Paper MWSUG 2019, IN-113

Sample Size and Design Considerations in Studies Assessing Non-Inferiority using Continuous Outcomes

Michael G. Wilson
Indianapolis IN, USA

ABSTRACT

Software developers employ incremental progress to cause radical development break throughs. The same is true in medicine, manufacturing and finance. For example, a new anti-diabetic medicine might not have superior outcomes of improved glycemic control, but it might be less expensive. Or a new device for use in hand surgery might not have superior digital mobility, but might be easier for the surgeon to implant. Or perhaps micro-loans to novice entrepreneurs might not raise the economic output for the county, but might cultivate cooperation among local businesses. These are examples where the outcome of a new method might not be objectively worse, that is non-inferior, but would have some reason to replace the current method and instigate incremental progress. SAS users are often asked to size and design studies to test this kind of non-inferiority. Such a design requires consideration of the frame-work of the hypothesis set-up, the directionality, the determination of the non-inferiority margin and the proper analysis method. In this review, the rationale for these considerations will be presented, common misunderstandings clarified and examples using SAS/STAT® given.

1. Introduction

Prospective experiments in a broad array of industries provide priceless information to the enterprise regarding causative effect. However, they are frequently expensive. Proper power estimation ensures that the experimental outcomes will be valuable.

Attribution of causative effects in experimental design is definitively demonstrated with the use of randomly assigning experimental units either to the experimental intervention group and a 'negative control group,' (for example, a placebo-controlled clinical trial) or by series of escalating intensity of the intervention (for example, a dose-response study or a puncture stress test). Hypothesis from these types of designs are called superiority.

In contrast, a 'positive control' study refers to those studies where a control treatment employed is active. The active comparator has a previously known causative effect on outcome and can serve as a reference.

Positive control studies have been chosen for use in one situation when a negative-control design would be unethical. For example, in anti-infection studies for the treatment of pneumonia, biological cure rates can be 80-90% and it would be inconceivable to assign some patients to a placebo to assess the effectiveness of a new interventional product. Equally unimaginable would be the assignment to placebo of patients suffering from any one of the many life-threatening cancers (Mukherjee, 2010) (see reference for free copy of his ebook).

But active controls can have a less dramatic but equally useful purpose as well. Active controls are also used to determine how experimental treatments compare and aim to demonstrate that interventions of interest have either superior effects or non-inferior effects to the control or reference.

But for as important and useful as this design is, there is as much confusion in the literature surrounding the execution and reporting of noninferiority studies. A recent study published in JAMA identified 162 reports of noninferiority trials published in 2003 and 2004 and found 78% either deficient methods or had misleading conclusions.

Offered here is an attempt to improve some of the clarity by examining design considerations in studies assessing non-inferiority using continuous outcomes.

This paper is about the continuous case. However, it is surprisingly simple to apply its contents to the binomial case. Also, this paper is intended for intermediate SAS users with completion of an introductory, collegiate-level statistical course.

2. Testing Statistical Hypotheses

First, a review of experimental outcomes, hypothesis testing framework and elements are reviewed because of their central role in non-inferiority testing. Readers regularly working with these outcomes, frameworks and elements may prefer to skip these three sections and go to Section 5. But a review, if for no other reason than the notation, is worthwhile.

2.1 Experimental Outcomes

The conclusion drawn at the end of the experiment can have two possible outcomes. However, at the beginning of the experiment, the two outcomes are written down and called 'conjectures.' Experiments are designed to test a conjecture the investigator believes to be true.

The opposing conjecture that the investigator believes is false, when stated mathematically, is the default hypothesis and is called the 'Null Hypothesis' (See Table A).

Table 2.1. Prospective Experimental Outcomes by Phase

Experimental Phase	Outcome
Before	Conjecture
During	Hypothesis
After	Conclusion

2.2 Hypotheses in the Testing Procedure

In statistical hypothesis testing, the alternative hypothesis specifies the researcher's belief that some effect exists in a well-defined population written in mathematical terms. The alternative hypothesis, written as H_{alt} or H_a, where the subscript 'alt' or 'a' are abbreviations for 'alternative.'

The null hypothesis, H_{null}, sometimes written as H_0, is the assertion that the effect does not truly exist. Written in mathematical terms as H_{null} or H_0, where the subscript zero '0' is a symbol for 'no effect.' Evidence is then gathered to reject H_{null} in favor of H_{alt}. A statistical test is used to assess H_{null}. If H_{null} is rejected, but there truly is no effect, this is called a Type I error.

Review of the hypothesis testing process might be worthwhile.

2.3 Elements of Hypothesis Testing

Seven basic elements of hypothesis testing can be adapted from previous authors (McClave & Dietrich, 1985) page 283.

1. Null hypothesis (H_{null}, Ho) A scientific theory mathematically phrased in terms of the values of one or more population parameters. This theory is the exact opposite of what the researcher wishes to prove. usually one that we wish to disprove. It is presumed to be the truth and the default conclusion of the study.

2. Alternative (research) hypothesis (H_{alt}, H_a, H_1). A theory that opposes the null hypothesis and that we wish to establish as true.

3. Test statistic: A sample statistic used to decide whether to reject the null hypothesis.

4. Rejection region: The numerical values of the test statistic for which the null hypothesis will be rejected. The rejection region is chosen so that the probability is alpha that it will contain the test statistic if the null hypothesis is true (thereby leading to an incorrect conclusion), where alpha is usually chosen to be small (say, .01, .05, or .10).

5. Experimentation and Data Collection: The sampling experiment is performed.

6. Calculation of the test statistic: The numerical value of the test statistic is determined.

Conclusion:

7a. Reject If the numerical value of the test statistic falls in the rejection region, we conclude that the alternative hypothesis is true (i.e. reject the null hypothesis), and we know that the test procedure will lead to this conclusion incorrectly only 100*alpha percent of the time it is used.

7b. Do Not Reject If the test statistic does not fall in the rejection region, we reserve judgment about which hypothesis is true. We 'Do Not Reject' the null hypothesis. We do not 'Accept' the null hypothesis, because we do not know the probability beta that our test procedure will lead us to accept the Null hypothesis (H_{null}, Ho) incorrectly.

3. Formulation of Hypothesis

3.1 Presumption of the Null

Experiments are conducted a little like the criminal justice system in the United States. One of the most sacred principles of the American justice system is that the defendant is presumed innocent until the prosecution proves guilt, beyond a reasonable doubt.

The alternative hypothesis, that the prosecution believes is true, is that the defendant is guilt. But in a court of law, the defendant enjoys the presumption of innocence, which is similar to the null hypothesis. When the jury returns a verdict of 'Not Guilt', it doesn't mean that the defendant is truly innocent, only that the jury decided there was insufficient evidence to prove guilt, beyond a reasonable doubt.

This presumption creates a situation where the jury's conclusion can result in one of two errors. Firstly, they can wrongfully convict an innocent (conclude the alternative, when the null is true). This is called a Type I error and is a false positive. Secondly, they can fail to convict a truly guilty criminal (fail to reject the null, when the alternative is true). This is called a Type II error and is a false negative.

3.2 Errors in Experimental Hypothesis Testing

Unlike the criminal justice system, errors in experimental outcome can be quantified. In order to quantify these errors, some notation is required.

Firstly, consider the two errors in hypothesis testing, Type I (false positive) and Type II (false negative).

Secondly, consider the general family of statistics, D, that are normally-distributed. Under the null hypothesis, $D \sim n(\mu_0, \Sigma_0^2)$.

When the primary measurement is the difference from baseline to the endpoint of the experiment a change of zero represents no difference. Often in superiority testing, the null hypothesis specifies 'an effect of zero,' so often $\mu_0 = 0$.

Under the alternative hypothesis, $D \sim n(\mu_1, \Sigma_1^2)$. Historically in the best designed experiments and in well-designed experiments, it is important to choose a value under the alternative hypothesis, μ_1, that is a value which is minimally relevant difference from the null value.

In general, the variance of the statistic Σ^2, is the variance of the statistic, D. It is a function of the more well-known variance of the 'N" individual observations, σ^2.

If $D \sim n(\mu, \Sigma^2)$ then, $Z = [X-\mu]*\Sigma^{-1}$ is $\sim n(0,1)$, which is the standard normal distribution.

3.3 Specifications of Decision Errors

The mathematical probability of committing a Type I error can be written as alpha, α. Likewise the probability of committing a Type II error can be written as beta, β.

$$\alpha = P(type\ I\ error) = P(reject\ H_0 | H_0\ is\ true)$$
$$\beta = P(type\ II\ error) = P(fail\ to\ reject\ H_0 | H_0\ is\ false)$$

(Equations 3.1 and 3.2)

The value of the alpha, α, is called the significance level. The power of an experiment at a specified alpha level is the probability of rejecting the null hypothesis when the null hypothesis is false and is written as:

$$Power = (1-\beta) * 100 = P(\ reject\ H_0 | H_0\ is\ false) * 100$$

(Equation 3.3)

If there truly is an effect in the population, but H_{null} is not rejected in the statistical test, then a Type II error has been made. The probability of avoiding a Type II error—that is, correctly rejecting H_{null} and achieving statistical significance when there truly is an effect—is called the power.

An important goal in study planning is to ensure an acceptably high level of power. Sample size plays a prominent role in power computations because the focus is often on determining a sufficient sample size to achieve power at a certain magnitude, or assessing the power for a range of different sample sizes.

3.4 Quantification of the Type I Error

Now given that the distribution under the null and alternative hypothesis from Section 3.2, the errors can be written algebraically, though not presented here. The distribution of the statistic, D, assuming the null hypothesis is true, the critical region dependent on alpha, and critical value can be represented graphically (Figure 3.1):

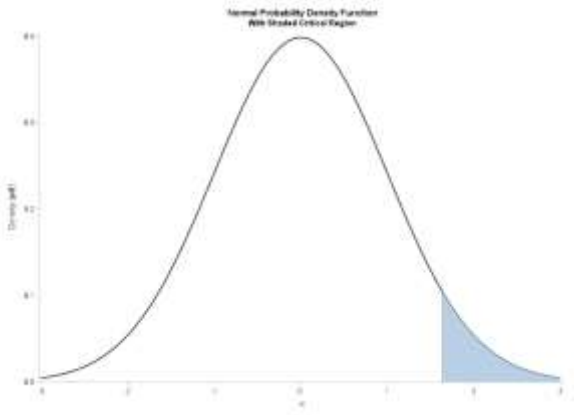


Figure 3.1. The distribution of a statistic, D, with variance, Σ2, under the null hypothesis and the probability of Type I error (alpha). Here alpha is illustrated to have a one-sided level of 0.05, so the critical region is bordered and determined by the critical value, $D_\alpha = \mu_0 + Z_\alpha \Sigma_0$.

The curve in Figure 3.1 can be created in SAS/Base using the pdf function in the data step (Snippet 3.1) and the sgplot procedure:

```
/* Normal Density Function    */
%let alpha = 0.05; %let sides = 1;
%let mu = 0; %let sigma = 1;
   do x=&mu - 3*&sigma to &mu + 3*&sigma by 0.01;
      density = pdf('normal',x,&mu,&sigma);
      output;
```

(Snippet 3.1)

4. Superiority Testing

4.1 Directionality in Superiority Testing

When designing a superiority study, endpoints are carefully chosen that have external validity and make a difference and impact to the researcher. Many endpoints are a change or a difference in the chosen measurement within an individual from the beginning of the study to end of the study.

$$Change = X_{end} - X_{beginning}$$
(Equation 4.1)

If the upper-case delta symbol represents the difference between the two comparator groups, investigational and control, the difference of interest is:

$$\Delta = \theta_{inv} - \theta_{cnt}$$
(Equation 4.1)

The use of the upper-case delta symbol represents the difference between groups based on the population. This emphasizes that the discipline of statistics is concerned with relative comparisons or in answering the question, 'Compared to what?' The parameters of each group are given a value of theta, which has the advantage of being general. It can represent either the population mean, when D from Equation 4.1 is continuous, or the population proportion, when D is binomial.

When smaller values of the change endpoints indicate improvement, the hypothesis to be tested is:

Null Hypothesis: Ho: $\Delta \geq 0$
Alternative Hypothesis: Ha: $\Delta < 0$
(Hypotheses Pair 4.1)

If smaller values are superior (better) then the alternative hypothesis parameters are less than the null value (which is zero in this case).

For example, in a solid tumor cancer study, a researcher could be interested in the change of the size of the tumor. In that case, values of change that are smaller at the end of the study could indicate improvement. The distribution of the alternative hypothesis would be to the left of the null hypothesis. Notice the direction of the less than sign in the alternative hypothesis.

Conversely, when larger values of the change endpoints denote improvement the hypothesis to be tested is are in the opposite direction and are written as:

Null Hypothesis: Ho: $\Delta \leq 0$
Alternative Hypothesis: Ha: $\Delta > 0$
(Hypotheses Pair 4.2)

For example, in a cardiovascular study of cardiomegaly, a researcher could be interested in the change in the volume of cardiac output from the beginning to the end of the study. In that case, values of change that are larger indicate improvement in the distressed population. In this case, the distribution of the alternative hypothesis would be to the right of the null hypothesis. Notice the direction of the greater than sign in the alternative hypothesis.

These hypotheses can be generalized by adding a zero-null value for the endpoint without changing the equality. In mathematical terms, the zero can be added without loss of generality. Adding zero-null values will be useful later. For smaller values that indicate improvement:

Null Hypothesis: Ho: $\Delta - (0) \geq 0$
Alternative Hypothesis: Ha: $\Delta - (0) < 0$
(Hypotheses Pair 4.3)

And for larger values that indicate improvement:

Null Hypothesis: Ho: $\Delta - (0) \leq 0$
Alternative Hypothesis: Ha: $\Delta - (0) > 0$
(Hypotheses Pair 4.4)

Notice how all of the terms are on the left side of the equation. When that happens, a value of zero is on the right side. Then the statistic of interest in the hypothesis test, which will be shown later, will be distributed according to a standard distribution.

5. Introducing Non-Inferiority Testing

Presumption of the null, directionality and tolerable-inferiority are three features that help to understand the contrasts between superiority and non-inferiority testing.

Part of the nature of hypothesis testing requires the presumption of the null. This precludes the use of the superiority framework in the non-inferiority design. Use of the superiority hypothesis test is inappropriate because the goal is to demonstrate similarity rather than difference (Blackwelder, 1982).

Non-inferiority testing shares with superiority testing the issue of directionality. In both testing procedures they are dependent on the preferred direction of the endpoint to be measured in the experiment. However, as will be shown, it is much more difficult to track issues related directionality when the alternative hypothesis to be tested is non-inferiority.

Often in a study it is not the purpose or the intent to show an experimental or investigational process or product is superior to the control. Sometimes the only intent is to show the investigational process is not intolerably worse. So, a value that is not intolerably worse must be pre-specified. This highlights the care that must be taken when using the term, 'Non-inferiority.' The investigational process might be inferior. However, that is not what is being tested. What is

being tested is that the investigational process is not more inferior than what can be tolerated.

5.1 Basic Difference is when the Null Value (δ) is Non-Zero

When designing non-inferiority study, it is most efficient to set-up the framework and orientation consistently and work methodically. Consider always writing down the directional of preference for the endpoint and the hypothesis formulation. It is not required. However, it has been seen that those who do not have gotten themselves helplessly confused.

In the case when consistent orientation of the $\Delta = \theta_{inv} - \theta_{cnt}$, is used, the basic difference between the classification of tests for superiority and non-inferiority is that the null value, δ is non-zero.

This δ is a parameter in the hypothesis testing equations and is often called, the 'non-inferiority margin' which can be abbreviated NIM. Because the null value NIM is a parameter in the hypothesis, it should be pre-specified before the experiment is begun. The null value NIM acts as the minimally tolerable difference. It is the smallest amount for which the experimental group can differ from the control and remain, in the opinion of the researcher, not intolerably worse. Often in the research literature a value of 20% shift of the distributional difference between the experimental and reference is common. But the best practice is pre-specified determination of the NIM, can be complex and an example will be given later.

Incorporation of the null-value, NIM into the hypothesis statements can be done be replacing the 0 with the non-zero δ parameter. Although only the magnitude of the δ is important for the hypothesis test its directionality needs to fit with the hypothesis framework of the specific study (Rothmann, Wiens, & Chan, 2012).

5.2 Directionality

How the δ appropriately fits with the hypothesis framework requires taking into consideration the preferred direction of your endpoint and the orientation of your difference parameter, Δ. For this paper and this section in particular, the experimental minus the control, $\Delta = \theta_{inv} - \theta_{cnt}$ will be used and all parameters will be on the left side of the hypothesis equation.

The preferred direction of the endpoint determines three things. These things might be obvious, but the careful analyst will take notice of each. Firstly, it determines on which side of the null hypothesis distribution the alternative distribution appears. When smaller values are preferred, the alternative distribution is to the left of the null. When larger values are preferred, the alternative distribution is positioned to the right. This is exactly as they are specified for superiority testing as shown in the previously.

Secondly, the preferred direction of the endpoint determines the direction of the relational operator or comparator sign for the alternative hypothesis. Alternative hypothesis for smaller preferred values uses a less than sign, while larger preferred values uses a greater than sign. This, too, is exactly as they are specified for superiority testing as shown in the previously.

Finally, preferred direction of the endpoint determines the direction of the negligible difference, δ. Preferred smaller value endpoints imply that the negligible difference would be in a positive direction, i.e. $\delta > 0$. This is logical because negative values would add to the overall magnitude of the difference, Δ, and not be negligible at all. Conversely, larger value endpoints imply that the negative values, $\delta < 0$ could be negligible.

How NIM is used in the hypothesis formulation is provided in the next sections. Endpoints where smaller values indicate the researcher's preference are demonstrated first and larger will follow.

5.3 When Improvement is Smaller

When smaller values of the outcome variable indicate improvement, the null hypothesis is that the investigative treatment group is worse by the null value, δ or more.

Null Hypothesis:	Ho: $\Delta - \delta \geq 0$
Alternative Hypothesis:	Ha: $\Delta - \delta < 0$

(Hypotheses Pair 5.1)

When smaller values are preferred then larger values are inferior. The non-inferiority test is one-sided because there is interest in only one direction. The lower-case delta represents the null value. Recall that when this value is zero, the hypothesis test is for superiority.

The alternative hypothesis is that the investigative group is either better or at least no more than δ worse, when delta has a positive value, $\delta > 0$. The interest is that the treatment difference is no smaller than δ, when improvement is shown by small values of the outcome.

The test statistic for this hypothesis is:

$$T_L = \frac{\hat{\Delta} - \delta}{\sqrt{s_p^2\left(\frac{1}{n_e} + \frac{1}{n_s}\right)}}$$

(Equation 5.1)

The numerator of the test statistic is the left-side expression of the hypothesis formulation after replacement of the parameter with the estimate.

The s_p in the denominator is the pooled estimate of the common standard deviations across the two groups and is given as,

$$s_p = \left[\frac{(n_e - 1)s_e^2 + (n_s - 1)s_s^2}{(n_e + n_s) - 2}\right]^{1/2}$$

(Equation 5.2)

Non-inferiority is claimed if T_L is smaller (not larger) than the T-critical value because the alternative hypothesis has a less than sign. The T-critical value from the t distribution has the $(n_e + n_s) - 2$ degrees of freedom.

When smaller values of the endpoint are preferred, the alternative distribution is positioned on the left of the null, the alternative hypothesis uses a less than sign and the negligible difference would be positive and these tests are called lower-tailed.

5.4 When Improvement is Larger

When larger values of the outcome are desirable, the signs are reversed and the null hypothesis is that the investigative treatment group is worse by the null value, δ or more.

Null Hypothesis:	Ho: $\Delta - \delta \leq 0$
Alternative Hypothesis:	Ha: $\Delta - \delta > 0$

(Hypotheses Pair 5.2)

When larger values are preferred then smaller values are inferior. The hypothesis formulation is still one-sided and the left side is the same but be careful here there are two subtle changes. Firstly, the delta has a negative value, $\delta < 0$. Also, notice the relational operators have changed relative to the case when improvement is smaller.

The alternative hypothesis is that it is either better or at least no more than δ worse. The non-inferiority test is one-sided because there is interest in only one direction. The interest is that the treatment difference is no smaller than δ, when large values of the outcome are desirable.

The expression of the alternative hypothesis as $\Delta - \delta > 0$, specifically by moving the δ to the left side leads to the t-test statistic to assess non-inferiority. In other words, whether the difference in means is above the lower limit of δ when delta has a negative value, $\delta < 0$.

The test statistic is the same but the value of delta is negative, $\delta < 0$.

$$T_L = \frac{\hat{\Delta} - \delta}{\sqrt{s_p^2\left(\frac{1}{n_e} + \frac{1}{n_s}\right)}}$$

(Still Equation 5.1)

Non-inferiority is claimed if T_L is larger (not smaller) than the T-critical value because the alternative hypothesis has a greater than sign. The T-critical value from the t distribution has the $(n_e + n_s) - 2$ degrees of freedom at $(1 - \alpha)$.

When larger values of the endpoint are preferred, the alternative distribution is positioned on the right of the null, the alternative hypothesis uses a greater than sign and the negligible difference would be negative and these tests are called upper-tailed.

Regrettably, some well-meaning authors of non-inferiority methods specify the direction of the null value of the NIM by using plus and minus operators (Mascha & Sessler, 2011). This is a lamentable practice only because it serves to confuse the SAS users. When they write $(-\delta)$, they do not refer to the hypothesis or the calculation of the statistic but only intend to emphasize that the direction of interest for the endpoint, which would be smaller. Likewise, they write $(+\delta)$ when they intend to emphasize that the direction of interest for the endpoint is larger.

5.5 A More Informative Way

More simply, when smaller values are desirable, non-inferiority can be claimed when the upper limit of the estimated confidence interval (CI) is below δ.

$$Upper\ CI\ Limit = \left(\hat{\Delta} - \delta\right) + t_{(1-\alpha, n_e + n_s - 2\ df)}\left(\widehat{SE}_{\hat{\Delta}}\right)$$

(Equation 5.2)

The calculation of the confidence interval using (Equation 5.2) is shown in the SAS Data Step Snippet 5.1.

Snippet 5.1Calculation of the Confidence Interval

```
%let alpha = 0.05; %let sides = 1;
data b03;
     set sumout05;
     format critlev tcrit 8.3
            llci ulci 8.4
            tstat 8.4 pval 8.3;
     NullDiff = &meandiff;
     alpha = &alpha;
     sides = &sides;
     critlev = 1 - (alpha/sides);
     tcrit = tinv(critlev,df);
     llci  = ((invmean - cntmean) - NullDiff) -
              (tcrit * poolstderr);
     ulci  = ((invmean - cntmean) - NullDiff) +
              (tcrit * poolstderr);
     tstat = ((invmean - cntmean) -
               NullDiff)/poolstderr;
     pval = (1 - cdf('t',abs(tstat),df,0)) * sides;
     run;
```

The null value of the minimally tolerable difference which in this case will have a positive value $\delta > 0$, is known to shift the null distribution left and toward the alternative (See Figure 5.1).

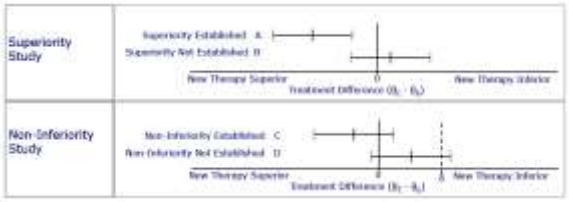


Figure 5.1 Comparison of Confidence Intervals for Superiority and Non-Inferiority Studies when Smaller Values are Preferred

A significant non-inferiority test (Hypotheses Pair 5.1) will coincide with the upper end of the estimated CI being below the specified δ (Walker & Nowacki, 2011).

When larger values are preferred, non-inferiority is claimed if the lower limit of the estimated confidence interval is above δ, when larger values are desirable.

Figure 5.2 Comparison of Confidence Intervals for Superiority and Non-Inferiority Studies when Larger Values are Preferred

A significant non-inferiority test (Hypotheses Pair 5.2) will coincide with the lower end of the estimated CI being above the specified δ.

Table 5.1. Directionality for Testing Statistical Hypotheses of Non-inferiority

		Preferred Direction of the Endpoint	
		Smaller	**Larger**
Superiority	Direction of the Alt Hypothesis	Less than, <	Greater than, >
	Side of Alt Hypothesis Distribution	Left of Null	Right of the Null
	Alt Confidence Intervals	Upper is below zero	Lower is above zero
Non-Inferiority	Direction of the Alt Hypothesis	Less than, <	Greater than, >
	Side on which the Alt Hypothesis Distribution	Left of Null	Right of the Null
	Direction of negligible difference	$\delta > 0$	$\delta < 0$
	Directional shift of the null toward the alternative $(\Delta - \delta)$	Left	Right
	One-sided Test Type	Lower-tailed	Upper-tailed
	Alt Confidence Intervals	Upper is below δ	Lower is Above δ

6. Estimating Power for Non-Inferiority Hypotheses

6.1 Quantification of the Type II Error

As previously shown, the probability of a Type II error depends on the value of alpha because alpha determines the critical value and was illustrated in Figure 3.1.

If the hypotheses are formulated as recommended, the μ_{null} value is $\Delta - \delta$ and would be centered at zero, but it doesn't matter. Figure 6.1 shows that the beta region shares a boundary with the alpha region.

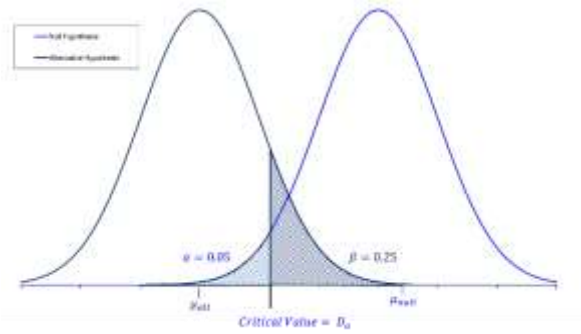


Figure 6.1. Distributions of a statistic, D, with variance, Σ^2, under the null hypothesis and under the alternative hypothesis and the Type I and Type II errors for a critical value of D_α when smaller values are preferred.

As the sample size increases, the spread of the curves in Figure 6.1 decrease and therefore beta decreases (power increases). The problem in planning a non-inferiority experiment is to determine the sample size, N, required such that testing H_{null} with significance level, alpha, that the probability of a Type II error, beta, is a desired, sufficiently-small level. To solve for N, six parameters are needed: α, β, μ_{null}, μ_{alt}, Σ_{null}^2, and Σ_{alt}^2.

The careful student of sizing studies with continuous endpoints, will recall that the assumption of equal variance, called homoscedasticity, is often reasonable and usually assumed. However, for this discussion two different variance parameters are specified, one under the null and one under the alternative hypotheses. This allows for generalization of the problem to binomial endpoints.

6.2 The Critical Regions are related by Sample Size

The sample size is N, that simultaneously satisfies $Pr(Z>Z_\alpha) = \alpha$ when H_{null} is true and $Pr(Z>Z_\beta) = 1-\beta$ when H_{alt} is true, where $Z = (X-\mu_0,)*\Sigma^{-1}$, is distributed by the standard normal distribution, and is the statistic one would use in testing.

Geometrically, the area of the two inequalities are illustrated in Figure 6.1, alpha is the critical region in light blue and beta in dark blue. Smaller values are preferred so the alternative distribution is to the left of the null. The alpha area is bordered and determined by the critical values of D at the alpha level of significance. This critical level for null distribution is given by, $D_\alpha = \mu_0 + Z_\alpha\Sigma_0$. The second dark blue region is bordered and determined by the same critical level but written in terms of the alternative distribution is, $D_\alpha = \mu_{alt} - Z_\beta\Sigma_1$, when $\mu_{alt} < D_\alpha$ and $D_\alpha = \mu_{alt} + Z_\beta\Sigma_1$ otherwise.

Both of these equations can be re-written to specify the distance as the difference between means as the sum of two parts (See Equation 6.1 below). Recall that distance is defined as the absolute values of the difference. The reason order doesn't matter is that the absolute value function renders subtraction communicative.

$$|\mu_{alt} - \mu_{null}| = |\mu_{alt} - D_\alpha| + |D_\alpha - \mu_{null}| \quad \text{(Equation 6.1)}$$

Values for the two sums on the right-hand side of (Equation 6.1) can be algebraically re-arranged and added together. Notice the critical value, D_α, drops out when alpha part is added to the beta part:

$$\begin{aligned} &\{|\mu_{alt} - D_\alpha| = Z_\beta\Sigma_1\} \\ &+\{|D_\alpha - \mu_{null}| = Z_\alpha\Sigma_0\} \\ &|\mu_{alt} - \mu_{null}| = Z_\beta\Sigma_1 + Z_\alpha\Sigma_0 \end{aligned} \quad \text{(Equations 6.2, 6.3, and 6.4)}$$

The result of the sum is (Equation 6.4). When working with an endpoint for which larger values are preferred, the algebra is the same even though the alternative distribution is to the right of the null (See Figure 6.2).

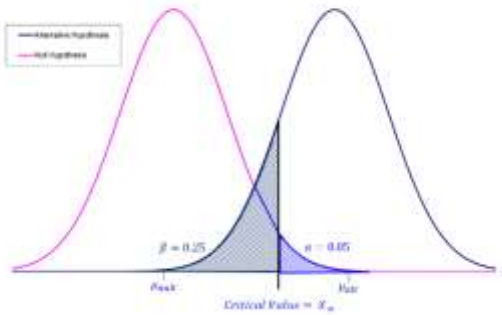


Figure 6.2. Distributions of a statistic, D, with variance, Σ^2, under the null hypothesis and under the alternative hypothesis and the Type I and Type II errors for a critical value of D_α when larger values are preferred.

With this result, all three tasks that we want to achieve are possible as will be seen in the next section.

7. Utility

The result in (Equation 6.4), $|\mu_{alt} - \mu_{null}| = Z_\beta\Sigma_1 + Z_\alpha\Sigma_0$, can be used to address three basic experimental design questions. Firstly, what sample size is required to ensure a specific power [(1-β)*100] of detecting a specific difference, μ_{alt}? Secondly, what is the power (in terms of Z_β) of the experiment in detecting a specific difference, μ_{alt} when a specific sample size, N, is used? Finally, what difference, μ_{alt}, can be detected with a given N.

7.1 Sample Size

Since it is known that as the sample size, N, increases, the spread of the curves decreases, as previously mentioned, an expression of the spread, Σ^2, is needed and it is often $\Sigma^2 = \sigma^2/N$, where σ^2 is the variance of the individual measurements and N is the total sample size. So, by substitution into Equation 2, the distance between means can be written in terms of N:

$$|\mu_{alt} - \mu_{null}| = Z_\alpha\sigma_0/\sqrt{N} + Z_\beta\sigma_1/\sqrt{N}. \quad \text{(Equation 7.1)}$$

Solving for N,

$$N = \left[\frac{(Z_\alpha\sigma_0 + Z_\beta\sigma_1)}{(\mu_1 - \mu_0)}\right]^2 \quad \text{(Equation 7.2)}$$

The order in the denominator will not matter (see Section 6.2).

7.2 Power

Solving Equation 6.4 for Z_β allows for the estimation of power.

$$Z_\beta = \frac{\sqrt{N}|\mu_1 - \mu_0| - Z_\alpha \sigma_0}{\sigma_1}$$

(Equation 7.3)

Power below 75-80% is generally considered underpowered.

7.3 Minimal Detectable Alternatives

Finally, μ_1, a minimally detectable alternative can be found, as shown in Equation 6:

$$\mu_1 = Z_\alpha \sigma_0/\sqrt{N} + Z_\beta \sigma_1/\sqrt{N} + \mu_0$$

(Equation 7.4)

Studies can be overpowered. In this case they have power to detect trivial differences that are below the minimal detectable alternative (See Figure 7.1).

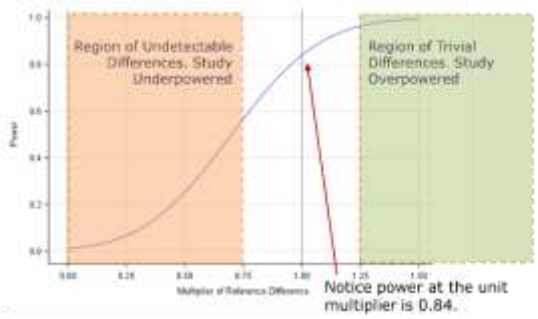


Figure 7.1. Regions of reasonably Distributions of a statistic, D, with variance, Σ^2, under the null hypothesis and under the alternative hypothesis and the Type I and Type II errors for a critical value of D_α when larger values are preferred.

7.4 Unequal Sample Size

An experiment with two groups can have and statistical procedures allow, by design or by practice, unequal group sizes for the interventional ($n_{inv} = Q_{inv}N$) and control groups ($n_{cnt} = Q_{cnt}N$), which are based on the group fractions $Q_{inv} + Q_{cnt} = 1$ and $n_{inv} + n_{cnt} = N$. For equal-sized groups, $Q_{inv} = Q_{cnt} = 0.5$ and $\left({Q_{inv}}^{-1} + {Q_{cnt}}^{-1}\right) = 4$.

7.5 Risk of Drop-outs

When there is a risk that only some fraction, R, of the sample size will complete the experiment or 'drop-out,' one adequate adjustment that has been suggested is:

$$N_d = N/(1-R)^2$$

(Equation 7.5)

Where N is the sample size calculated assuming no dropouts.

7.6 Applications

This approach affords a wide variety of application. It can be used with the one-sample mean, with paired observations, two independent groups with paired observations, single-group proportions, two independent proportions, two independent groups for correlational analysis and even survival analysis of two groups with censoring.

7.7 Small Sample Size Consideration

When using the t-distribution to approximate the normal distribution with small sample sizes, this approach is known to overestimate power slightly. An adjustment factors have been suggested (Cochran & Cox, 1964).

7.8. Sample Size Example

Consider an experiment where two samples are to be compared with equally sized groups. The measurement, Y, is assumed to be normally-distributed and variance of the measurement is assumed to be equal between the groups and known not to exceed 1.0. The difference to be detected is 0.20 with 90% power and a one-sided alpha level of 0.05.

Because the difference to be detected is 0.20 and that is positive, the relational operator for the alternative hypothesis is greater than and the preferred direction is larger. The hypotheses to be tested are in Hypotheses Pair 7.1.

Null Hypothesis:	Ho: $\Delta - \delta \leq 0.0$
Alternative Hypothesis:	Ha: $\Delta - \delta > 0.2$

(Hypotheses Pair 7.1)

The SAS code to find the Sample Size is provided in Snippet 7.1.

Snippet 7.1Sample Size Calculation

```
/*   Sample Size for the Test of   */;
/*   Two Independent Means         */;
data ss01;
     alpha = 0.05;
      sides = 1;
      beta  = 0.1;
      delta = 0.2;
      sigma = 1;
      q_e   = 0.5;
      q_c   = 0.5;
      z_alpha = probit(1-(alpha/sides));
      z_beta  = probit(1-beta);
      alloc   = (1/q_e) + (1/q_c);
      sqrt_n  = sigma*(z_alpha +
     z_beta)/delta;
      ntotal  = ceil(sqrt_n*sqrt_n) * alloc;
      put z_alpha z_beta ntotal;
      run;
```

The partial log for this data step is:

```
1.644853627 1.2815515655 860
NOTE: The data set WORK.SS01 has 1 observations and 12
variables.
NOTE: DATA statement used (Total process time):
      real time          0.10 seconds
      cpu time           0.04 seconds
```

A sample size of 430 per group would provide 90% with a one-sided, alpha level of 0.05 power to detect a difference of 0.20, assuming a drop-out rate of R=0 and that such a difference truly exists.

7.9 Power Procedure in SAS/STAT

For the upper one-sided case of continuous variables, the SAS/STAT documentation states that the exact power formulae are from O'Brien, R. G., and Muller, K. E. (1993). Equation 7.6 provides the one-sample case and Equation 7.7 provides the two-sample case. They are given below.

Upper One-Sided, One-Sample:

$$Power = P\left\{t(N-1,\delta) \geq t_{(1-\alpha)}(N-1)\right\}$$

(Equation 7.6)

$$t = N^{\frac{1}{2}}\left(\frac{\bar{x}-\mu_0}{s}\right)$$

It is stated in the documentation that solutions for sample size (N), alpha, and delta are obtained by numerically inverting the power equations (See SAS/STAT User Guide, Chapter 90, page 7366).

Upper One-Sided, Two-Sample:

$$Power = P\left\{t(N-2,\delta) \geq t_{(1-\alpha)}(N-2)\right\}$$

(Equation 7.7)

$$t = N^{\frac{1}{2}}(\omega_1\omega_2)^{\frac{1}{2}}\left(\frac{\bar{x}_2-\bar{x}_1-\mu_0}{s_p}\right)$$

(See SAS/STAT User Guide, Chapter 90, page 7384).

8. Example from Surgical Medicine

8.1 Clinical Background

A client requested an evaluation of the statistical power for a study design comparing two surgical techniques to repair ruptured, zone 2, flexor tendons. In the human, the flexor tendon connects to the muscles responsible for moving the fingers and thumb. Zone 2 is a particularly challenging location for primary repair in the human due to the presence of some complex anatomy.

The traditional technique of using sutures to repair tendons was first documented by the Greek physician Galen of Pergamum during the 2nd century AD following an attempted repair on a lacerated tendon of the foot (1). Now it has become the world-wide standard of care. However, suture repair requires one of the most intensive training curriculums of all surgical skills and the number of properly trained experts is limited, especially in the developing world.

The experimental technique uses an implant and deployment device that takes advantage of intervening 1,800 years of advancements in biomaterials and perioperative protocols. Expertise with the device requires far fewer years of training than suture repair. If it could be shown that post-operative mobility of the repaired digit was no worse than the reference technique, access to care could be expanded.

Enrollment for this study is planned for the African continent where tendon ruptures from accidents with knives, machetes, and sickles are common and access to intensively-trained, hand surgeons can be limited. Rigorous regulatory review, however, will be conducted at the US-FDA.

One of the two techniques will be randomly assigned to each of the 72 subjects in a one-to-one allocation ratio. The one-sided, alpha (type I) error rate will be 0.025.

8.2 Mobility Metric

Mobility can be measured using Strickland's Revised score (SRS). This score is comprised of four range of motion measurements of the proximal and distal interphalangeal joints, flexion and extension each, using a goniometer, proximal interphalangeal flexion (PIF), distal interphalangeal flexion (DIF), proximal interphalangeal extension deficit (PID) and distal interphalangeal extension deficit (DID). These measurements are used to calculate Strickland's Revised score using the following equation:

SRS = 100 * ((PIF + DIF)
- (PID + DID)).

(Equation 2)

SRS scores range from 0 to 175. Larger SRS scores are associated with improved clinical outcomes.

The units for SRS are angular degrees. The distribution of these scores in the population are assumed to be distributed approximately normal (that is, Gaussian, bell-shaped curve). Final scores will be assessed 24-weeks, post-surgery.

8.3 Determination of Non-Inferiority Margin

First, a decrease in 21.8 degrees was determined minimally acceptable. We arrived at that value by examining the associated classification system used by hand surgeons. To get the category, they calculate a percentage of target effectiveness by dividing the SRS by its total (175) and multiply that quotient by 100. They then classify repairs as: excellent (75-100%), good (50-74%), fair (25-49%), or poor (<25%) (Su et al., 2005), (Elliot & Giesen, 2013). The width of each of those intervals is 25%. Therefore, one-half of the category width (12.5%) could be considered sub-clinical and is a minimally acceptable difference. The value in degrees is back-calculated as the product of half the width of the category times the conversion factor from percent to degrees of seven-fourths.

8.4 Observed Performance of the Reference Group

Secondly, a systematic literature search was recommended and conducted. Primary research articles from medical journals identified in that search were examined for quality. Only five studies were adequate and they represented (n=525) repaired digits. The traditional method was associated with an average SRS mobility score of 99.7 and standard deviation of 31.3 degrees. Based on pre-clinical studies, the surgeon's opinion was that the true average mobility of the subjects using the experimental technique is similar (99.7 +/- 31.3 degrees).

8.5 Elements for Hypothesis Testing Framework

The research objective is to compare two surgical techniques statistically. Hypothesis testing is the quantification of that objective. The default conclusion is the null hypothesis is that the mobility scores for the experimental technique are strictly inferior to those of traditional technique. The usual alternative hypothesis is to rule out inferiority and conclude that the mobility scores for the experimental technique are superior to those of traditional technique. But that is not necessary in this study. It is sufficient that the alternative hypothesis is to rule out inferiority that the mobility scores for the experimental technique are not substantially worse to those of traditional technique. Therefore, a non-inferiority hypothesis testing framework, and not the more common superiority hypothesis framework, is appropriate. The 'not substantially worse' is quantified in the 'minimally acceptable decrease' as the 'non-inferiority margin' of 21.8 degrees.

Now we have all of the elements necessary (Table A) to build the hypothesis testing statements.

Table 8.1 Values of Hypothesis Testing Elements in the Hand Surgery Example

Element	Value
Study Objective	Non-inferiority
Primary Endpoint (PE)	SRS
Direction of Improvement	Larger values
Nature of PE	Continuous
Control Group Mean	99.7
Standard Deviation	31.3
Investigational Mean	99.7
Non-Inferiority Margin	21.8
Alpha	0.025
Sides	1
Sample Size per group	36

So, we can write the hypothesis statements, as follows:

Null Hypothesis: Ho: $\mu_{Inv} - \mu_{Cnt} \leq -21.8$

Alternative Hypothesis: Ha: $\mu_{Inv} - \mu_{Cnt} > -21.8$

(Hypotheses Pair 8.1)

The alternative hypothesis in this set of statements is one-sided, or one-tailed. Specifically, the null hypothesis will be rejected for larger values than the non-inferiority margin and the rejection region will be in the 'upper tail' of the sampling distribution. When the data have been collected, the analysis involves generating a one-sided confidence interval in the direction of the alternative hypothesis (See Section 2.6 above). That is, the confidence interval has a fixed lower limit but the upper limit is positive infinity.

9. Manual (Data Step) Method

Snippet 9.1 Manual Sample Size Calculation

```
97   data ss01;
98        alpha = 0.025;
99        sides = 1;
100       ntotal = 72;
101       delta = -21.8;
102       sigma = 31.3;
103       q_e   = 0.5;
104       q_c   = 0.5;
105       z_alpha = probit(1-(alpha/sides));
106       alloc   = (1/q_e) + (1/q_c);
107       sqrtalloc = sqrt(alloc);
108       absdelta = abs(delta);
109       z_beta  = (absdelta*sqrt(ntotal) ╥
                     z_alpha*sigma*sqrtalloc)/
                     (sigma*sqrtalloc);
110       beta = 1-probnorm(z_beta);
111       power = (1-beta)*100;
112       put z_alpha z_beta beta power ntotal;
113  run;

1.9599639845 0.9949742576 15.987439231 0.8401256077 72
NOTE: The data set WORK.SS01 has 1 observations and 14
variables.
```

```
NOTE: DATA statement used (Total process time):
      real time           0.02 seconds
      cpu time            0.03 seconds
```

10. The Power Procedure Method

In SAS/STAT, the Power and TTest procedures do not have explicit options for the Non-Inferiority study design with continuous outcomes (Castelloe & Watts, 2015). Does that seem strange? For example, the Freq procedure has NONINF and MARGIN= options for binary outcomes. You can use implicitly use Power and TTest procedures for the Non-Inferiority study design but be careful because you have to reverse the sign on the non-inferiority margin because the two procedures use different hypothesis testing statement formulations.

It is invocated using the NULLDIFF option, the direction of the NIM and the direction of the statistical test using the SIDES option. But you have to watch the sign and the position of the values in the groupmeans list.

In SAS/STAT, you can use the following statements to determine the estimated power:

Snippet 10.1 SAS procedure code to Estimate the Power of a Non-Inferiority Study with a Continuous Endpoint Using PROC POWER

```
ods output output=out01;
proc power;
     twosamplemeans
            test      = diff
            meandiff  = 0
            nulldiff  = -21.8
            sides     = u
            stddev    = 31.3
            alpha     = 0.025
            npergroup = 36
            power     = .
            ;
          run;
```

(Source: Table13_NonInfEff.sas).

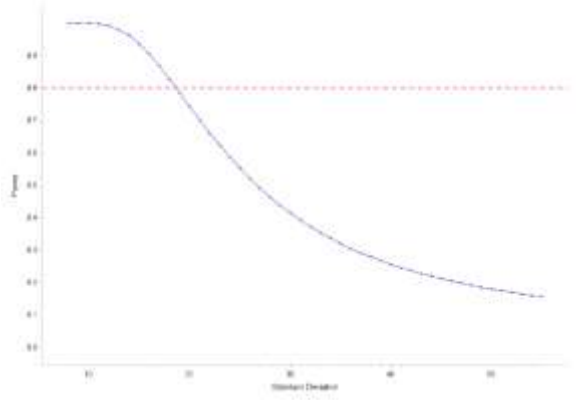
The output from the SAS/STAT statements in Snippet 10.1 are provided in Figure 10.1.

Figure 10.1. Output for Estimated Power for a Non-Inferiority Study with a Continuous Endpoint

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

Fixed Scenario Elements

Distribution                    Normal
Method                           Exact
Number of Sides                      U
Null Difference                 -21.8
Alpha                           0.025
Mean Difference                      0
Standard Deviation                31.3
Sample Size per Group               36

            Computed Power
                Power
                0.830
```

The effect of mis-specifying the standard deviation can be assessed by modifying the code from Snippet 10.1 to accommodate several values of the unknown standard deviation as follows: stddev = 8 to 55. These results are shown in Figure 10.2.

Figure 10.2. Power for Various values of the Standard Deviation



In addition, the effect of larger non-inferiority margins can also be assessed by using a number list for the NULLDIFF option. The power for various magnitudes of non-inferiority margins are provided in Figure 10.3.

Figure 10.3. Computed Power for Various Magnitudes of Non-Inferiority Margin

Obs.	NIM	Power
1	0	0.025
2	-5	0.098
3	-10	0.267
4	-15	0.518
5	-20	0.762
6	-25	0.916
7	-30	0.980

From Figure 10.3, it can be seen larger values of the non-inferiority margin are associated with larger power. Also, notice that if the NIM = 0, the computed power is the Type I error for the superiority design.

11. Example Simulation Program to Estimate Power

It is also possible to estimate the power of a non-inferiority design using simulations. There are several advantages to using a simulation method, including the ability to size

increasingly complex designs. An outline for conducting such a simulation is provided in Figure 11.1. Creating an outline before beginning a SAS program is a recommended practice.

This simulation generates and analyzes data for a continuous endpoint for a non-inferiority study design for both upper and lower tails. Equivalence is also simulated but the analysis steps below must be slightly modified to summarize the analysis. The simulation is achieved in nine steps (see Figure 11.1).

Figure 11.1 Outline of Simulation Program to Estimate Power

1. Parameter Inputs for Each Scenario
2. Create Macro Variables for Each Row
3. Simulate the Data using the Rand Function
4. Summarize the Simulations with Descriptive Statistics
5. Structure the Summaries as One Observation per Rep
6. Calculate the pooled standard error for each Rep
7. Calculate the confidence interval for the difference
8. Score the Simulations
9. Tally the Simulations

Figure 11.2 Outline of Simulation Program to Estimate Power

Snippet 11.1 Parameter Inputs – Enter One row observation per scenario

```
*************************************************************************;
*                                                                       *;
*    Input nine (9) parameters:                                         *;
*                                                                       *;
*    scenario provides an identification number.                        *;
*    invmean  Mean for the interventional arm.                          *;
*    cntmean  Mean for the control arm.                                 *;
*    stddev   Standard Deviation of the measurement.                    *;
*    nim      non-inferiority margin.                                   *;
*    alpha    Type I error rate (usually either 0.025 or 0.050).        *;
*    sides    Number of sides in the test.                              *;
*    invsize  The sample size for the investigational arm.              *;
*    cntsize  The sample size for the control arm.                      *;
*                                (Source: Table 14_NonInfEff.SAS)       *;
*                                                                       *;
*************************************************************************;

data ss01;
     format scenario 3.
            invmean cntmean nim 8.1
            alpha 8.3 sides 1. nim 8.4 invsize cntsize 8.;
     input  scenario invmean cntmean nim alpha sides invsize cntsize;
     cards;
1   99.7   99.7   -21.9   0.025   1   36   36
;
run;
```

Snippet 11.2 Create Macro Variables for some Parameters

```
data psp01;
     set ss01;
     format sd 8.3;
     if scenario = 1;
     call symput('invmean',put(invmean,8.1));
     call symput('invsize',put(invsize,8.));
     call symput('cntmean',put(cntmean,8.1));
     call symput('cntsize',put(cntsize,8.));
     call symput('nim',put(nim,8.1));
     call symput('alpha',put(alpha,8.3));
     call symput('sides',put(sides,1.));
     call symput('meandiff',put((invmean-cntmean),8.1));
     scenario = 1;
     do sd = 30 to 32 by 0.1;
        scenario = scenario + 1;
        output;
        end;
        run;
```

Snippet 11.3 Use the Rand Function to Simulate the Data

```
*************************************************************************;
*                                                                       *;
*       Data Simulation.                                                *;
*       Nobs = #scenarios * #rep * #trt * #subj.                        *;
*       Need to use an array when the sample size is unbalanced.        *;
*       Need to use an array when the means are different.              *;
*       Allow trt to be numeric for use in array.                       *;
*       then format trt with trtfmt. 0 = ref, 1 = inv.                  *;
*       During program development use rep = 1000.                      *;
*                                                                       *;
*************************************************************************;
data sim01;
     set psp01;
     format trt trtfmt.;
     array sampsize{2} (&cntsize &invsize);  /* needed when unbalanced */
     array mean{2}     (&cntmean &invmean);  /* needed when unequal    */

     seed = 20190328;
     call streaminit(seed);

     do rep = 1 to 10000;
        do trt = 0 to 1;
           do subj = 1 to sampsize{(trt+1)};
              x = rand('normal',mean{(trt+1)},sd);
              output;
              end;
              end;
              end;
              run;
```

Snippet 11.4 Summarize the results of the Simulations

```
proc sort data = sim01;
     by scenario rep;
     run;

ods output Summary = summ01;
proc means data = sim01;
     by scenario rep trt;
     var x;
     output out = sumout01
              n = n
              mean = mean
              stddev = stdev
              ;
              run;

proc sort data = sumout01;
     by scenario rep;
     run;
```

Snippet 11.5 Structure the data so there is one observation per Study

```
*************************************************************************;
*                                                                       *;
*    Manually transpose the dataset to create one                       *;
*    observation per study (rep).                                       *;
*                                                                       *;
*************************************************************************;
data sumout02;
     set sumout01;
     format cntmean invmean 8.1 cntstdev invstdev 8.2;
     retain cntn cntmean cntstdev;
     if trt = 1 then do;
     invn = n;
     invmean = mean;
     invstdev = stdev;
     end;
     else do;
     cntn = n;
     cntmean = mean;
     cntstdev = stdev;
     end;
     drop _type_ _freq_;
     run;
```

```
data sumout03;
     set sumout02;
     if trt = 1;
     keep scenario rep
          cntn cntmean cntstdev
          invn invmean invstdev;
     run;
```

Snippet 11.6 Calculate the pooled standard error

```
************************************************************************;
*                                                                      *;
*    Calculate the pooled standard error.                              *;
*    Also see McClave and Dietrick, page 346.                          *;
*    Reorder variables using PROC SQL.                                 *;
*                                                                      *;
************************************************************************;

data sumout04;
     set sumout03;
     format cntvar invvar 8.3
            poolvar poolstderr 8.4;
     cntvar = cntstdev*cntstdev;
     invvar = invstdev*invstdev;
     df         = cntn + invn - 2;
     poolvar    = ((invn - 1) * invvar + (cntn - 1) * cntvar)/df;
     poolstderr = sqrt(poolvar * (1/invn + 1/cntn));
     run;

proc sql;
     create table sumout05 as
     select
          scenario,
          rep,
          cntn, cntmean, cntstdev,
          invn, invmean, invstdev,
          df, poolvar, poolstderr
      from work.sumout04;
  quit;
```

Snippet 11.7 Calculate the confidence interval for the difference

```
************************************************************************;
*                                                                      *;
*    Calculate Confidence Interval for the difference.                 *;
*    Only one side is needed for non-inferiority.                      *;
*    Both sides are needed for equivalence.                            *;
*    Also see McClave and Dietrick, page 351.                          *;
*    P-value calculation.                                              *;
*    Could use probt but the cdf function is more flexible for         *;
*    other distributions.                                              *;
*                                                                      *;
************************************************************************;
data b03;
     set sumout0Snippet 11;
     format critlev tcrit 8.3
            llci ulci 8.4
            tstat 8.4 pval 8.3;
     NullDiff = &meandiff;
     alpha = &alpha;
     sides = &sides;
     critlev = 1 - (alpha/sides);
     tcrit = tinv(critlev,df);
     llci  = ((invmean - cntmean) - NullDiff) - (tcrit * poolstderr);
     ulci  = ((invmean - cntmean) - NullDiff) + (tcrit * poolstderr);
     tstat = ((invmean - cntmean) - NullDiff)/poolstderr;
     pval = (1 - cdf('t',abs(tstat),df,0)) * sides;
     run;
```

Snippet 11.8 Score the simulations

```
************************************************************************;
*                                                                      *;
*    Score the simulations.                                            *;
*    When improvement is smaller, use u < nim, for nim > 0.            *;
*    When improvement is larger,  use l > nim, for nim < 0.            *;
*    When testing equivalence, use both.                               *;
*                                                                      *;
************************************************************************;
data pe04;
     set b03;
     length lresult uresult eresult $13.;
     nim = &nim;
     if llci = . then lresult          = '             ';
     if llci <= nim then lresult       = 'Do Not Reject';
     else if llci > nim then lresult   = 'Reject       ';
     if ulci = . then uresult          = '             ';
     if ulci => nim then uresult       = 'Do Not Reject';
     else if ulci < nim then uresult   = 'Reject       ';
     if lresult = '             ' or
        uresult = '             ' then
        eresult = '             ';
     else if lresult = 'Do Not Reject' or
             uresult = 'Do Not Reject' then
             eresult = 'Do Not Reject';
     else if lresult = 'Reject       ' and
             uresult = 'Reject       ' then
             eresult = 'Reject       ';
     run;

proc sort data = pe04;
     by scenario;
     run;
```

Snippet 11.9 Tally the simulations

```
ods output OneWayFreqs = owf01;
proc freq data = pe04;
     by scenario;
     table lresult uresult eresult;
     run;
```

Figure 11.3 Tabulated Results. The SAS code snippets (11.1-9) were executed on my desktop which has an Intel® i7-6700 CPU with 16.0GB of physical RAM and running Microsoft Windows 10 Pro with SAS 9.4 installed. The program was run in interactive mode and executed in 2 minutes and 20 seconds.

MWSUG 2109, IN-113
Power Estimation in Non-Inferiority Study Designs
Example Power Simulation

Obs	scenario	invmean	cntmean	nim	alpha	sides	invsize	cntsize	sd	lresultp	uresultp	eresultp
1	2	99.7	99.7	-21.9000	0.025	1	36	36	30.000	86.79	0.00	0.00
2	3	99.7	99.7	-21.9000	0.025	1	36	36	30.100	85.97	0.00	0.00
3	4	99.7	99.7	-21.9000	0.025	1	36	36	30.200	86.40	0.00	0.00
4	5	99.7	99.7	-21.9000	0.025	1	36	36	30.300	85.65	0.00	0.00
5	6	99.7	99.7	-21.9000	0.025	1	36	36	30.400	85.96	0.00	0.00
6	7	99.7	99.7	-21.9000	0.025	1	36	36	30.500	85.25	0.00	0.00
7	8	99.7	99.7	-21.9000	0.025	1	36	36	30.600	85.09	0.00	0.00
8	9	99.7	99.7	-21.9000	0.025	1	36	36	30.700	84.69	0.00	0.00
9	10	99.7	99.7	-21.9000	0.025	1	36	36	30.800	84.63	0.00	0.00
10	11	99.7	99.7	-21.9000	0.025	1	36	36	30.900	84.82	0.00	0.00
11	12	99.7	99.7	-21.9000	0.025	1	36	36	31.000	83.65	0.00	0.00
12	13	99.7	99.7	-21.9000	0.025	1	36	36	31.100	83.93	0.00	0.00
13	14	99.7	99.7	-21.9000	0.025	1	36	36	31.200	83.68	0.00	0.00
14	15	99.7	99.7	-21.9000	0.025	1	36	36	31.300	83.72	0.00	0.00
15	16	99.7	99.7	-21.9000	0.025	1	36	36	31.400	82.92	0.00	0.00
16	17	99.7	99.7	-21.9000	0.025	1	36	36	31.500	82.71	0.00	0.00
17	18	99.7	99.7	-21.9000	0.025	1	36	36	31.600	82.65	0.00	0.00
18	19	99.7	99.7	-21.9000	0.025	1	36	36	31.700	81.76	0.00	0.00
19	20	99.7	99.7	-21.9000	0.025	1	36	36	31.800	81.90	0.00	0.00
20	21	99.7	99.7	-21.9000	0.025	1	36	36	31.900	82.20	0.00	0.00
21	22	99.7	99.7	-21.9000	0.025	1	36	36	32.000	81.46	0.00	0.00

CONFIDENTIAL, DRAFT
Program Name - \\...\saspgm\Table14_NonInfEff.sas
Data Source - \\...\sasdata\
Output Name - \\...\sasout\Table14_NonInfEff.pdf

12. Summary

The one-sided hypothesis testing framework was reviewed. Two methods for calculating power for the Non-Inferiority Study Design were provided including the Power Procedure in SAS and example SAS program for simulations. The estimation of power for this example is similar among the manual calculation, the Power Procedure and to the simulation result.

13. References

Chow, S.-C., Chang, M., & Pong, A. (2005). Statistical Consideration of Adaptive Methods in Clinical Development. Journal of Biopharmaceutical Statisics, 15, 575–591.

Elliot D, Giesen T. Primary flexor tendon surgery: the search for a perfect result. Hand clin 2013;29(2):191-206.

Goldstein, D. J., & Wilson, M. G. (1993). Adverse event frequencies generate hypotheses of efficacy and safety. *Clinical Pharmacology and Therapeutics*, *54*(3), 245–251.

Hogg, R. V. (1979). Statistical Robustness: One View of Its Use in Applications Today. *The American Statistician*, *33*(3), 108–115.

McClave, J. T., & Dietrich, F. H. I. I. (1985). Statistics. San Francisco: Dellen Publishing Company, Division of Macmillan, Inc.

O'Brien, R. G., and Muller, K. E. (1993). "Unified Power Analysis for t-Tests through Multivariate Hypotheses." In Applied Analysis of Variance in Behavioral Science, edited by L. K. Edwards, 297–344. New York: Marcel Dekker.

SAS Institute Inc. 2016. SAS/STAT® 14.2 User's Guide. Chapter 90, The Power Procedure. Cary, NC: SAS Institute Inc.

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. http://doi.org/10.1007/BF01068419

Su BW, Solomons M, Barrow A, et al. Device for zone-II flexor tendon repair. J Bone Joint Surg Am 2005;87:923-35. Doi: 10.2106/JBJS.C.01483.

Wilson, M. G. (2000). Lilly Reference Ranges. In S. S-C (Ed.), *Encyclopedia of Biopharmaceutical Statistics*. New York: Marcel Dekker, Inc.

Wilson, M. G. (2010). Assessing and Modeling Time to Event Data with Non-Proportional Hazards. *Procedings of the Mid-West SAS Users Group*, Paper 125–2010.

Blackwelder, W. C. (1982). 'Proving the Null Hypothesis' in Clinical Trials. *Controlled Clinical Trials*, *3*, 345–353.

Castelloe, J., & Watts, D. (2015). Equivalence and Noninferiority Testing Using SAS/STAT® Software. *Paper SAS1911-2015*, 1–23.

Cochran, W. G., & Cox, G. M. (1964). *Experimental Designs*. New York: Wiley.

Elliot, & Giesen. (2013). Primary flexor tendon surgery: the search for a perfect result. *Hand Clin*, *29*(2), 191–206.

Mascha, E. J., & Sessler, D. I. (2011). Equivalence and noninferiority testing in regression models and repeated-measures designs. *Anesthesia and Analgesia*, *112*(3), 678–687. https://doi.org/10.1213/ANE.0b013e318206f872

McClave, J. T., & Dietrich, F. H. I. I. (1985). *Statistics.* San Francisco: Dellen Publishing Company, Division of Macmillan, Inc.

Mukherjee, S. (2010). *The Emperor of All Maladies: A Biography of Cancer*. New York, New York: Scribner. Retrieved from https://archive.org/details/pdfy-cWnUvQsgyf0XYuPn/page/n3

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

Su, B. W., Solomons, M., Barrow, A., Senoge, M. E., Gilberti, M., Lubbers, L., … Rosenwasser, M. P. (2005). Device for zone-II flexor tendon repair. A multicenter, randomized, blinded, clinical trial. *The Journal of Bone and Joint Surgery. American Volume*. https://doi.org/10.2106/JBJS.C.01483

Walker, & Nowacki. (2011). Understanding Equivalence and Noninferiority Testing. *J Gen Intern Med*, *26*(2), 192–196.

14. Notices

TRADEMARK INFORMATION

EDUCATIONAL USE ONLY

ACKNOWLEDGMENTS

The author would like to thank the 2019 MWSUG Co-chairs, Jessica and Adrian, and the Industry Section Co-chair, Ge Guo, for accepting this manuscript.

CONTACT INFORMATION

Your comments and questions are encouraged.
Contact me at:

Michael G. Wilson
11630 Diamond Pointe Ct.
Indianapolis, IN 46236
317.620.0077
michael.g.wilson@biostat.us