Equal	**Overall**				<.0001
	Ratio (1/2)	**Satterthwaite**	Unequal	**Upper**	0.8	66.163	5.59	<.0001
	Ratio (1/2)	**Satterthwaite**	Unequal	**Lower**	1.25	66.163	-4.72	<.0001
	Ratio (1/2)	**Satterthwaite**	Unequal	**Overall**				<.0001

WHERE TO FIND MORE INFORMATION

The statistical techniques that this paper presents have a great number of moving parts. To the already hefty machinery of typical tests for significant differences, equivalence and noninferiority tests add margins of various types, with different possibilities for how you incorporate them into the analysis. Also, power and sample size analysis adds the need to carefully consider and specify alternative hypotheses and the probability of failing to reject the null. Indeed, the complete process of planning equivalence and noninferiority tests of sufficient power is not for the faint of heart!

The general discussion of combinations of options in the FREQ, POWER, and TTEST procedures in the sections "NORMAL AND LOGNORMAL MEANS" on page 5 and "PROPORTIONS, RISK DIFFERENCES, AND RELATIVE RISKS" on page 9 gives you a catalog of the tools you will use to navigate this process. The three extended examples in the section “EXAMPLES” on page 12 demonstrate these tools in action for some common equivalence and noninferiority tasks, showing how their results can be interpreted.

However, there are many more noninferiority and equivalence tasks than this paper can discuss in detail. For more information about the design and analysis of equivalence and noninferiority tests, see the references listed at the end of this paper. There are also two SAS usage notes that provide more details and examples, SAS Institute Inc. (2013a) on noninferiority and SAS Institute Inc. (2013b) on equivalence. Finally, for complete details about the FREQ, POWER, and TTEST procedure syntax that is required to perform these analyses in SAS, see the *SAS/STAT User's Guide*.

REFERENCES

Agresti, A. (2013). *Categorical Data Analysis*. 3rd ed. Hoboken, NJ: John Wiley & Sons.

Barker, L., Rolka, H., Rolka, D., and Brown, C. (2001). “Equivalence Testing for Binomial Random Variables: Which Test to Use?” *American Statistician* 55:279–287.

Blackwelder, W. C. (1982). “‘Proving the Null Hypothesis’ in Clinical Trials.” *Controlled Clinical Trials* 3:345–353.

Chow, S.-C., and Liu, J.-P. (2009). *Design and Analysis of Bioavailability and Bioequivalence Studies*. 3rd ed. Boca Raton, FL: CRC Press.

Chow, S.-C., Shao, J., and Wang, H. (2008). *Sample Size Calculations in Clinical Research*. 2nd ed. Boca Raton, FL: Chapman & Hall/CRC.

Dann, R. S., and Koch, G. G. (2008). “Methods for One-Sided Testing of the Difference between Proportions and Sample Size Considerations Related to Non-inferiority Clinical Trials.” *Pharmaceutical Statistics* 7:130–141.

Diletti, D., Hauschke, D., and Steinijans, V. W. (1991). “Sample Size Determination for Bioequivalence Assessment by Means of Confidence Intervals.” *International Journal of Clinical Pharmacology, Therapy, and Toxicology* 29:1–8.

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* 9:1447–1454.

Hauschke, D., Kieser, M., Diletti, E., and Burke, M. (1999). “Sample Size Determination for Proving Equivalence Based on the Ratio of Two Means for Normally Distributed Data.” *Statistics in Medicine* 18:93–105.

Hauschke, D., Steinijans, V., and Pigeot, I. (2007). *Bioequivalence Studies in Drug Development: Methods and Applications*. Chichester, UK: John Wiley & Sons.

Miettinen, O. S., and Nurminen, M. M. (1985). “Comparative Analysis of Two Rates.” *Statistics in Medicine* 4:213–226.

Patterson, S. D., and Jones, B. (2006a). “Bioequivalence: A Review of Study Design and Statistical Analysis for Orally Administered Products.” *International Journal of Pharmaceutical Medicine* 20:243–250.

Patterson, S. D., and Jones, B. (2006b). *Bioequivalence and Statistics in Clinical Pharmacology*. Boca Raton, FL: Chapman & Hall/CRC.

Phillips, K. F. (1990). “Power of the Two One-Sided Tests Procedure in Bioequivalence.” *Journal of Pharmacokinetics and Biopharmaceutics* 18:137–144.

Rothmann, M. D., Wiens, B. L., and Chan, I. S. F. (2012). *Design and Analysis of Non-inferiority Trials*. Boca Raton, FL: Chapman & Hall/CRC.

SAS Institute Inc. (2013a). “Usage Note 48616: Design and Analysis of Noninferiority Studies.” SAS Institute Inc., Cary, NC. `http://support.sas.com/kb/48/616.html`.

SAS Institute Inc. (2013b). “Usage Note 50700: Design and Analysis of Equivalence Tests.” SAS Institute Inc., Cary, NC. `http://support.sas.com/kb/50/700.html`.

Schuirmann, D. J. (1987). “A Comparison of the Two One-Sided Tests Procedure and the Power Approach for Assessing the Equivalence of Average Bioavailability.” *Journal of Pharmacokinetics and Biopharmaceutics* 15:657–680.

Schuirmann, D. J. (1999). “Confidence Interval Methods for Bioequivalence Testing with Binomial Endpoints.” In *Proceedings of the Biopharmaceutical Section*, 227–232. Alexandria, VA: American Statistical Association.

Wellek, S. (2010). *Testing Statistical Hypotheses of Equivalence and Noninferiority*. 2nd ed. Boca Raton, FL: CRC Press.

ACKNOWLEDGMENTS

The authors are grateful to Randy Tobias, Ed Huddleston, and Tim Arnold of the Advanced Analytics Division at SAS Institute Inc., and to David Schlotzhauer and Jill Tao of the Technical Support Division at SAS, for their valuable assistance in the preparation of this paper.

CONTACT INFORMATION

Your comments and questions are valued and encouraged. Contact the author:

John Castelloe
SAS Institute Inc.
SAS Campus Drive
Cary, NC 27513
919-531-5728
john.castelloe@sas.com