Sample Size Determination in Clinical Trials

HRM-733 CLass Notes

Lehana Thabane, BSc, MSc, PhD
Biostatistician
Center for Evaluation of Medicines
St. Joseph's Heathcare
105 Main Street East, Level P1
Hamilton ON L8N 1G6

thabanl@mcmaster.ca
Fax: (905)528-7386
Tel: (905)522-1155 x3720
http://www.lehanathabane.com

Assistant Professor
Department of Clinical Epidemiology & Biostatistics
Faculty of Health Sciences
McMaster University
Hamilton ON

November 19, 2004

Contents

1 Introduction

1.1 Learning Objectives

Specific Learning Objectives

- Learn about the important elements of a sample size calculation in the design of a clinical trial
  - Why is the sample size calculation important?
  - How to calculate the required sample size?
- Gain some knowledge about the basic statistical rules of thumb in sample size calculations
- Learn how to report the results of sample size calculation for a granting agency, research ethics board submission, etc.

1.2 Introductory Remarks

Sample size calculations form an integral part of a vast majority of quantitative studies. There are three main parts to sample size calculation: (i) sample size estimation, which depends on a host of items (see Section 3.1); (ii) sample size justification, which often involves justification of the calculated number in the light of budgetary and other biological considerations, and (iii) sample size adjustment, increasing the sample to account for potential dropouts or effect of covariates.

- Sample size calculations may not be required for some pilot or exploratory studies. It is important to note that
  - A pilot study is a preliminary study intended to test the feasibility of a larger study, data collection methods, collect information for sample size calculations, and therefore should always have a main study to which it leads.
  - In other words, pilot studies cannot exist on their own, but only in relation to a larger studies with the aim to facilitate the larger studies.
  - A pilot study SHOULD NOT be regarded as a study which is too small to produce a definitive answer to the question of interest.

- Sample size is just one part of a study design. There are several other parts that are needed to determine the sample size (see Section 4).
- As part of sample size discussions, it is crucial to know what the consequences of 'getting it wrong' are: these may be ethical, economic or scientific
- Sample size problems can be approached in two ways:
  - *Patients I need* approach: based on calculation of the sample size for a given power, level of significance and clinically meaningful difference
  - *Patients I can get* approach: based on calculation of power or detectable difference for a given sample size and level of significance

2 Why is sample size calculation important?

Sample size is important for two main reasons:

- Economic reasons: See Altman (1980),
  - An *undersized* study may result in a waste of resources due to their incapability to yield useful results. Recall that without a large enough a sample, an important relationship or effect/difference may exist, but the collected data be not be sufficient to detect it (ie the study may be under-powered to detect the effect).
  - An *oversized* study can result in unnecessary waste of resources, while at the same time yielding significant results that may not have much practical importance. Note that if a study is based on a very large sample, it will almost always lead to statistically significant results.
- Ethical reasons: See Altman (1980)
  - An *undersized* study can expose subjects to unnecessary (sometimes potentially harmful or futile) treatments without the capability to advance knowledge
  - An *oversized* study has the potential to expose an unnecessarily large number of subjects to potentially harmful or futile treatments
- Scientific reasons: (Moher et al (1994))

- If a trial with negative results has a *sufficient sample size* to detect a clinically important effect, then the negative results are interpretable–the treatment did not have an effect at least as large as the effect considered to be clinically relevant.
- If a trial with negative results has *insufficient power (insufficient sample size)*, a clinically important (but statistically nonsignificant) effect is usually ignored or, worse, is taken to mean that the treatment under study made no difference

Overall sample size calculation is an important part of the study design to ensure validity, accuracy, reliability and, scientific and ethical integrity of the study.

3 Approaches to sample size calculation

There are two major classical approaches to sample size calculations in the design of quantitative studies:

- Precision of estimation of an unknown characteristic/parameter of a Population
- Hypothesis testing of treatment effects/population parameters

3.1 Precision of Estimation: Precision Analysis

In studies concerned with estimating some parameter of a population (e.g. the prevalence of a medical condition in the population), sample size calculations are important to ensure that estimates are obtained with required precision/accuracy or level of confidence. Recall that the smaller the margin of error in the estimation, the more informative or precise the estimate is. For example,

- a prevalence of 10% from a sample of size 20 would have a 95% confidence interval (CI) of (1%, 31%), which may not be considered very precise or informative.
- However, a prevalence of 10% from a sample of size 400 would have a 95% CI of (7%, 13%), which may be considered more accurate or informative.

3.2 Hypothesis Testing of effects/relationships: Power Analysis

In studies concerned with detecting an effect (e.g. a difference between two treatments, or relative risk of a diagnosis if a certain risk factor is present versus absent), sample size calculations are important to ensure that if an effect deemed to be clinically meaningful exists,

then there is a high chance of it being detected, i.e. that the analysis will be statistically significant. If the sample is too small, then even if large differences are observed, it will be impossible to show that these are due to anything more than sampling variation. There are different types of hypothesis testing problems depending on the goal of the research. Let μ_S = mean of standard treatment, μ_T = mean of new treatment, and δ = the minimum clinically important difference.

1. *Test for Equality*: Here the goal is to detect a clinically meaningful difference/effects is such a difference/effects exists

2. *Test for Non-inferiority*: To demonstrate that the new drug is as less effective as the standard treatment (ie the difference between the new treatment and the standard is less than the smallest clinically meaningful difference)

3. *Test for Superiority*: To demonstrate that the new treatment is more superior that standard treatment (ie the difference between the new treatment and the standard is greater than the smallest clinically meaningful difference).

4. *Test for equivalence*: To demonstrate the difference between the new treatment and standard treatment has no clinical importance

Test for	Null Hypothesis	Alternative Hypothesis
Equality	$H_0 : \mu_T - \mu_S = 0$	$H_a : \mu_T - \mu_S \neq 0$
Non-inferiority	$H_0 : \mu_T - \mu_S \geq \delta$	$H_0 : \mu_T - \mu_S < \delta$
Superiority	$H_0 : \mu_T - \mu_S \leq \delta$	$H_0 : \mu_T - \mu_S > \delta$
Equivalence	$H_0 : \lvert\mu_T - \mu_S\rvert \geq \delta$	$H_0 : \lvert\mu_T - \mu_S\rvert < \delta$

It is important to note that

- the test for superiority is often referred to as the test for *clinical* superiority
- If $\delta = 0$, it is called the test of *statistical* superiority
- Equivalence is taken to be the alternative hypothesis, and the null hypothesis is nonequivalence

4 Information required to calculate a sample size

It is highly recommended that you ask a professional statistician to conduct the sample size calculation.

4.1 Factors that influence sample size calculation: A checklist

1. The objective(s) of the research: Is the research dealing with an estimation, hypothesis or equivalence testing problem?

2. Are the control and intervention(s) described in detail?

3. The outcome(s) of the research:

    - Is/Are the outcome(s) categorical or continuous?
    - Is it a multiple or single outcome study?
    - What is(are) the primary outcome(s)?
    - What is(are) the secondary outcome(s)?
    - Are the outcomes clinically relevant?
    - Can the outcomes be measured for all subjects?
    - Are the frequency and duration of the outcome measurements explicit?
    - Are there any surrogate outcomes?
        - What is the rationale for using surrogate outcomes?
        - Will they accurately reflect the main outcomes?
        - How can the observed benefit or harm made on surrogate outcomes translate into corresponding benefit or harm on the main outcome?

4. Are there any covariates or factors for which to control?

5. What is the unit of randomization? Is it individual subjects, family practices, hospital wards, communities, families, etc?

6. What is the unit of analysis? Is it individual subjects or clusters (eg family practices, hospital wards, communities, families)?

7. What is the research design? Is it

    - a simple randomized controlled trial (RCT)
    - a cluster randomized trial
    - an equivalence trial

- a non-randomized intervention study
- an observational study
- a prevalence study
- a study measuring sensitivity and specificity
- a paired design study (ie paired comparison)
- a repeated-measures design study (ie does your study include repeated measures)?

The following additional factors are equally important

- are groups of equal sizes?
- are the data hierarchical?

8. Research subjects
    - what is the target population?
    - what is the inclusion and exclusion criteria?
    - what is the likely patient compliance rate?
    - what is the baseline risk (poor or good prognosis)?
    - what is the chance of treatment response?
    - what is the potential drop-out rate?
9. How long is the duration of the follow-up? Is it long enough to be of any clinical relevance?
10. What is the desired level of significance?
11. What is the desired power?
12. What type of summary or test statistic will be used for analysis? Will it be a one- or two-tailed test?
13. The smallest difference (see Spiegelhalter and Freedman [9] and Spiegelhalter et al [8])
    - Does this reflect the degree of benefit from the intervention against the control over the specified time frame?
    - Is it stated as

- the smallest clinically important difference? (Lachin [12])
- the difference that investigators think is worth detecting? (Fleiss [13])
- the difference that investigators think is likely to be detected? (Halperin et al [10]).

14. Justification: Most importantly, is the justification provided on how the various prior estimates used in the calculations were obtained and their usefulness in the context of the study? This also deals with the clinical relevance of the estimates depending on the source (ie published data, previous work, review of records, expert opinions, etc).

5 Explanation of Statistical Terms

Below are some brief descriptions of some of above statistical terms.

1. *Null and alternative hypothesis*: Many statistical analyses involve the comparison of two treatments, procedures or subgroups of subjects. The numerical value summarizing the difference of interest is called the *effect*. In other study designs the effect may be

   - Odds ratio (OR): $H_0:$ OR=1
   - Relative risk (RR): $H_0:$ RR=1
   - Risk Difference (RD): $H_0:$ RD=0
   - Difference between means ($\mu_1 - \mu_2$): $H_0: \mu_1 - \mu_2 = 0$
   - Correlation coefficient (ρ): $H_0 : \rho = 0$

   Note that usually, the null hypothesis H_0 states that there is no effect and the alternative hypothesis that there is an effect.

2. *P-value of a test*: The p-value is the probability of obtaining the effect as extreme or more extreme than what is observed in the study if the null hypothesis of no effect is actually true. It is usually expressed as a proportion (e.g. p=0.001).

3. *Significance level of a test*: Also called the Type error probability, the significance level is a cut-off point for the p-value, below which the null hypothesis will be rejected and it will be concluded that there is evidence of an effect. The conventional significance level is $\alpha = 0.05$ or 5%.

4. *Power of a test*: Power is the probability

   - that the null hypothesis will be correctly rejected i.e. rejected when there is indeed a real difference or association
   - the test will detect a difference or association of a particular magnitude when it exists
   - $1-\beta$, where β is the P(Type II error), the chance of missing a clinically meaningful difference

   The higher the power, the lower the chance of missing a real effect. Power is typically set to be at least 80%.

	Reject H_0	Accept H_a
H_0 True	Type I error (Probability=α)	Correct Decision
H_0 False	Correct Decision	Type II error (Probability=β)

5. *Effect size of clinical importance*: This is the smallest difference between the group means or proportions (or odds ratio/relative risk closest to unity) which would be considered to be clinically meaningful. The sample size should be estimated so that if such a difference exists, then the chance that a statistically significant result would be obtained is very high.

6. *One-sided and two-sided tests of significance*: In a two-sided test, the null hypothesis states there is no effect, and the alternative hypothesis is that a difference exists in either direction. In a one-sided test the alternative hypothesis does specify a direction, for example that an active treatment is better than a placebo, and the null hypothesis then includes both no effect and placebo better than active treatment.

   - Two-sided tests should be used unless there is a very good reason for doing otherwise.
   - One-sided tests may be appropriate in situations where is it completely inconceivable that the results could go in either direction, or the only true concern with outcomes in one tail of the distribution.

     (a) Examples include:

i. toxicity studies

ii. safety evaluations

iii. analysis of occurrences of adverse drug reactions

iv. risk analysis

(b) References:

i. Bland JM, Altman DG. One and two sided tests of significance. *BMJ* 1994; 309: 248.

ii. Dubey SD. Some Thoughts on the One-sided and Two-sided Tests. *Journal of Biopharmaceutical Statistics* 1991; 1: 139-150.

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

The expectation that the difference will be in a particular direction is not adequate justification for one-sided tests.

6 Formulae for Sample Size Calculations

Table 1: Formulae for Sample Size Calculations for Comparisons Between Means

Design	Hypothesis	Hypotheses and Sample Size rules		
		H_0	H_a	Basic Rule
One-sample	Equality	$\mu - \mu_0 = 0$	$\mu - \mu_0 \neq 0$	$n = \frac{\left(z_{\frac{\alpha}{2}} + z_\beta\right)^2 \sigma^2}{(\mu - \mu_0)^2}$
	Superiority	$\mu - \mu_0 \leq \delta$	$\mu - \mu_0 > \delta$	$n = \frac{\left(z_\alpha + z_\beta\right)^2 \sigma^2}{(\mu - \mu_0 - \delta)^2}$
	Equivalence	$\lvert\mu - \mu_0\rvert \geq \delta$	$\lvert\mu - \mu_0\rvert < \delta$	$n = \frac{\left(z_\alpha + z_\beta\right)^2 \sigma^2}{(\lvert\mu - \mu_0\rvert - \delta)^2}$
Two-sample Parallel	Equality	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$	$n_i = \frac{2\left(z_{\frac{\alpha}{2}} + z_\beta\right)^2 \sigma^2}{(\mu_1 - \mu_2)^2}$
	Non-inferiority	$\mu_1 - \mu_2 \geq \delta$	$\mu_1 - \mu_2 < \delta$	$n_i = \frac{2\left(z_\alpha + z_\beta\right)^2 \sigma^2}{(\mu_1 - \mu_2 - \delta)^2}$
	Superiority	$\mu_1 - \mu_2 \leq \delta$	$\mu_1 - \mu_2 > \delta$	$n_i = \frac{2\left(z_\alpha + z_\beta\right)^2 \sigma^2}{(\mu_1 - \mu_2 - \delta)^2}$
	Equivalence	$\lvert\mu_1 - \mu_2\rvert \geq \delta$	$\lvert\mu_1 - \mu_2\rvert < \delta$	$n_i = \frac{2\left(z_\alpha + z_\beta\right)^2 \sigma^2}{(\lvert\mu_1 - \mu_2\rvert - \delta)^2}$
Two-sample Crossover	Equality	$\mu_1 - \mu_2 = 0$	$\mu_1 - \mu_2 \neq 0$	$n_i = \frac{\left(z_{\frac{\alpha}{2}} + z_\beta\right)^2 \sigma^2}{2(\mu_1 - \mu_2)^2}$
	Non-inferiority	$\mu_1 - \mu_2 \geq \delta$	$\mu_1 - \mu_2 < \delta$	$n_i = \frac{\left(z_\alpha + z_\beta\right)^2 \sigma^2}{2(\mu_1 - \mu_2 - \delta)^2}$
	Superiority	$\mu_1 - \mu_2 \leq \delta$	$\mu_1 - \mu_2 > \delta$	$n_i = \frac{\left(z_\alpha + z_\beta\right)^2 \sigma^2}{2(\mu_1 - \mu_2 - \delta)^2}$
	Equivalence	$\lvert\mu_1 - \mu_2\rvert \geq \delta$	$\lvert\mu_1 - \mu_2\rvert < \delta$	$n_i = \frac{\left(z_\alpha + z_\beta\right)^2 \sigma^2}{2(\lvert\mu_1 - \mu_2\rvert - \delta)^2}$

Table 2: Formulae for Sample Size Calculations for Comparisons Between Proportions

Design	Hypothesis	Hypotheses and Sample Size rules	
		H_0	Basic Rule
One-sample	Equality	$\pi - \pi_0 = 0$	$n = \frac{\left(z_{\frac{\alpha}{2}} + z_\beta\right)^2 \pi(1-\pi)}{(\pi-\pi_0)^2}$
	Superiority	$\pi - \pi_0 \leq \delta$	$n = \frac{\left(z_\alpha + z_\beta\right)^2 \pi(1-\pi)}{(\pi-\pi_0-\delta)^2}$
	Equivalence	$\lvert\pi - \pi_0\rvert \geq \delta$	$n = \frac{\left(z_\alpha + z_\beta\right)^2 \pi(1-\pi)}{(\lvert\pi-\pi_0\rvert-\delta)^2}$
Two-sample Parallel	Equality	$\pi_1 - \pi_2 = 0$	$n_i = \frac{\left(z_{\frac{\alpha}{2}} + z_\beta\right)^2 (\pi_1(1-\pi_2)+\pi_2(1-\pi_2))}{(\pi_1-\pi_2)^2}$
	Non-inferiority	$\pi_1 - \pi_2 \geq \delta$	$n_i = \frac{\left(z_\alpha + z_\beta\right)^2 (\pi_1(1-\pi_2)+\pi_2(1-\pi_2))}{(\pi_1-\pi_2-\delta)^2}$
	Superiority	$\pi_1 - \pi_2 \leq \delta$	$n_i = \frac{\left(z_\alpha + z_\beta\right)^2 (\pi_1(1-\pi_2)+\pi_2(1-\pi_2))}{(\pi_1-\pi_2-\delta)^2}$
	Equivalence	$\lvert\pi_1 - \pi_2\rvert \geq \delta$	$n_i = \frac{\left(z_\alpha + z_\beta\right)^2 (\pi_1(1-\pi_2)+\pi_2(1-\pi_2))}{(\lvert\pi_1-\pi_2\rvert-\delta)^2}$
Two-sample Crossover	Equality	$\pi_1 - \pi_2 = 0$	$n_i = \frac{\left(z_{\frac{\alpha}{2}} + z_\beta\right)^2 \sigma_d^2}{2(\pi_1-\pi_2)^2}$
	Non-inferiority	$\pi_1 - \pi_2 \geq \delta$	$n_i = \frac{\left(z_\alpha + z_\beta\right)^2 \sigma_d^2}{2(\pi_1-\pi_2-\delta)^2}$
	Superiority	$\pi_1 - \pi_2 \leq \delta$	$n_i = \frac{\left(z_\alpha + z_\beta\right)^2 \sigma_d^2}{2(\pi_1-\pi_2-\delta)^2}$
	Equivalence	$\lvert\pi_1 - \pi_2\rvert \geq \delta$	$n_i = \frac{\left(z_\alpha + z_{\beta/2}\right)^2 \sigma_d^2}{2(\lvert\pi_1-\pi_2\rvert-\delta)^2}$

6.1 Sample Size Adjustments

It is important to note that sample-size problems will vary from study to study depending on the context. The sample size may need to be adjusted to account for the effects of other variables, and the uncertainty of predictable practical and ethical factors.

- *Which variables should be included in the sample size calculation?*
  - The sample size calculation should relate to the study's primary outcome variable.
  - Ideally, separate sample size calculations should be provided for each *important variable*: the sample size should also be sufficient for the analyses of all important variables.
  - A simpler conservative approach is to estimate sample sizes for all important outcomes and then use the maximum estimate.
  - As a rule of thumb, when the correlation of a covariate with the response variable is ρ, then the sample size can be reduced by a factor of $1 - \rho^2$. That is,

$$n_{new} = n(1 - \rho^2).$$

- *Multiplicity and Sample Size Adjustment*: Multiplicity adjustment using the Bonferroni method to the level of significance should be made when at least one significant result (eg one of several primary outcomes or several pairwise comparisons) is required to draw a conclusion

- *Allowing for response rates and other losses to the sample*

  The sample size calculation should relate to the final, achieved sample. Therefore, the initial sample size may need to be adjusted in order to account for

  - the expected response rate
  - loss to follow up
  - lack of compliance
  - any other unforseen reasons for loss of subjects

  For example to adjust the sample size for the anticipated loss to follow-up rare: Suppose n is the total number of subjects in each group not accounting for loss to follow-up, and L is the loss to follow-up rate, then the adjusted sample size is given by

$$n_{new} = \frac{n}{1 - L}$$

  It is important to state clearly what factors were taken into consideration in the sample size adjustment and the justification should also be explicit.

- *Adjustment for Unequal Group Size*

- First calculate n the per group sample size assuming equal number per group
- If we require $n_1/n_2 = k$, then

$$n_2 = \frac{1}{2}n(1 + 1/k) \text{ and } n_1 = \frac{1}{2}n(1 + k)$$

- The question of whether to use unequal sample sizes matters when multiple sizes can be obtained in one group (see Lachin 2000, van Belle 2003)

7 Reporting the results of sample size calculation in the protocol

The protocol should provide sufficient details on how the sample size was determined. This should cover

1. clear statements of the (primary) objectives of the study
2. the desired level of significance
3. the desired power
4. type of summary or test statistic will be used for analysis
5. whether the test will one- or two-tailed
6. the smallest difference and a clear statement of whether it is
   - the smallest clinically important difference
   - the difference that investigators think is worth detecting
   - the difference that investigators think is likely to be detected
7. justification provided on how the various prior estimates of the variance and the effect used in the calculations were obtained and their usefulness in the context of the study
8. clear statements about the assumptions made about the distribution or variability of the outcomes
9. clear statement about the scheduled duration of the study
10. clear statements about how the sample size calculation was adjusted for
    - the expected response rate
    - loss to follow up
    - lack of compliance
    - any other unforseen reasons for loss of subjects
11. any other information that formed the basis for the sample size calculation.

8 Specific examples of samples size calculations

If your study requires the estimation of a single proportion, comparison of two means, or comparison of two proportions, the sample size calculations for these situations are (generally) relatively straightforward, and are therefore presented here. However, it is still strongly recommended that you ask a statistician to conduct the sample size calculation. The following examples (taken from St. George's Hospital Medical School website: http://www.sghms.ac.uk/depts/phs/guide/size.htm) are meant to illustrate how to calculate, justify and report sample size calculations.

8.1 Example 1: Comparing two proportions

- *Goal:* The following calculation only applies when you intend to compare two groups of the same size.
- *Scenario*: A placebo-controlled randomized trial proposes to assess the effectiveness of colony stimulating factors (CESS) in reducing sepsis in premature babies. A previous study has shown the underlying rate of sepsis to be about 50% in such infants around 2 weeks after birth, and a reduction of this rate to 34% would be of clinical importance.
- *Required information*:
  - Primary outcome variable = presence/absence of sepsis at 14 days after treatment (treatment is for a maximum of 72 hours after birth).
  - Hence, a categorical variable summarized by proportions.
  - Size of difference of clinical importance = 16%, or 0.16 (i.e. 50%-34%)
  - Significance level = 5%
  - Power = 80%
  - Type of test = two-sided

  The formula for the sample size for comparison of 2 proportions (two-sided) is as follows:

$$n = \frac{[z_{\frac{\alpha}{2}} + z_{\beta}]^2 \times [\pi_1(1-\pi_1) + \pi_2(1-\pi_2)]}{(\pi_1 - \pi_2)^2}$$

  where

  - n = the sample size required in each group (double this for total sample)
  - π_1 = first proportion=0.50,
  - π_2 = second proportion=0.34,
  - $\pi_1 - \pi_2$ = size of difference of clinical importance = 0.16
  - $z_{\frac{\alpha}{2}}$ depends on desired significance level = 1.96

– z_β depends on desired power = 0.84

Inserting the required information into the formula gives: -

$$n = \frac{[1.96 + 0.84]^2 \times [(0.50 \times 0.50) + (0.34 \times 0.66)]}{[0.16]^2} = 146$$

This gives the number required in each of the trial's two groups. Therefore the total sample size is double this, i.e. 292.

- *Suggested description of this sample size calculation*:

> "A sample size of 292 babies (146 in each of the treatment and placebo groups) will be sufficient to detect a clinically important difference of 16% between groups in the sepsis rate at 14 days, using a two-sided Z-test of the difference between proportions with 80% power and a 5% significance level. This 16% difference represents the difference between a 50% sepsis rate in the placebo group and a 34% rate in the treatment group."

8.2 Example 2: Comparing two means

- *Goal*: The following calculation only applies when you intend to compare two groups of the same size.

- *Scenario*: A randomized controlled trial has been planned to evaluate a brief psychological intervention in comparison to usual treatment in the reduction of suicidal ideation amongst patients presenting at hospital with deliberate self-poisoning. Suicidal ideation will be measured on the Beck scale; the standard deviation of this scale in a previous study was 7.7, and a difference of 5 points is considered to be of clinical importance. It is anticipated that around one third of patients may drop out of treatment (Guthrie et al. 2001)

- *Required information*:
  - Primary outcome variable = The Beck scale for suicidal ideation.
  - A continuous variable summarized by means.
  - Standard deviation = 7.7 points
  - Size of difference of clinical importance = 5 points
  - Significance level = 5%
  - Power = 80%
  - Type of test = two-sided

The formula for the sample size for comparison of 2 means (2-sided) is as follows: -

$$n = \frac{[z_{\frac{\alpha}{2}} + z_{\beta}]^2 \times 2\sigma^2}{\delta^2}$$

where n = the sample size required in each group (double this for total sample).

σ = standard deviation, of the primary outcome variable = 7.7.

δ = size of difference of clinical importance = 5.0.

$z_{\frac{\alpha}{2}}$= 1.96.

z_{β}= 0.84.

Inserting the required information into the formula gives: -

$$n = \frac{[1.96 + 0.84]^2 \times 2 \times 7.7^2}{5.0^2} = 38$$

This gives the number required in each of the trial's two groups. Therefore the total sample size is double this, i.e. 76.

To allow for the predicted dropout rate of around one third, the sample size was increased to 60 in each group, a total sample of 120.

- *Suggested wording of this sample size calculation*:

  > " A sample size of 38 in each group will be sufficient to detect a clinically important difference of 5 points on the Beck scale of suicidal ideation, assuming a standard deviation of 7.7 points, using a tow-tailed t-test of the difference between means, a power of 80%, and a significance level of 5%. The calculation is based on the assumption that the measurements on Beck scale are normally distributed. This number has been increased to 60 per group (total of 120), to allow for a predicted drop-out from treatment of around one third".

- *Wording from Power and Precision*

  Power for a test of the null hypothesis

  One goal of the proposed study is to test the null hypothesis that the two population means are equal. The criterion for significance (alpha) has been set at 0.050. The test is 2-tailed, which means that an effect in either direction will be interpreted.

  With the proposed sample size of 39 and 39 for the two groups, the study will have power of 80.8% to yield a statistically significant result.

  This computation assumes that the mean difference is 5.0 and the common within-group standard deviation is 7.7.

  This effect was selected as the smallest effect that would be important to detect, in the sense that any smaller effect would not be of clinical or substantive significance. It is also assumed that this effect size is reasonable, in the sense that an effect of this magnitude could be anticipated in this field of research.

  Precision for estimating the effect size

  A second goal of this study is to estimate the mean difference between the two populations. On average, a study of this design would enable us to report the mean difference with a precision (95.0% confidence level) of plus/minus 3.46 points.

  For example, an observed difference of 5.0 would be reported with a 95.0% confidence interval of 1.54 to 8.46.

  The precision estimated here is the median precision. Precision will vary as a function of the observed standard deviation (as well as sample size), and in any single study will be narrower or wider than this estimate.

9 Inappropriate wording or reporting of Sample size calculations

1. **Example 1:** "A previous study in this area recruited 150 subjects and found highly significant results (p=0.014), and therefore a similar sample size should be sufficient here."

   - *Why is this a problem?*: Previous studies may have been 'lucky' to find significant results, due to random sampling variation
   - *Solution:* Calculations of sample size specific to the present, proposed study should be provided

2. **Example 2:** "Sample sizes are not provided because there is no prior information on which to base them."

   - If the study is a preliminary pilot aimed at assessing feasibility or gathering the information required to calculate sample sizes for a full-scale study, then sample size calculations are not necessary
   - Where prior information on standard deviations is unavailable, then standard deviation can be estimated from the range as

     $$\text{Standard deviation} = \frac{\text{Max-Min}}{4}.$$

     Then sample size calculations can be given in very general terms, i.e. by giving the size of difference that may be detected in terms of a number of standard deviations

3. **Example 3:** "The throughput of the clinic is around 50 patients a year, of whom 10% may refuse to take part in the study. Therefore over the 2 years of the study, the sample size will be 90 patients. "

   - Although most studies need to balance feasibility with study power, the sample size should not be decided on the number of available patients alone.
   - If the number of available patients is a known limiting factor, a apply the *patients I can get* approach to indicate either
     (a) the power which the study will have to detect the desired difference of clinical importance, or
     (b) the difference which will be detected when the desired power is applied

   Where the number of available patients is too small to provide sufficient power to detect differences of clinical importance, you may wish to consider extending the length of the study, or collaborating with a colleague to conduct a multi-centre study.

4. **Other Examples**

   - "The results of a pilot study have been submitted for publication and the reviewers indicate that our sample size is adequate."
   - "We aim to recruit 100 participants. This sample size was determined to detect a small to moderate mean difference of xx points between the treatment groups on at least one of the key outcomes with a 70% power."

10 Important Remarks About Achieving the required sample size

10.1 Common Recruitment Strategies

Recruitment of suitable subjects for participation in clinical trials can also achieved through use of effective recruitment strategies (taken from the 2000 Office of Inspector General (OIG) of the US Department of Health Services Report):

1. through the use of financial and non-financial incentives
2. by physicians flagging patients in their practice through chart reviews or when they appear for an appointment
3. by furnishing trial information to other local clinicians or to disease advocacy and other groups
4. through advertising and promotion such as
    - media ads
    - press releases
    - televised segments
    - speakers at local health fairs

Recruitment through the Internet or Web in also increasingly becoming a popular option.

10.2 Reasons for failure to achieve the required sample size

The sample size required for a clinical trial may be very hard to recruit or recruitment has been much slower than anticipated. This is quite common in clinical studies. Sometimes this can lead to pre-mature ending of the trial, which could lead inconclusive findings because of lack of power. In order to avoid or address this problem, it is important to understand why it happen:

- patients' refusal to consent to participate in the study
- bad time of the study: snowy weather may also discourage potential patients from participating especially if the trial involves clinic visits
- adverse media publicity: sometimes adverse media publicity about medicine in general and trials in particular may discourage potential subjects from taking part in a research endeavor
- failure of recruiting staff to identify and approach potential research subjects

- lack of genuine commitment to the project: sometimes this which might be caused by honest doubts about the safety or efficacy of the new treatment
- poor recruitment may also be due to staffing problems: eg clinical unit is under pressure from excessive patient numbers or understaffing
- too many projects going after the same subjects

For further details on barriers to patient participation in clinical studies, see Ross et al 1999.

10.3 Possible or Common Solutions

Possible solutions include

- Pilot studies are very helpful in providing insights into some of these issues
- good planning: devise a clear plan of how to monitor recruitment. This may also help the proposal for funding since funders may well be impressed by a proposal which shows that this issue has been considered by applicants and some there are plans to deal with it
- request to the funders for an extension in time or for an extension in funding
- check with potential collaborators what their other trial commitments are
- maintain regular visits to trial sites and good contact with staff who are responsible for recruitment
- to have recruitment targets (milestones) to enable the research team to monitor how well recruitment is going, so that problems can be detected as soon as possible

11 Retrospective Sample Size Calculations

Sometimes, people try to estimate the sample size or perform power analysis after the study is completed.

- Avoid retrospective planning; it's bad science!
- IMPORTANT: "Observed power" (ie power calculated based on the observed effect) is a decreasing function of the p-value of the test (Hoenig and Heisey, 2001). That is,
  - the observed power increases as the p-value decreases
  - the higher the observed power, the greater the evidence *against* the null hypothesis
- The problem with observed power:

- It is associated with this common mis-interpretation or misconception: if the test is nonsignificant (pvalue is large), but the observed power is high, then this is interpreted to mean that there is strong evidence in support of the null hypothesis
- It causes great confusion because it is often used inappropriately to add interpretation to a non-significant test result

12 Important Rules of Thumb/Cautions

12.1 General

1. *Multiplicity and Sample Size Adjustment*: Multiplicity adjustment using the Bonferroni method to the level of significance should be made when at least one significant result (eg one of several primary outcomes or several pairwise comparisons) is required to draw a conclusion

2. *Overlapping Confidence Intervals Do not imply non-significance*

   Basic Rule: "Confidence intervals associated with statistics can overlap as much as 29% and the statistics can still be significantly different" (van Belle 2002)

3. *Sample size calculations should be based on the statistics used in the analysis.* Example, if sample size calculations were based on assumption that the outcome is continuous, then dichotomizing the outcome for the analysis would not be appropriate. Why?

   - using a different statistic for analysis may alter the anticipated power
   - the anticipated treatment effect may no longer be meaningful in the scale of the new statistic

4. The basic rule of thumb for estimating the sample size for testing equality of two means is
$$n_1 = n_2 = \frac{8\sigma^2}{\delta^2}; \text{ where } \delta = \mu_1 - \mu_2$$

5. The basic rule of thumb for estimating the sample size for testing equality of two proportions is
$$n_1 = n_2 = \frac{8\pi(1-\pi)}{(\pi_1 - \pi_2)^2}; \text{ where } \pi = \frac{\pi_1 + \pi_2}{2}$$

6. Since sample calculations are estimates,

   - *it is usually better to be conservative.* For example, it is usually better to assume a two-sided test than a one-sided test
   - *it is better to adopt a simple approach* even for complex problems. For example, it is simpler to use difference between proportions than logistic regression in sample size calculation

7. Although larger sample sizes are generally desired, it is important to *be aware of the statistical (over-power), ethical, and economic consequences of too large a sample*

8. *Rare Incidence rates*: If the primary outcome is an extremely rare event (eg. one per 10, 000), then sample size calculations will indicate that a very large sample is required

9. It is worth noting that *observational or non-randomized studies looking for differences or associations will generally require a much larger sample* in order to allow adjustment for confounding factors within the analysis

10. *It is the absolute sample size which is of most interest*, not the sample size as a proportion of the whole population

11. *Consistency with study aims and statistical analysis*

    - The adequacy of a sample size should be assessed according to the purpose of the study. Check whether the purpose of the study is
      - to test for no difference (equality)
      - to test for non-inferiority
      - to test for superiority
      - to test for equivalence

      Note that the sample size required to demonstrate equivalence will be larger than that required to demonstrate a difference.
    - Sample size calculations should relate to the study's stated objectives, and be based on the study's primary outcome variable or all important outcome variables
    - Sample size calculations should also be consistent with the proposed method of analysis

12. The following rules of thumb have been recommended by vanVoorhis and Morgan 2001:

    - For moderate to large effect size (ie $0.50 \leq$effect size≤ 0.80), 30 subjects per group are required
    - For comparisons between three or more groups, then to detect an effect size of 0.5
      (a) with 80% power, will require 14 subjects/group
      (b) with 50% power, will require 7 subjects/group

13. *Sensitivity Analysis*: It is best to create a sample size table for different values of the level of significance (α), power or different effect sizes, and then ponder the table to select the optimal sample size

12.2 Rules of Thumb for Relationships/Association

In regression problems, *power* refers to the ability to find a specified regression coefficient or level of R^2 statistically significant at a specified level of significance and specified sample size.

1. For multiple regression: Hair et al, 2000 state
   (a) that with 80% power, and $\alpha = 0.05$, one can detect a
      i. $R^2 \geq 0.23$ based on $n = 50$
      ii. $R^2 \geq 0.12$ based on $n = 100$
   (b) The general rule is that the ratio of number of subjects to number of independent variables should be about 5:1. There is substantial risk of "overfitting" if it falls below.
   (c) The desired ratio is usually about 15 to 20 subjects for each independent variable
2. Sample size for examining relationships: Green (1991) recommends
   (a) Rule 1: for testing multiple correlations
   $$n > 50 + 8m$$
   where m is the number of independent variables
   (b) Rule 2: for testing relationship of outcome with individual predictors
   $$n > 104 + m$$
3. Harris (1985) recommends
   (a) Rule 1: For 5 or less predictors, the number of subjects should exceed the number of independent variables by 50
   $$n > 50 + m$$
   (b) Rule 2: For equations involving 6 or more predictors, an absolute number of 10 subjects per predictor is recommended
   $$n > 104 + m$$
4. Large samples are needed (see Tabachnick and Fidell, 1996) if
   (a) the dependent variable is skewed
   (b) the effect size is small
   (c) there is substantial measurement error

(d) stepwise regression is used

5. Rules for chi-squared testing

   (a) Sample size should be such that no expected frequency in a cell should drop below 5: small expected cell frequencies can limit power substantially and inflate Type I error.

   (b) Overall sample size should be at least 20

   (c) The number of cells (ie degrees of freedom of the chi-squared test) is indirectly related with power (see Cohen (1988))

6. Rules for Factor Analysis

   (a) At least 300 cases/subjects (Tabachnick and Fidell, 1996)

   (b) At least 50 participants/subjects per variable (Pedhazur and Schmelkin, 1991)

   (c) Comrey and Lee (1992) guide:

   - n=50: very poor
   - n=100: poor
   - n=200: fair
   - n=300: good
   - n=500: very good

   (d) The higher the cases-per-variable ratio, the smaller the chance of "overfitting" (ie creating factors that are not generalizable beyond the specific sample)

13 Tips on Elicitation of Effect Sizes and Variances for Sample Size Calculations

The following tips are taken from "Some Practical Guidelines for Effective Sample Size Determination" by Lenth (Lenth, 2001)

1. Elicitation of information on effect sizes to calculate sample size: Important questions

   - What results do you expect (hope to) see?
   - Would an effect of half that magnitude [specify] be of any scientific interest?
   - Would a increase/decrease/difference of this magnitude [specify] be of any practical importance?
   - What is the range of clinical indifference?
   - If you were a patient, would the benefits of reducing/increasing the primary outcome [specify] by this magnitude [specify] outweigh the cost, inconvenience and potential side effects of this treatment?

2. Elicitation of information on standard deviations/variances to calculate sample size: Important questions

   - What is the usual range of the primary outcome?
   - Tell me about the smallest and largest values that you have seen?
   - Discuss stories about extreme observations to determine the extent to which they may represent ordinary dispersion
   - Do you have any studies that you have done or done by others on using this outcome? Do you have any historical or pilot data?
   - What are the possible sources of variation based on past studies?

3. Effect sizes can also be expressed in terms of the standard deviation. For example, the general sample size formula for comparing two means is

$$n = \frac{\left(Z_{\alpha/2} + Z_{\beta}\right)(\sigma_1^2 + \sigma_2^2)}{\delta^2}$$

   where $\delta = \mu_1 - \mu_2$. If $\sigma_1 = \sigma_2 = \sigma$, this reduces to

$$n = \frac{\left(Z_{\alpha/2} + Z_{\beta}\right) 2\sigma^2}{\delta^2}.$$

   If $\delta = k\sigma$ for some constant k, the formula can be rewritten as

$$n = \frac{2\left(Z_{\alpha/2} + Z_{\beta}\right)}{k^2}.$$

   Thus, your elicitation about the δ involves elicitation of k.

14 Sample Size Calculations for Cluster Randomized Controlled Studies

Cluster randomized designs are increasingly used in community healthcare interventions and health services research. Cluster randomized designs are designs in which intact social units or clusters of other units are allocated to treatment groups or interventions.

14.1 Reasons for Using Cluster-randomized Designs

Reasons for using cluster-randomized designs include (from Hutton 2001; Donner and Klar 2000):

- Scientific reasons

1. *Treatment contamination*: In cases where intervention is aimed at changing human behavior or knowledge transmission, cluster randomized designs are used to avoid contamination or influence of personal interactions among cluster members.
2. *Enhancing compliance/adherence*: Informal discussions about the intervention applied to a family practice, school, community setting, etc, might enhance subject adherence.
3. *Cluster level intervention*: Some interventions such as those aimed at family physicians can only be applied at the cluster level. The cluster effect or influence of cluster-level covariates is such that individuals within a cluster are often treated in a similar fashion or exposed to a common environment.
4. *Cluster action of an intervention*: Interventions such as vaccines, treatment of river blindness, applied at community level reduce the likelihood of infections or transmissions of diseases within the community. This is because infections tend to spread more quickly within communities/families than between communities/families.

- Logistical and political reasons

1. *Administrative convenience*: Using clusters facilitates the administration of a trial in many ways: Fewer units (clusters) to contact; access to patients through family practices is easier; recruitment and randomization of practices is easier and faster than that of patients.
2. *Political*: Sometimes community leaders, local or national leaders may have to provide permission before individual subjects within a community can be contacted for trial purposes
3. *Access to routine data:* Sometimes it's easier to randomize practices or communities in order to access relevant information

- Ethical Reasons

1. *Randomizing part of the family*: For trials that deal with vaccines or food interventions, it would appear unethical to randomize some members of a family or community to one intervention instead of the whole family.

14.2 Sample Size Formulae for Cluster-randomized Designs

Let

$$\begin{aligned} k &= \text{number of clusters} \\ m &= \text{average cluster size} \\ \rho &= \text{intra-cluster correlation coefficient} \\ IF &= 1+(m-1)\rho = \text{inflation factor} \end{aligned}$$

- Testing Equality of Means: $\mu_1 - \mu_2 = 0$

$$n_1 = n_2 = \frac{\left(z_{\alpha/2} + z_\beta\right)^2 2\sigma^2 \times IF}{\left(\mu_1 - \mu_2\right)^2}$$

$$k = \frac{\left(z_{\alpha/2} + z_\beta\right)^2 2\sigma^2 \times IF}{m\left(\mu_1 - \mu_2\right)^2}$$

- Testing Equality of Proportions: $\pi_1 - \pi_2 = 0$

$$n_1 = n_2 = \frac{\left(z_{\alpha/2} + z_\beta\right)^2 \left[\pi_1(1+\pi_1) + \pi_2(1-\pi_2)\right] \times IF}{\left(\pi_1 - \pi_2\right)^2}$$

$$k = \frac{\left(z_{\alpha/2} + z_\beta\right)^2 \left[\pi_1(1+\pi_1) + \pi_2(1-\pi_2)\right] \times IF}{m\left(\pi_1 - \pi_2\right)^2}$$

- Testing Equality of Incidence Rates: $\lambda_1 = \lambda_2$

$$n_1 = n_2 = \frac{\left(z_{\alpha/2} + z_\beta\right)^2 \left[\lambda_1 + \lambda_2\right] \times IF_t}{t\left(\lambda_1 - \lambda_2\right)^2}$$

where

$$\begin{aligned} IF_t &= 1 + \frac{CV^2\left(\lambda_1^2 + \lambda_2^2\right)t}{\left(\lambda_1 + \lambda_2\right)} \\ CV &= \frac{\sigma_1}{\lambda_1} = \frac{\sigma_2}{\lambda_2} \end{aligned}$$

and σ_i^2 is the between-cluster variation in incidence rates for the ith group, t is the person-years, and CV is the coefficient of variation, which plays the same role as the intra-class correlation coefficient. Note that if CV=0, then $IF_t = 1$

15 Sample Size Calculations for Other Types of Studies

15.1 Analysis of Change From Baseline

Let

$$\begin{aligned} y &= \text{Response variable} \\ y_b &= \text{Corresponding baseline measurement} \end{aligned}$$

$$
\begin{aligned}
\mu &= \text{Mean of the distribution of the response variable} \\
\mu_b &= \text{Corresponding baseline mean} \\
\rho &= \text{correlation between response and baseline measurements} \\
d &= y - y_b = \text{change from baseline} \\
\delta &= \mu - \mu_b \\
\sigma &= \text{standard deviation of the distribution of } d
\end{aligned}
$$

The sample size calculation for analysis based on d is given by

$$n = \frac{\left(Z_{\alpha/2} + Z_{\beta}\right)^2 2\sigma^2(1-\rho)}{\delta^2}$$

Note that if $\rho > 0.5$, it is advantageous, in terms of the sample size, to evaluate the change from baseline instead of comparing two groups in a parallel design.

15.2 Analysis of Times to Failure

Let

$$
\begin{aligned}
M_t &= \text{Mean survival time in Treatment Group} \\
M_c &= \text{Mean survival time in Control Group}
\end{aligned}
$$

Assuming the survival times are exponentially distributed, the required sample size is given by

$$n = \frac{2\left(Z_{\alpha/2} + Z_{\beta}\right)^2}{\left(\ln(M_t/M_c)\right)^2}$$

15.3 Comparisons of Means for two Poisson Populations

Let

$$
\begin{aligned}
\theta_1 &= \text{Mean of Population 1} \\
\theta_1 &= \text{Mean of Population 2}
\end{aligned}
$$

Using a two- sample test of equality of means based on samples from two Poisson populations, the required number of observations per sample is

$$n = \frac{4}{\left(\sqrt{\theta_1} - \sqrt{\theta_2}\right)^2}$$

15.4 Testing for a single correlation coefficient

Let

$$\begin{aligned} H_0 &: \rho = 0 \\ H_a &: \rho \neq 0 \end{aligned}$$

Using The Fisher's arctanh (Z) transformation

$$Z = \frac{1}{2} \ln \left(\frac{1+\rho}{1-\rho} \right)$$

and normal approximation, then the required sample is

$$n = 3 + \frac{2\left(Z_{\alpha/2} + Z_{\beta}\right)^2}{\left(\ln\left(\frac{1+\rho}{1-\rho}\right)\right)^2}$$

where ρ is regarded as the clinically meaningful value of the correlation coefficient. Note that the same formula can be used to determine the sample size for testing that the slope of a regression line is not equal to zero.

15.5 Comparing Correlation Coefficients for Two Independent Samples

Let

$$\begin{aligned} H_0 &: \rho_1 = \rho_2 \\ H_a &: \rho_1 \neq \rho_2 \end{aligned}$$

The required number of observations per sample is given by

$$n = 3 + \frac{2\left(Z_{\alpha/2} + Z_{\beta}\right)^2}{\left(\ln\left(\frac{1+\rho_1}{1-\rho_1}\right) - \ln\left(\frac{1+\rho_2}{1-\rho_2}\right)\right)^2}$$

15.6 Estimation Problems

If there are no comparisons being made but a parameter is being estimated, then confidence interval approach is used in calculating the sample size. Here we require the prior estimate of the variance and the margin of error or required accuracy. For estimating the population

mean μ, using a $(1-\alpha)100\%$ confidence interval and the desired margin of error of E, the sample size is given by

$$n = \left(\frac{Z_{\alpha/2}\sigma}{E}\right)^2$$

where σ is the prior estimate of the standard deviation of the population. The corresponding formula for estimating the population proportion is given by

$$n = \frac{Z^2_{\alpha/2}\pi(1-\pi)}{E^2}$$

where π is the prior estimate.

16 Sample Size Calculation based on Transformations

Most of the statistical testing and corresponding sample size calculation procedures are based on the normal distribution or some specific distribution of the responses or data. However, quite often the assumed distribution may not fit the data, changing the scale of the original data (transformations) and assuming the distribution for the transformed data may provides a solution. Thus, if the analysis of the data is to be done on transformed data, it is equally important to base the sample calculations on the scale of the transformed data.

1. For example if instead using risk different $p_1 - p_2$, one could use the odds ratio

$$OR = \frac{p_1(1-p_2)}{p_2(1-p_1)}.$$

   In this case the required sample size to test $H_0 : OR = 1$ versus $H_a : OR \neq 1$ is

$$n = \frac{\left(Z_{\frac{\alpha}{2}} + Z_\beta\right)^2}{\log^2(OR)}\left(\frac{1}{p_1(1-p_1)} + \frac{1}{p_2(1-p_2)}\right)$$

2. The distribution of certain outcomes such as duration of symptoms, cost, etc, is often skewed, but the log-transformation may normalize the distribution leading to a log-normal distribution. Therefore it would be important to perform the sample size calculations on the transformed scale which would used for inferences.

17 Sample Size Calculations and Non-parametric Tests

Use of non-parametric tests is also quite common in statistical analysis of RCT data.

- Most of the statistical procedures discussed so far, including those under s, have been developed under the assumption of normality or some other distribution

- Non-parametric (also called *distribution-free*) methods are designed to avoid distributional assumptions
- Advantages of Non-parametric Methods
    1. Fewer assumptions are required (ie no distributional assumptions or assumptions about equality of variances)
    2. Only nominal (categorical data) or ordinal (ranked) are required, rather than numerical (interval) data
- Disadvantages of Non-parametric Methods
    1. They are less efficient
        (a) less powerful than parametric counterparts
        (b) often lead to overestimation of variances of test statistics when there are large proportions of tied observations
    2. They don't lend themselves easily to CIs and sample size calculations
    3. Interpretation of non-parametric results is quite hard

For latest developments in sample size calculations for nonparametric tests, see Chapter 11 of

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

18 Software for Sample Size Calculations

The following article by Len Thomas and Charles J. Krebs provides an indepth review of some software of sample size calculation software:

- Thomas L, Krebs CJ. A Review of Statistical power analysis software. *Bulletin of the Ecological Society of America* 1997; 78(2): 126-139.
- Commercial Software Many more options are provided by the commercial computer package that include
    1. nQuery advisor: $http://www.statsol.ie/nquery/samplesize.htm$
    2. Power and Precision: $http://www.power-analysis.com/home.htm$
    3. PASS 2002: $http://www.ncss.com/pass.html$
- Freeware on the web (User Beware!)

http://www.stat.ucla.edu/~jbond/HTMLPOWER/index.html
http://www.health.ucalgary.ca/~rollin/stats/ssize/
http://www.stat.uiowa.edu/%7Erlenth/Power/index.html
http://www.dssresearch.com/SampleSize/
http://www.stat.ucla.edu/calculators/powercalc/
http://hedwig.mgh.harvard.edu/sample_size/size.html
http://www.bobwheeler.com/stat/SSize/ssize.html
http://www.math.yorku.ca/SCS/Online/power/
http://www.surveysystem.com/sscalc.htm
http://www.researchinfo.com/docs/calculators/samplesize.cfm
http://espse.ed.psu.edu/spsy/Watkins/Watkins3.ssi
http://www.mc.vanderbilt.edu/prevmed/ps/index.htm

References

[1] Chow S-C, Shao J, Wang H. *Sample Size Calculations in Clinical Research* Marcel Dekker: New York, NY 2003

[2] van Belle G. *Statistical Rules of Thumb*. Wiley: New York, NY 2002.

[3] Dubey SD. Some Thoughts on the One-sided and Two-sided Tests. *Journal of Biopharmaceutical Statistics* 1991; 1: 139-150.

[4] Moher D, Dulberg CS, Wells GA. Statistical Power, Sample Size, and Their Reporting in Randomized Controlled Trials. *JAMA* 1994;272:122-124

[5] Altman DG. Statistics and ethics in medical research, III: how large a sample? *BMJ* 1980;281:1336-1338.

[6] Altman DG. *Practical Statistics for Medical Research*. Chapman and Hall, London, 1991.

[7] Lenth RV. Some Practical Guidelines for Effective Sample Size Determination. *American Statistician* 2001; 55: 187-93

[8] Spiegelhalter DJ, Freedman LS, Parmar MKB. Bayesian approaches to randomized trials. In *Bayesian Biostatistics*. DA Berry, DK Stangl (eds). Marcel Dekker: New York, NY 1996

[9] Spiegelhalter DJ, Freedman LS. A predictive approach to selecting the size of a clinical trial, based on subjective clinical opinion. *Stat Med* 1986; **5**: 1-13.

[10] Halperin M, Lan KKG,Ware JH, Johnson NJ, DeMets DL. An aid to data monitoring in long-term clinical trials. *Contr Clin Trials* 1982; **3**: 311-323

[11] Lachin JM. *Biostatistical Methids*. Wiley and Sons, New York, NY 2000

[12] Lachin JM. Introduction to sample size determination and power analysis for clinical trials. *Contr Clin Trials* 1981; **1**: 13-28.

[13] Fleiss JL. *Statistical methods for rates and proportions, 2nd ed.* John Wiley & Sons: New York, NY 1981

[14] Armitage P, Berry G, Matthews JNS. *Statistical Methods in Medical Research*, 4th ed. Blackwell, Oxford, 2002.

[15] Bland JM, Altman DG. One and two sided tests of significance. *BMJ* 1994; 309: 248.

[16] Bland M. *An Introduction to Medical Statistics*, 3rd. ed. Oxford University Press, Oxford, 2000.

[17] Elashoff JD. *nQuery Advisor Version 4.0 User's Guide*. Los Angeles, CA, 2000.

[18] Guthrie E, Kapur N, Mackway-Jones K, Chew-Graham C, Moorey J, Mendel E, Marino-Francis F, Sanderson S, Turpin C, Boddy G, Tomenson B. Randomised controlled trial of brief psychological intervention after deliberate self poisoning. *BMJ* 2001; 323: 135-138.

[19] Lemeshow S, Hosmer DW, Klar J, Lwanga SK. *Adequacy of sample size in health studies*. John Wiley & Sons, Chichester, 1996.

[20] Thomas L, Krebs CJ. A Review of Statistical power analysis software. *Bulletin of the Ecological Society of America* 1997; 78(2): 126-139.

[21] Machin D, Campbell MJ, Fayers P, Pinol, A. (1998) Statistical Tables for the Design of Clinical Studies, Second Edition Blackwell, Oxford.

[22] Pocock SJ. *Clinical Trials: A Practical Approach*. John Wiley and Sons, Chichester 1983.

[23] Thomas M, McKinley RK, Freeman E, Foy C. Prevalence of dysfunctional breathing in patients treated for asthma in primary care: cross sectional survey. *BMJ* 2001; 322: 1098-1100.

[24] Whitehead, J. *The Design and Analysis of Sequential Clinical Trials*, revised 2nd. ed., Wiley, Chichester 1997.

[25] St. George's Hospital Medical School. *Statistics Guide for Research Grant Applicants*. Available at http://www.sghms.ac.uk/depts/phs/guide/size.htm#which (Last accessed on September 1, 2003)

[26] Willan AR. Power function arguments in support of an alternative approach for analyzing management trials. *Contr Clin Trials* 1994; 15:211-219.

[27] Cassagrande JT, Pike MC, Smith PG. The power function of the "exact" test for comparing two binomial distributions. *Applied Statistics* 1978; 27:176-189.

[28] Freedman LS, Lowe D,Macaskill P. Stopping rules for clinical trials incorporating clinical opinion. *Biometric*s 1984; 40:575-586.

[29] George SL, Desu MM: Planning the size and duration of a clinical trial studying the time to some critical event. *J Chron Dis* 1974; 27:15-24.

[30] Rubinstein LV, Gail MH, Santner TJ: Planning the duration of a comparative clinical trial with loss to follow-up and a period of continued observation. *J. Chron Dis* 1981; 34:469-479.

[31] Shuster JJ. *CRC Handbook of Sample Size Guidelines for Clinical Trials*. CRC Press, 1990.

[32] Lachin JM. Introduction to sample size determination and power analysis for clinical trials. *Contr Clin Trials* 1981; 2:93-113. Last JM (Ed.). A Dictionary of Epidemiology, 3rd Edition. New York: Oxford University Press, Inc., 1995. ISBN: 0-19-509668-1.

[33] Aday LA. Chapter 7: Deciding how many will be in the sample. In *Designing and Conducting Health Surveys: A Comprehensive Guide*, 2nd Edition. San Francisco: Jossey-Bass Publishers, 1996.

[34] Gordis L. *Epidemiology.* Philadelphia: W.B. Saunders Company, 1996.

[35] Fletcher RH, Fletcher SW, Wagner EH. *Clinical Epidemiology, The Essentials*, 3rd Edition. Philadelphia: Williams & Wilkins, 1996.

[36] Campbell M, Grimshaw J, Steen N, for the Changing Professional Practice in Europe Group. Sample size calculations for cluster randomised trials. *J Health Serv Res Policy*, 2000;5(1):12-16.

[37] Ray JG, Vermeulen MJ. Sample size estimation for the sorcerers apprentice. *Can Fam Phys*, July 1999;45:1999.

[38] Lwanga S, Lemshow S. *Sample Size Determination in Health Studies: a Practical Manual.* Geneva, Switzerland: World Health Organization, 1991.

[39] Schulz KF, Grimes DA. Sample size slippages in randomised trials: Exclusions and the lost and wayward. *Lancet* 2002; 358: 781-5.

[40] Boen JR, Zahn DA. *The Human Side of Statistical Consulting*, Lifetime Learning Publications, Belmont, CA 1982.

[41] Borenstein M, Rothstein H, Cohen J. *Power and Precision*, Biostat, Teaneck, NJ, Software for MS-DOS systems 1997.

[42] Castelloe J. Sample Size Computations and Power Analysis with the SAS System, in *Proceedingsof the Twenty-Fifth Annual SAS Users Group International Conference* 2000, Cary, NC, SAS Institute, Inc.,Paper 265-25.

[43] Cohen J. *Statistical Power Analysis for the Behavioral Sciences*, 2nd Edn, Academic Press, New York, 1988.

[44] Desu MM, Raghavarao D. *Sample Size Methodology*, Academic Press, Boston 1990.

[45] Elashoff J. *nQuery Advisor Release 4.0*, Statistical Solutions, Cork, Ireland, Software for MS-DOS systems 2000.

[46] Freiman JA 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 Controlled Trial: Survey of 71 "Negative Trials", in *Medical Uses of Statistics*, eds. J. C. Bailar III and F. Mosteller, chap. 14, pp. 289304, NEJM Books, Waltham, Mass. 1986.

[47] Hintze J. *PASS 2000*, Number Cruncher Statistical Systems, Kaysville, UT, Software for MS-DOS systems 2000.

[48] Hoenig JM, Heisey DM. The Abuse of Power: The Pervasive Fallacy of Power Calculationsin Data Analysis, *The American Statistician* 2001; 55: 1924.

[49] Kraemer HC, Thiemann S. *How Many Subjects? Statistical Power Analysis in Research* ,Sage Publications, Newbury Park, CA 1987.

[50] Lenth RV. (2000), Java applets for power and sample size. Avalialble at http://www.stat.uiowa.edu/ rlenth/Power/ 2000. (Last accessed on September 7, 2003)

[51] Lipsey MW. *Design Sensitivity: Statistical Power for Experimental Research*, Sage Publications, Newbury Park, CA 1990.

[52] Mace AE. (1964), *Sample-size determination*, Reinhold, New York, 1964.

[53] Muller KE, Benignus V. A. Increasing scientific power with statistical power, *Neurotoxicology and Teratology* 1992; 14: 211219.

[54] OBrien RG. *UnifyPow.sas Version 98.08.25*, Department of Biostatistics and Epidemiology,Cleveland Clinic Foundation, Cleveland, OH, 1998. Available for download from http://www.bio.ri.ccf.org/power.html (Last accessed on September 7, 2003)

[55] Odeh RE, Fox M. *Sample Size Choice: Charts for Experiments with Linear Models*, 2nd edn. Marcel Dekker, New York 1991

[56] Schuirmann D. A compromise test for equivalence of average bioavailability, *ASA Proceedings of the Biopharmaceutical Section* 1987; 137142.

[57] Taylor DJ, Muller KE. Computing Confidence Bounds for Power and Sample Size of theGeneral Linear Univariate Model, *The American Statistician* 1995; 49: 4347.

[58] Thomas L. Retrospective Power Analysis, *Conservation Biology* 1997; 11: 276280.

[59] Thomas L. Statistical power analysis software. 1998. Avalialble at http://www.forestry.ubc.ca/ conservation/power/ (Last accessed on September 7 2003)

[60] Thornley B, Adams C. Content and quality of 2000 controlled trials in schizophrenia over 50years, *BMJ* 1998; 317: 11811184.

[61] Wheeler RE. Portable Power, *Technometrics* 1974; 16:193-201.

[62] Wright TA. simple algorithm for tighter exact upper confidence bounds with rare attributes infinite universes, *Statistics and Probability Letters* 1997; 36: 5967

[63] Whitley E, Ball J. Statistics review 4: Sample size calculations. *Critical Care* 2002; 6:335-341

[64] Dupont WD, Plummer WD, Jr. Power and Sample Size Calculations: A Review and Computer Program. *Controlled Clinical Trials* 11:116-128, 1990

[65] Dupont WD, Plummer WD, Jr. Power and Sample Size Calculations for Studies Involving Linear Regression. *Controlled Clinical Trials* 19:589-601, 1998

[66] Schoenfeld DA, Richter JR. Nomograms for calculating the number of patients needed for a clinical trial with survival as an endpoint. *Biometrics* 38:163-170, 1982

[67] Pearson ES, Hartley HO. *Biometrika Tables for Statisticians* Vol. I 3rd Ed. Cambridge: Cambridge University Press, 1970

[68] Schlesselman JJ. *Case-Control Studies: Design, Conduct, Analysis*. New York: Oxford University Press, 1982

[69] Casagrande JT, Pike MC, Smith PG. An improved approximate formula for calculating sample sizes for comparing two binomial distributions. *Biometrics* 1978; 34:483-486

[70] Dupont WD. Power calculations for matched case-control studies. *Biometrics* 44:1157-1168,

[71] Gore SM. Assessing clinical trials trial size. *BMJ* 1981; 282: 1687-1689.

[72] Friedman L, Furberg C, DeMets D. *Fundamentals of clinical trials*. 3rd ed. New York: Springer-Verlag; 1998.

[73] Gebski V, Marschner I, Keech AC. Specifying objectives and outcomes for clinical trials. *Med J Aust* 2002; 176: 491-492.

[74] Kirby A,Gebski V, Keech AC. Determining the sample size in a clinical trial. *MJA* 2002 177 (5): 256-257

[75] Beal SL. Sample Size Determination for Confidence In-tervals on the Population Mean and on the Differences Between Two Population Means. *Biometrics* 1989; 45: 969977.

[76] Diletti E, Hauschke D, Steinijans VW. Sample Size Determination for Bioequivalence Assessment by Means of Confidence Intervals, *International Journal of Clinical Pharmacology, Therapy and Toxicology* 1991; 29: 18.

[77] OBrien R, Lohr V. Power Analysis For Linear Models: The Time Has Come. *Proceedings of the Ninth Annual SAS Users Group International Conference* 1984, 840846.

[78] Owen DB. A Special Case of a Bivariate Non-centralt-distribution. *Biometrika* 1965; 52: 437446.

[79] Phillips KF. Power of the Two One-Sided Tests Procedurein Bioequivalence. *Journal of Pharmacokinetics and Biophar-maceutics* 1990; 18(2): 137144.

[80] Gordon D, Finch SJ, Nothnagel M, Ott J. Power and Sample Size Calculations for Case-Control Genetic Association Tests when Errors Are Present: Application to Single Nucleotide Polymorphisms. *Human Heredity* 2002;54:22-33

[81] Kirby A, Gerski V, Keech AC. Determining the Smaple Size in a Clinical Trial. **MJA** 2002; 177: 256-257

[82] Donner A, Klar N. *Design and Analysis of Cluster Randomized Trials in Health Research*. Arnold, London 2000.

[83] Hutton LJ. Are Distinctive Ethical Principles Required for Cluster Randomized Controled Trials. *Statistics in Medicine* 2001; 20: 473-488

[84] Hayes RJ, Bennett S. Simple Sample Size Calculations for Cluster-randomized Trials. *Int J Epidermiology* 1999; 28:319-326

[85] *Recruiting Human Subjects: Sample Guidelines for Practice*. OEI-01-97-00196, June 2000

[86] American Psychological Association. *Publication manual of the American Psychological Association* (5th ed.). Washington, DC: Author 2001.

[87] Aron A, Aron EN. *Statistics for psychology* (2nd ed.). Upper Saddle River, NJ: Prentice Hall 1999.

[88] Cohen J. *Statistical power analysis for the behavioral sciences* (2nd ed.). Hillsdale, NJ: Erlbaum 1988.

[89] Cohen J. Things I have learned (so far). *American Psychologist* 1990; 45: 1304-1312.

[90] Cohen J. A power primer. *Psychological Bulletin* 1992;112: 155-159.

[91] Cohen J, Cohen P. *Applied multiple regression/correlation analysis for the behavioral sciences*. Hillsdale, NJ: Erlbaum 1975.

[92] Comrey AL, Lee HB. *A first course in factor analysis* (2nd ed.). Hillsdale, NJ: Erlbaum 1992.

[93] Green SB. How many subjects does it take to do a regression analysis? *Multivariate Behavioral Research* 1991; 26: 499-510.

[94] Guadagnoli E, Velicer WF. Relation of sample size to the stability of component patterns. *Psychological Bulletin* 1988; 103: 265-275.

[95] Harris RJ. *A primer of multivariate statistics* (2nd ed.). New York: Academic Press 1985.

[96] Howell DC. *Statistical methods for psychology* (4th ed.). Belmont, CA: Wadsworth 1997.

[97] Hoyle EH (Ed.). *Statistical strategies for small sample research*. Thousand Oaks, CA: Sage 1999.

[98] Kraemer HC, Thiemann S. *How many subjects? Statistical power analysis in research*. Newbury Park, CA: Sage 1987.

[99] Pedhazur EJ, Schmelkin LP. *Measurement, design, and analysis: An integrated approach*. Hillsdale, NJ: Erlbaum 1991.

[100] Tabachnick BG, Fidell LS. *Using multivariate statistics* (3rd ed.). New York: HarperCollins 1996.

[101] Wilkinson L, Task Force on Statistical Inference, APA Board of Scientific Affairs. Statistical methods in psychology journals: Guidelines and explanations. *American Psychologist* 1999; 54: 594-604.

[102] Wolins L. *Research mistakes in the social and behavioral sciences*. Ames: Iowa State University Press 1982.

[103] Shuster JJ. *Handbook of Smaple Size Guidelines for Clinical Trials*. CRC Press, Boca Raton: Florida 1990

[104] Lemeshow S, Hosmer DW (Jr.), Klar J, Lwanga SK. *Adequacy of Smaple Size in Health Studies*. World Health Organization, Wiley, New York: NY 1990.

[105] Hair JF, Anderson RE, Tatham RL, Black WC. *Multivariate Data Analysis*, 5th Edition, Prentice Hall, New Jersey; 1998.