SAMPLE SIZE CONVERSATION

by

Michael J. Symons, Ph.D.

Department of Biostatistics
University of North Carolina at Chapel Hill

Institute of Statistics Mimeo Series No. 1864

July, 1989

Sample Size Conversation

Michael J. Symons, Ph.D.

Department of Biostatistics

School of Public Health, CB# 7400

University of North Carolina at Chapel Hill

Chapel Hill, North Carolina 27599-7400

919-966-1107

ABSTRACT

A consultation on sample size provides an excellent opportunity to elaborate upon basic statistical ideas. And when the calculated sample size exceeds the client's constraints, statistical responses in a natural dialogue between client and statistician are illustrated.

KEY WORDS: Consulting, Hypothesis Testing, Detectable Difference

SHORT TITLE: Sample Size Conversation

1. Introduction and Methods

The guaging of the size of a research project at a statistical consulting session can heighten a client's interest in the statistical concepts of variance in observations, a detectable difference, and the possible errors in a testing situation. How these ideas affect the sample size determination should be discussed with the client rather than delivering results of calculations.

The comparison of two treatments, and their proportions of successes, provides an illustrative scenario. The paramount datum is the minimal difference in treatments that is clinically important. If the standard treatment is successful in about 50% of patients, suppose that for clinical recognition a new treatment will need to be successful in about 70% of patients. A schematic of 10 patients per treatment group makes explicit the study design, outcome measures, and the features of a clinically important result in the experiment, as presented in Table 1. Although the specification of the important difference to be detected is expected of the subject matter expert, the engineering type imputation of Freiman, Chalmers, Smith, and Kuebler (1) may be a useful soliciting strategy; namely, seek the client's reaction to a 25% or 50% improvement over the standard.

The variance of the outcome measure needs to be quantified by the statistician. With dichotomous outcomes the variance is a fairly stable function ($p(1-p)$) of the mean (p). However, when the outcome is a continuous measurement, a sample variance from a pilot study or the literature can be used to approximate the variance. Or, a calculation based upon the range of the measurement may be used. For example, if the range is b-a and normality of the observations is

reasonable, then the standard deviation can be approximated by $(b-a)/6$. With a uniform distribution, the variance can be very conservatively taken as $(b-a)^2/12$.

The central limit theorem assures the approximate normality of the proportion of successes in each treatment group. Then with

(a) two independent samples of equal numbers of patients, as provided by randomization,

(b) a Type I error of size α, the level of significance, and

(c) a Type II error of size β,

the required sample size for each group satisfies the equality

$$z_{1-\alpha}\sqrt{2p_o(1-p_o)/n} = \delta - z_{1-\beta}\sqrt{[p_o(1-p_o) + p_1(1-p_1)]/n} \qquad [1]$$

where

p_o is the proportion of successes with the standard treatment;

p_1 is the proportion of successes with the experimental treatment;

δ is the minimal difference, p_1-p_o, of clinical importance; and

z_ε is the 100ε percentile of the standard normal.

The left side of equation [1] is the critical difference in the proportions of treatment success determined by the size of the test, α, for p_o and sample size n. The right side is set equal to this critical difference so that the desired Type II error is assured for specified p_o, δ, and sample size n.

Notice that two elements of [1] differ from many presentations. First, $z_{1-\alpha}$ indicates a one-tail application while a two-tail test requires the larger percentile, $z_{1-\alpha/2}$. Although conservatism supports the use of the $100(1-\alpha/2)$ percentile, ethical comparisons with an established therapy can only entertain alternatives that may be improvements and thereby correspond to a one-tail test. Second, the variance under the null hypothesis is usually approximated by $2\bar{p}(1-\bar{p})$, where $\bar{p}$ is an average of p_o and p_1; see Schlesselman (2). For a comparison with an established therapy, $2p_o(1-p_o)$ is the proper null hypothesis variance expression and yields a simplier result later; see the discussion with equation [6].

The literature on sample size determination is large. Fortunately, surveys by Lachin (3) and Donner (4) provide comprehensive technical reviews of sample size estimation for clinical trials. In addition, Freiman, Chalmers, Smith and Kuebler (1) provides an excellent exposition of the key statistical concepts.

2. Results: Essence Of A Dialogue

A natural negotiation between client and statistician during a consultation on the scale of a research investigation will now be illustrated.

2.1 Sample Size

With interest in a clinical difference δ, and desiring power $1-\beta$ with level of significance α, relationship [1] can be solved for the required sample size, yielding

$$n = \frac{[z_{1-\alpha}\sqrt{2p_o(1-p_o)} + z_{1-\beta}\sqrt{p_o(1-p_o) + p_1(1-p_1)}\,]^2}{\delta^2} \quad . \qquad [2]$$

Using the proportions in Table 1 and a 0.05 level of significance, 104 subjects per treatment will detect a difference of 20% more successes by the experimental therapy with 0.90 power. Shown in Figure 1 are the two sampling distributions of the difference in proportions of successes for the treatments being compared, when the therapies each have 50% success and when the experimental treatment delivers 70% successes and the standard delivers 50% successes. Also shown are the critical difference in proportions of success and the tail areas corresponding to possible errors in the decision process. The two therapies are claimed as the same if the observed difference in the proportions of treatment success is less than 0.114. Larger differences suggest that the experimental therapy is better than the standard. There are two possible errors with this division of the sample space. The level of significance is 0.05, the chance of concluding that the experimental therapy is better when, in fact, the treatments are the same. The probability of a Type II error is 0.10, the chance of concluding that the treatments are the same when, in fact, the experimental treatment is 20% better.

2.2 Power

Sample sizes calculated by [2] "are often larger than what the investigator has in mind. When the size required is beyond practical limits, something has to give," as is stated so well by Freiman, Chalmers, Smith and Kuebler (1, p. 691). Suppose that 50 subjects per treatment is within the resources (time, money, staff) of the investigator. Relationship [1] can be solved for the power to detect the difference of interest at a specified level of significance. The

power is

$$1 - \beta = \Phi(z_{1-\beta}) \qquad [3]$$

where $\Phi(\cdot)$ is the cumulative distribution function for the standard normal distribution and

$$z_{1-\beta} = \frac{\delta - z_{1-\alpha}\sqrt{2p_o(1-p_o) / n}}{\sqrt{[p_o(1-p_o) + p_1(1-p_1)] / n}} . \qquad [4]$$

The situation with 50 subjects per treatment with the Table 1 scenario is portrayed in Figure 2. The power drops from 0.90 with 104 subjects to 0.64 with 50 subjects, due to the increased variance of each sampling distribution.

2.3 Detectable Difference

With such a large Type II error, 0.36, it may be informative to determine the clinical difference that could be detected with 50 patients per treatment and with more reasonable power. Substituting $p_o + \delta$ for p_1 in expression [1], yields a quadratic equation in δ ,

$$a\,\delta^2 + b\,\delta + c = o, \qquad [5]$$

where

$$a = n + z^2_{1-\beta}$$

and $$b = -2\left\{ z_{1-\alpha}\sqrt{2np_o(1-p_o)} + z^2_{1-\beta}\,(1/2 - p_o)\right\};$$

$$c = 2p_o(1-p_o)\,[\, z^2_{1-\alpha} - z^2_{1-\beta}\,].$$

The solution of [5] is

$$\delta = \frac{-b \pm \sqrt{b^2 - 4ac}}{2a}, \qquad [6]$$

using the required sign so that δ has the appropriate magnitude for the application. The closed form solution results from the use of the null hypothesis variance, $2p_o(1-p_o)$, for the difference in proportions. Iterative calculations are indicated by Lachin (3, p.96) when the variance $2\bar{p}(1-\bar{p})$ is used, $\bar{p}$ being an average of p_o and p_1.

At the 0.05 level of significance, 50 patients per treatment, and desiring 0.90 power, the experimental therapy would have to provide 78.2% (=50% + 28.2%) successes when the standard therapy helped 50% of the patients. Displayed in Figure 3 are the sampling distributions, cut point, and tail areas for the desired errors with the comparable variances of Figure 2, but with an increased separation. The detectable difference may be so large with the available resources and tolerable errors that to proceed with the study would be a waste of time, money, and effort. Such is a virtue of a thorough review of a study; harsh alternatives should be faced before the study is embarked upon. Again, see Freiman, et al. (1, p.691).

2.4 Error Sizes

Although the clinically important difference is the major determinant of the scope of an investigation, selection of Type I and Type II error sizes can also be reviewed. Decision theory offers some guidance for specifying the relative sizes of these errors, but the magnitude of the losses corresponding to these error is needed. Suppose that the loss incurred by making a Type I (II) error is A(B); then equating the risk of rejecting a true null hypothesis and the risk of accepting a false null hypothesis gives

$$\alpha A = \beta B. \qquad [7]$$

With the size of the test, α, specified and presuming that costs A and B are available, the size of the Type II error is given by [7]. However, except in business applications, the costs of these mistakes are difficult to assess on a common scale of measurement. For example in health applications, the units of A and B may be in dollars, in days of suffering saved, or in years of productivity preserved, and hence the impasse. Rejection of a worthwhile drug has a cost in human suffering not avoided. But promoting a new, but not better, drug has financial costs that are usually impossible to equate with human suffering or lost life.

But the spirit of a principle should not be lost because the requisite inputs may not be available. Equal losses suggest equal error sizes and higher costs should be balanced with smaller error sizes, as by [7]. Common specifications of Type I and Type II errors have β larger than α. Does this imply greater losses associated

with the financial costs of marketing a new, but not better, drug than with the cost of human suffering not averted by a new, better drug? Perhaps a case-study review of clinical trial plans may shed some light on the answer to this question or the historic precedent for such a pattern of choices.

Jerrell (5) provides computer software illustrating this idea by determining the best rejection region as the one minimizing the sum of the two risks equated in [7]. For the one tail presentation and with $\alpha = \beta$, then $z_{1-\alpha} = z_{1-\beta}$ in equation [1]. Denoting by ε the common error size,

$$\varepsilon = 1 - \Phi(z_{1-\varepsilon}) \quad , \qquad [8]$$

where

$$z_{1-\varepsilon} = \frac{\sqrt{n}\,\delta}{\sqrt{2p_o(1-p_o)} + \sqrt{p_o(1-p_o) + p_1(1-p_1)}} \quad . \qquad [9]$$

3. Discussion and Summary

The essence of the consulting dialogue presented above expands upon the list of three questions presented by Lachin (3), adding a discussion of the choice of error sizes, and provides a closed form solution for calculating a detectable difference with proportions. Further, the posture is in consultation with a subject matter expert at the planning stage of an investigation. Finally, attention is drawn to two prototypes for conveying the statistical concepts of

hypothesis testing to one's clients or students of introductory statistics courses. The split-figure presentation in Dixon and Massey (6) clearly separates the sampling situations as under the null hypothesis or under the alternative hypothesis. Schor (7, p.162) also uses this presentation very effectively to keep distinct the areas for each error type and yet shows the sample space cut-point(s) as common to each sampling distribution. And, the pedagogy of Freiman, et al. (1) in their "The Problem" section is excellent, especially for the statistical novice, but suitable for all levels of statistical edification.

A common layout of possible outcomes in a statistical testing situation can be amplified slightly and provide a useful overview for a client. Clear statements of the null and alternative hypotheses are needed, including the minimal clinical effect of interest. For each of four logical outcomes, three entries are provided in Table 2: (i) a statement of whether or not the decision is correct and the associated loss; (ii) the probability of each decision given the State of Nature; and (iii) the usual terminology for the outcome.

3.1 For Sale: Statistical Significance

An investigator should realize that many null hypotheses are strawmen and with a large enough sample size they can be rejected. Setting the left side of equation [1] equal to a difference of interest and solving for n illustrates this point. The sample size so calculated puts the difference of interest at the lower edge of the rejection region for an upper tail test. The question then is whether statistical significance is worth the cost of the required sample size. The investigator's commitment should not exceed what

will detect a clinically important difference. For more on the tangle of sample size and statistical significance, see Royal (8), who provides a distinction between "statistical" and "practical" (or "clinical") significance as part of a comprehensive discussion of the meaning of statistical significance.

A range of sample size situations is presented by the three panels of Figure 4. Panel A represents a waste of resources, presenting highly statistically significant results for a difference that is too small to be of practical importance. A designed experiment is portrayed in Panel B, suggesting controlled error sizes for a difference that is clinically of interest. In Panel C, the sample size is too small to provide a reasonable Type II error, the concern of the Freiman, et al. (1) paper.

3.2 Variations

With large samples, the central limit theorem yields normal approximations to a variety of applications. The conversation of Section 2 could be adapted according to the details of those situations. For example, examining the difference between two means with equal variances in each group, given by Snedecor and Cochran (9 p.111ff), is even simplier than for a difference of two proportions. Lachin (3) and Donner (4) provide the requisite statistical details for a wide variety of applications.

A refinement worthy of special notice is a provision for unequal sample sizes. Although applications with variances of similar magnitudes are not unusual, some situations tend toward unequal sample sizes in each group, such as

(i) unequal variances in each group, suggesting a larger sample size allocated to the group with greater variance of measurement. Providing comparable precision for the mean of each group is a reasonable objective.

(ii) case-control comparisons where the number of cases is fixed. Additional resources spent on enrolling more controls is informative.

(iii) optimal allocation based upon differential costs of case and control measurements is nicely overviewed by Schlesselman (2), beginning with Walter (10).

Lachin (3) accommodates unequal group sizes for a variety of applications for which sample size or power calculations are desired.

As a final note related to the specific form [1] used for proportions, various approximations are noted by Donner (4, p.201). These can often simplify the calculations without substantial effect on the resulting sample size. On this same point and as an extension to more complex designs, the recent work by Rochon (11) describes sample size calculation within the context of the weighted least squares linear modeling, as per Grizzle, Starmer, and Koch (12). The approach should unify a long list of special cases as well as provide a framework for extensions to sample size calculations when additional design features are to be incorporated.

Selected References

1. Freiman JE, Chalmers TC, Smith Jr. H, Kuebler RR: The Importance of Beta, The Type II Error and Sample Size in the Design and Interpretation of the Randomized Control Trial. The New England Journal of Medicine, 299: 690-694, 1978

2. Schlesselman JJ: Sample size requirements in cohort and case-control studies of disease. American Journal of Epidemiology, 99:381-384, 1974

3. Lachin JM: Introduction to Sample Size Determination and Power Analysis for Clinical Trials. Controlled Clinical Trials, 2:93-113,1981

4. Donner A: Approaches to Sample Size Estimation in the Design of Clinical Trials - A Review. Statistics in Medicine, 3: 199-214, 1984

5. Jerrell ME: Computer Programs to Demonstrate Some Hypothesis Testing Issues. The American Statistician, 42: 80-81, 1988

6. Dixon WJ, Massey Jr. FJ: Introduction to Statistical Analysis. (Second Edition, Ch. 14) New York: McGraw-Hill, 1957

7. Schor SS: Fundamentals of Biostatistics. New York: G.P. Putnam's and Sons, 1968

8. Royall RM: The effect of sample size on the meaning of significance tests. The American Statistician, 40:313-315, 1986

9. Snedecor GW, Cochran WG: Statistical Methods (Sixth Edition). Iowa: State University Press, 1967

10. Walter SD: Determination of significant relative risks and optimal sampling procedures in prospective and retrospective comparative studies of various sizes. American Journal of Epidemiology 105: 387-397, 1977

11. Rochon J: The application of the GSK method to the determination of minimum sample size. Biometrics, 45:193-206, 1989

12. Grizzle JE, Starmer CF, Koch GG: Analysis of Categorical data by linear models. Biometrics, 25: 489-504, 1969

Acknowledgements

The Freiman, Chalmers, Smith, and Kuebler (1978) paper is the source of much of the stimulation for this work. Their "The Problem" section, written by Roy R. Kuebler, is an inspiration to pedagogic excellence and is loaded with common sense for a sample size determination. This work was supported by NIMH Center Grant (MH33127), Biostatistics Core.

Table 1. Pattern of Clinically Important Results

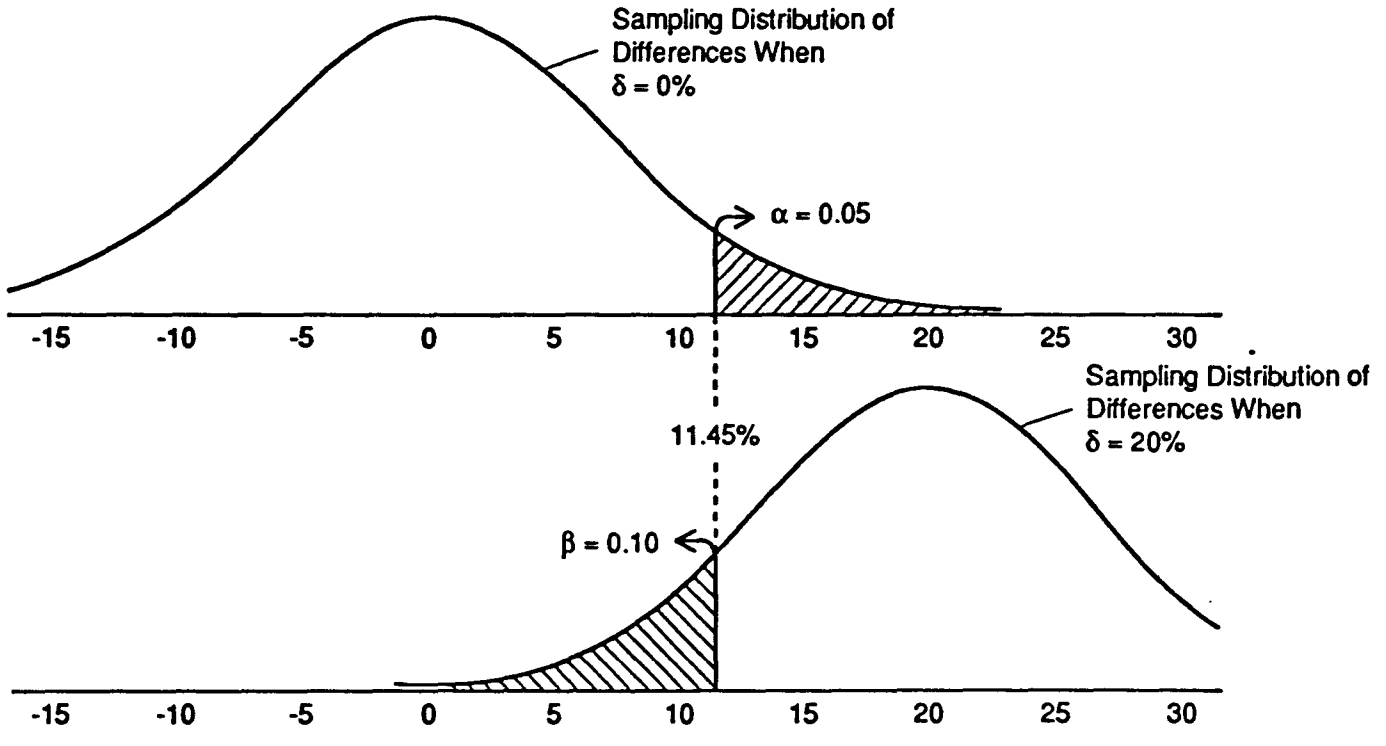
Treatment Group	Number of Patients by Outcome		Total	Proportion of Successes
	Success*	Failure		
Standard	5	5	10	50% (= p_o)
Experimental	7	3	10	70% (= $p_o + \delta$)

* Success is generic for a clinical definition of the outcome of interest for the treatment. Failure is "not success".

Table 2. Loss, Probability, and Terminology for Decisions by State of Nature

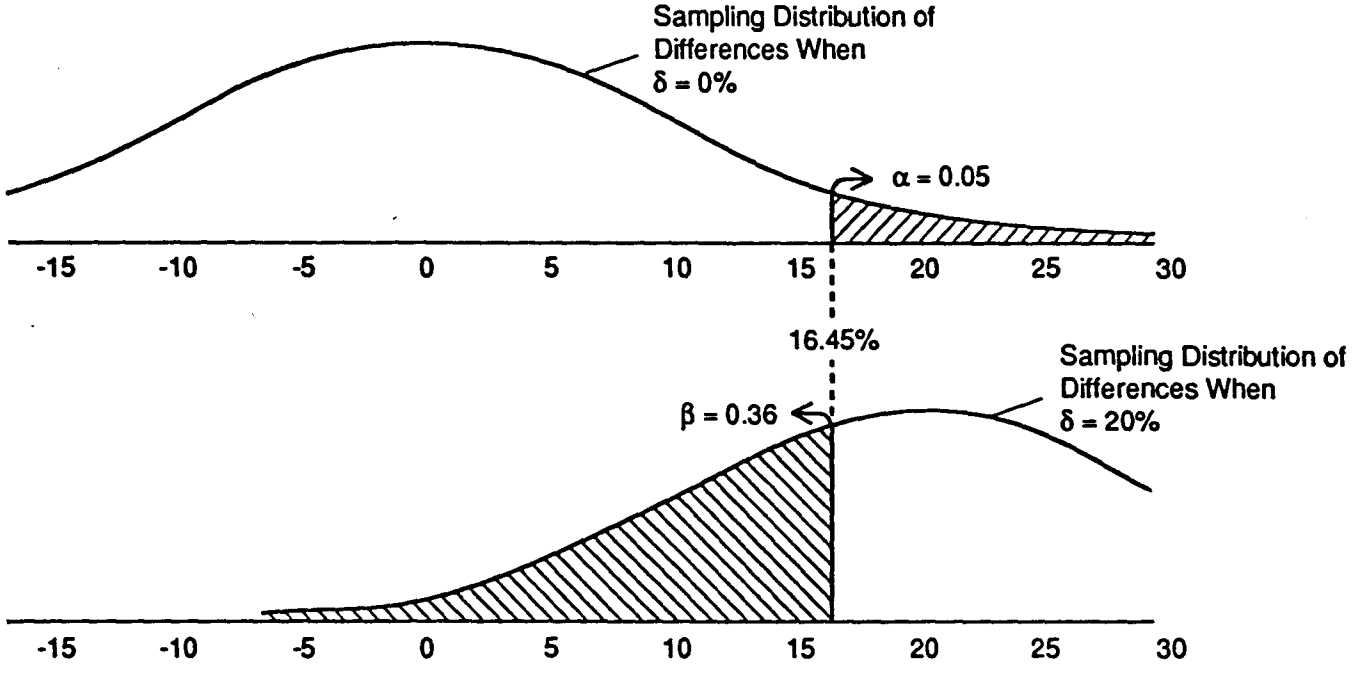
State of Nature	Statistical Decision: Accept H_o	Statistical Decision: Reject H_o
H_o is true	. correct decision: no loss . probability = $1 - \alpha$. specificity	. mistake: lose $\$_A$. probability = α . level of significance
H_o is false	. mistake: lose $\$_B$. probability = β . Beta error	. correct decision: no loss . probability = $1 - \beta$. power or sensitivity

Figure 1. Sampling Distributions of Difference in Treatment Proportions* of Success with 104 Subjects Per Treatment When $\delta = 0\%$ and When $\delta = 20\%$



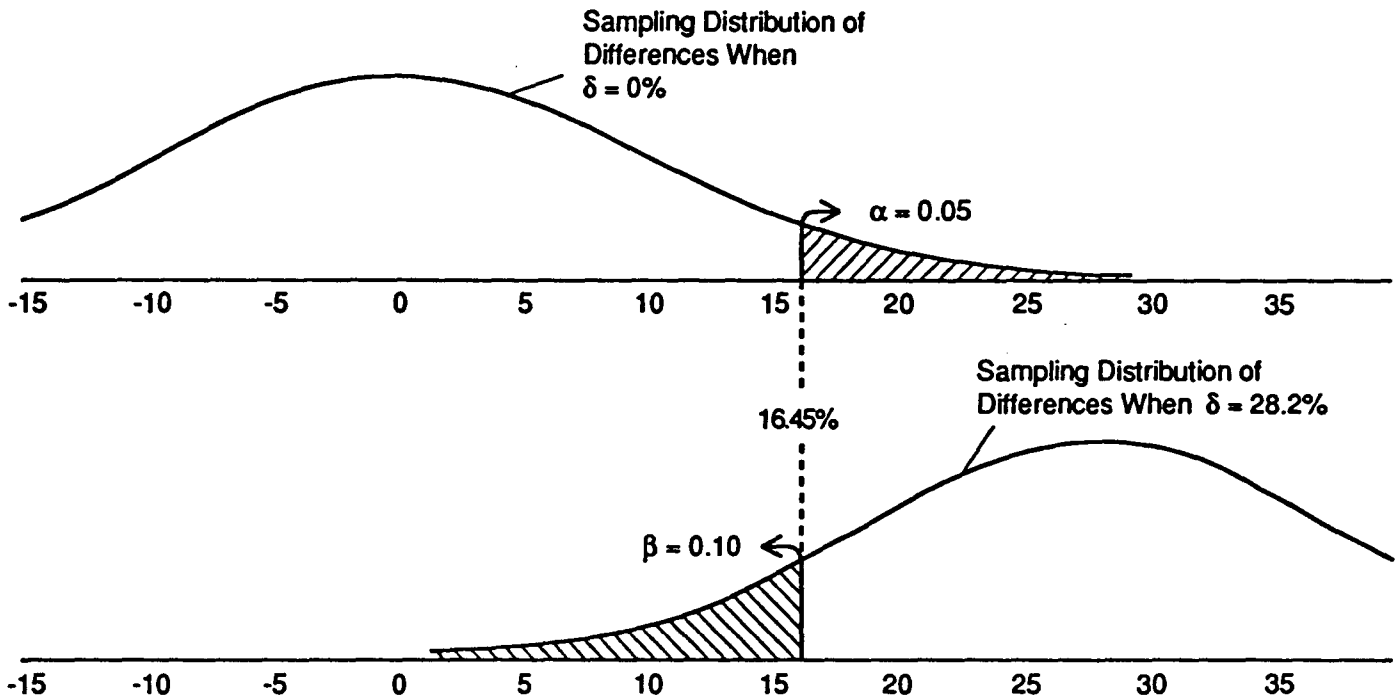
* Standard treatment has 50% success; see Table 1. The above abscissas are the difference in treatment proportions: $\hat{p}_1 - \hat{p}_0$ as a percentage.

Figure 2. Sampling Distributions of Differences in Treatment Proportions* of Success With 50 Subjects Per Treatment When $\delta = 0\%$ and When $\delta = 20\%$



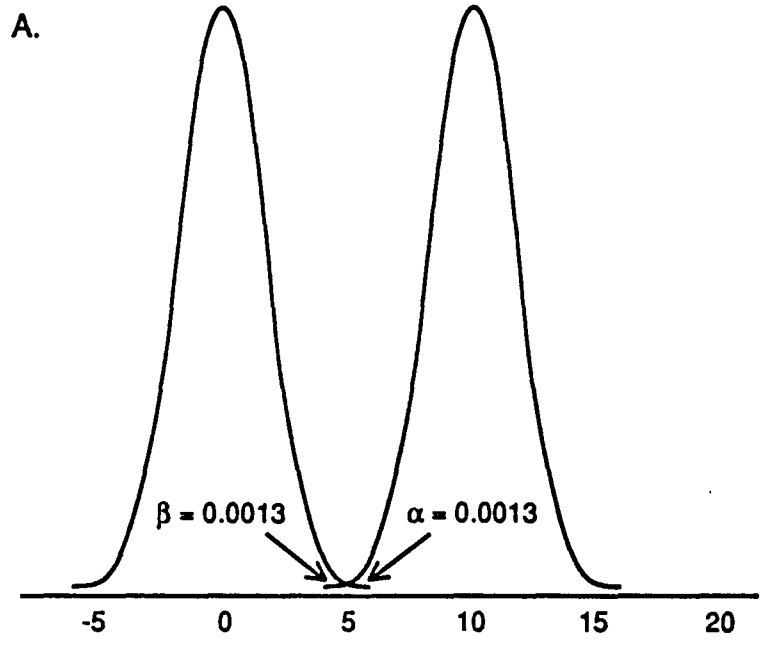
* Standard treatment has 50% success; see Table 1. The above abscissas are the difference in treatment proportions: $\hat{p}_1 - \hat{p}_0$ as a percentage.

Figure 3. Sampling Distributions of Differences in Treatment Proportions* of Success With 50 Subjects Per Treatment When $\delta = 0\%$ and When $\delta = 20\%$



* Standard treatment has 50% success; see Table 1. The above abscissas are the difference in treatment proportions: $\hat{p}_1 - \hat{p}_0$ as a percentage.

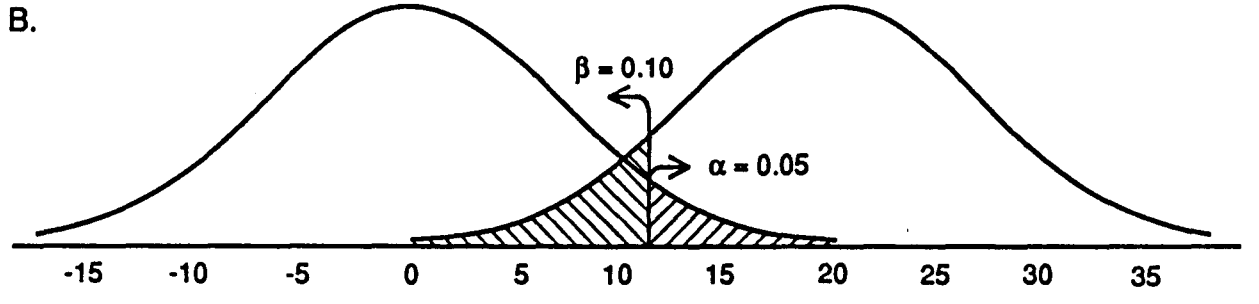
Figure 4. Range of Sample Size Situations With Clinically Important Difference of 20%; See Table 1

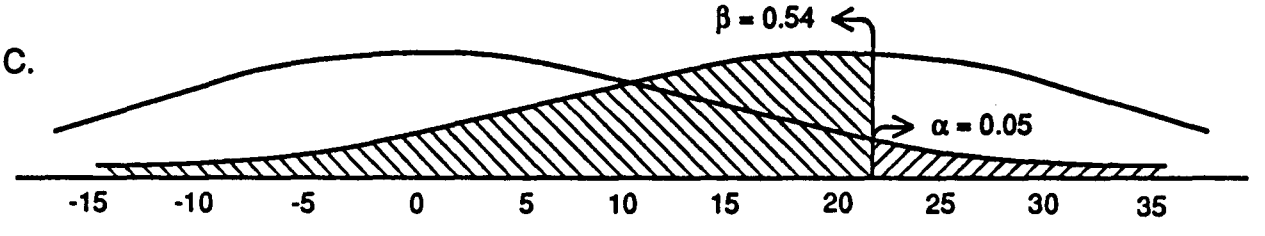


Panel A.
Sampling Distribution of Differences When δ = 0% and 10%. Sample Size is 1800.

Panel B.
Sampling Distributions of Differences When δ =0% and δ = 20%. Sample Size is 104.

Panel C.
Sampling Distributions of Differences When δ = 0% and δ = 20%. Sample Size is 30.





List of Figures