How to report on sample size estimation and power in a study protocol

1. Practical Issues

- Careful consideration for and justification of sample size is a necessary but never a sufficient condition for funding, passing a comprehensive exam, or developing a defensible thesis proposal.

- However, it is worth asking whether there is always a good scientific rationale for always providing such a justification.

- Be sure that the sample size/power section of your protocol is based on the study objectives.

- Try to justify and reference the selected parameters although often these are often based on little information.

- Use a range of values in a sensitivity analysis exploring the effect of varying parameters on the required sample size/power (even if only a subset of this information appears in your protocol.)

- It is often necessary to omit some details of the study design (e.g. adjusting for confounders or stratification) when estimating sample size due to difficulties in obtaining accurate values for required parameters.

- Be sure to state whether type I error is one or two-sided.

- Provide a strong justification if you decide to use a one-sided type I error and/or lower (or greater?) than 80% power.

- Consider discussing strategies for preventing missing data or loss to follow-up for the variables of primary interest. (Also consider discussing how you will adjust for loss to follow-up and/or missing data at data analysis.)
2. Randomized vs. observational studies

- Review papers summarizing sample size estimation tend to focus either on randomized trials (e.g., Donner (1984), Lachin (1981)) or on observational studies (e.g. Wickramaratne, 1995).

- This is a bit artificial since sample size formula used when designing randomized trials may also be directly applied to the design of case-control studies, cross-sectional studies and cohort studies.

- Why is any distinction made?

- Perhaps, in part, to simplify application of the formulae and to account for any differences in notation or nomenclature.

- Sample size estimation is also seen as simpler for randomized trials than for observational studies since random assignment offers the opportunity of obtaining unbiased estimates of the effect of interest.

- Therefore, omitting consideration for stratification variables or for baseline predictors when designing a randomized trial should result in a conservatively large study assuming that these are appropriately accounted for during data analysis.
3. **Two Independent Groups**

- Consider a study comparing two independent groups having a normally distributed outcome with a mean given by $\mu_i$, $i = 1, 2$, and a common variance $\sigma^2$.

- The required sample size per group is then given by...

\[
n = \frac{(1.96 + 0.841)^2 \ 2\sigma^2}{(\mu_1 - \mu_2)^2}
\]

assuming a two-tailed 5% type I error rate and 80% power.

- If the outcome follows a binomial distribution the required sample size per group may be estimated as...

\[
n = \frac{(1.96 + 0.841)^2 \ [\pi_1(1 - \pi_1) + \pi_2(1 - \pi_2)]}{(\pi_1 - \pi_2)^2}
\]

again assuming a two-tailed 5% type I error rate and 80% power (Pocock, 1983 page 125).

- Numerous alternative sample size formula have been developed for sample size estimation of binomial data (see, e.g. Gordon, 1994).
4. Cohen’s Effect Size Approach

- Investigators often have difficulty obtaining the parameters needed to estimate sample size.

- Cohen’s (1992) effect size approach simplifies such calculations.

- For example, in a study comparing means of normally distributed outcomes for two independent groups Cohen (1992) defines the effect size to be \((\mu_1 - \mu_2)/\sigma\) and provides small, medium and large values equal to 0.2, 0.5, and 0.8 respectively so that the required sample size per group is then given by...

\[
\begin{align*}
  n &= \frac{(1.96 + 0.841)^2 \times 2 \sigma^2}{(\mu_1 - \mu_2)^2} \\
  &= \frac{2(1.96 + 0.841)^2}{ES^2} \text{ where} \\
  ES &= (\mu_1 - \mu_2)/\sigma
\end{align*}
\]

assuming a two-tailed 5% type I error rate and 80% power.

- Why should the specified effect sizes be of equal importance across all potential study outcomes?

- The effect size used to compare two independent groups each having a binomial outcome is

\[
ES = \arcsin(\sqrt{\hat{\pi}_1}) - \arcsin(\sqrt{\hat{\pi}_2})
\]

\[
= \sin^{-1}(\sqrt{\hat{\pi}_1}) - \sin^{-1}(\sqrt{\hat{\pi}_2})
\]

selected since \(Var[\arcsin(\sqrt{\hat{\pi}_1}) - \arcsin(\sqrt{\hat{\pi}_2})] = 1/4n\) when expressing the transformed proportions in radians and where \(\hat{\pi}_i\) denotes the observed proportion for the \(i\)’th group \(i = 1, 2\) (Steel and Torrie, 1960 page 158) so that the resulting sample size formula is given by (Breslow and Day, 1987 page 290)

\[
n = (1.96 + 0.841)^2/(2 \times ES^2)
\]
5. **Two Independent Groups: A Sensitivity Analysis**

Consider the following two sample size estimates for a cohort study comparing smoking quit rates of men ($\pi_2$) and women ($\pi_1$) who participated in a smoking cessation program.

These formula are provided, respectively, by Pocock (1983, 125) and Breslow and Day (1987, page 289), Breslow and Day (1987, 290)

\[
n = \frac{(1.96 + 0.841)^2 \left[ \pi_1(1 - \pi_1) + \pi_2(1 - \pi_2) \right]}{(\pi_1 - \pi_2)^2}
\]

\[
m = \frac{\left(1.96\sqrt{2\overline{\pi}(1 - \overline{\pi})} + 0.841\sqrt{\pi_1(1 - \pi_1) + \pi_2(1 - \pi_2)}\right)^2}{(\pi_1 - \pi_2)^2}
\]

where

\[
\bar{\pi} = (\pi_1 + \pi_2)/2
\]

\[
o = (1.96 + 0.841)^2/(2 \times ES^2)
\]

where

\[
ES = \arcsin(\sqrt{\pi_1}) - \arcsin(\sqrt{\pi_2})
\]

<table>
<thead>
<tr>
<th>$\pi_1$</th>
<th>$\pi_2$</th>
<th>$\pi_1 - \pi_2$</th>
<th>n</th>
<th>m</th>
<th>o</th>
<th>ES</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.15</td>
<td>0.10</td>
<td>0.05</td>
<td>682.57</td>
<td>685.31</td>
<td>680.07</td>
<td>0.07595</td>
</tr>
<tr>
<td>0.20</td>
<td>0.10</td>
<td>0.10</td>
<td>196.14</td>
<td>198.88</td>
<td>194.83</td>
<td>0.14190</td>
</tr>
<tr>
<td>0.25</td>
<td>0.10</td>
<td>0.15</td>
<td>96.76</td>
<td>99.50</td>
<td>96.28</td>
<td>0.20185</td>
</tr>
<tr>
<td>0.30</td>
<td>0.10</td>
<td>0.20</td>
<td>58.84</td>
<td>61.57</td>
<td>58.98</td>
<td>0.25789</td>
</tr>
<tr>
<td>0.20</td>
<td>0.15</td>
<td>0.05</td>
<td>902.24</td>
<td>904.99</td>
<td>901.97</td>
<td>0.06595</td>
</tr>
<tr>
<td>0.25</td>
<td>0.15</td>
<td>0.10</td>
<td>247.14</td>
<td>249.88</td>
<td>247.49</td>
<td>0.12590</td>
</tr>
<tr>
<td>0.30</td>
<td>0.15</td>
<td>0.15</td>
<td>117.68</td>
<td>120.42</td>
<td>118.51</td>
<td>0.18194</td>
</tr>
<tr>
<td>0.35</td>
<td>0.15</td>
<td>0.20</td>
<td>69.63</td>
<td>72.36</td>
<td>70.82</td>
<td>0.23535</td>
</tr>
<tr>
<td>0.25</td>
<td>0.20</td>
<td>0.05</td>
<td>1090.54</td>
<td>1093.28</td>
<td>1091.44</td>
<td>0.05995</td>
</tr>
<tr>
<td>0.30</td>
<td>0.20</td>
<td>0.10</td>
<td>290.29</td>
<td>293.03</td>
<td>291.57</td>
<td>0.11599</td>
</tr>
<tr>
<td>0.35</td>
<td>0.20</td>
<td>0.15</td>
<td>135.12</td>
<td>137.86</td>
<td>136.69</td>
<td>0.16940</td>
</tr>
<tr>
<td>0.40</td>
<td>0.20</td>
<td>0.20</td>
<td>78.46</td>
<td>81.19</td>
<td>80.27</td>
<td>0.22107</td>
</tr>
</tbody>
</table>
6. Sample size vs. Power

• Consider a study comparing binomially distributed data from two independent groups of size \(n_i, \ i = 1, 2\).

• The test statistic for the null hypothesis \(H_0 : \pi_1 = \pi_2\) vs. \(H_A : \pi_1 \neq \pi_2\) is given by

\[
Z = \frac{\hat{\pi}_1 - \hat{\pi}_2}{\sqrt{\hat{\pi}_1(1 - \hat{\pi}_1)/n_1 + (\hat{\pi}_2(1 - \hat{\pi}_2)/n_2)}},
\]

where \(Z \sim N(0,1)\) under the null hypothesis and \(\hat{\pi}_i\) denotes the observed proportion of subjects with the outcome of interest in the \(i\)th group, \(i = 1, 2\).

• If under \(H_A : \pi_1 - \pi_2 > 0\) ...

\[
Power \approx Pr \left( Z \leq -1.96 + \frac{\pi_1 - \pi_2}{\sqrt{\hat{\pi}_1(1 - \hat{\pi}_1)/n_1 + (\hat{\pi}_2(1 - \hat{\pi}_2)/n_2)}} \right) + Pr \left( Z \leq -1.96 - \frac{\pi_1 - \pi_2}{\sqrt{\hat{\pi}_1(1 - \hat{\pi}_1)/n_1 + (\hat{\pi}_2(1 - \hat{\pi}_2)/n_2)}} \right)
\]

• Power may be calculated using the SAS commands...

\[
nc=(pi1-pi2)/sqrt((pi1*(1-pi1)/n1) + (pi2*(1-pi2)/n2));
power=probnorm(-1.96+nc) + probnorm(-1.96-nc);
\]

• The corresponding sample size is given by

\[
n = \frac{(1.96 + 0.841)^2 [\pi_1(1 - \pi_1) + \pi_2(1 - \pi_2)]}{(\pi_1 - \pi_2)^2}
\]

assuming equal numbers of subjects per group and a two-sided 5% type I error rate and 80% power.
7. Justifying Sample Size using Confidence Intervals

- Most surveys are designed primarily to estimate population parameters (e.g. percent of smokers, number of cigarettes smoked per day by current smokers).

- It is usually reasonable to justify their size based on the anticipated confidence interval width for the variables of primary importance.

- This is particularly easy to do for binomial outcome data as then for a specified sample size, $n$, the maximum width for a confidence interval will occur when the true probability of a positive response is equal to 50%.

- A more controversial approach for estimating sample size of clinical trials based on confidence intervals has also been debated in the literature (Goodman and Berlin (1994), Daly (1991)).
8. Unequal numbers of subjects per group

- Clinical trials are most often designed with equal numbers of subjects in each of the intervention groups.

- An advantage of such ‘balanced’ designs is that they tend to have greater power than would otherwise be the case.

- In case-control and cross-sectional studies, however, there are almost always unequal numbers of subjects per group complicating sample size estimation.

- Let $k = n_2/n_1$ denote the anticipated degree of imbalance in sample size for a study comparing two independent groups.

- The required sample size in each group can be estimated in two steps:
  - Calculate the value of $n$ needed for equal group sizes.
  - The number of subjects in the $i$’th group is then given by

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

$$
n_2 = \frac{1}{2}n(1 + k)
$$

- Note that the total required sample size $n_1 + n_2 \geq 2n$ with equality holding only if $k = 1$.

- For example, suppose that $k = n_2/n_1 = 2$ then

$$
n_1 = \frac{1}{2}n(1 + (1/2)) = \frac{3n}{4}
$$

$$
n_2 = \frac{1}{2}n(1 + 2) = \frac{6n}{4}
$$

so that $n_1 + n_2 = 9n/4 > 2n$. 

9. Stratified data

- Breslow and Day (1987, Section 7.8), Rosner (1995, Section 10.11) and Wickramaratne (1995) discuss sample size estimation for comparisons of two independent, binomially distributed samples selected from each of $k$ strata.

- The resulting sample size formula is quite complicated reflecting the complexity of the variance of the Mantel-Haenszel odds ratio estimator.

- Sample size estimation requires specification of the following parameters:
  - Type I and type II error rates;
  - proportion of exposed subjects in the $i$'th stratum, $i = 1, ..., k$;
  - proportion of diseased subjects in the $i$'th stratum, $i = 1, ..., k$;
  - proportion of total study population in the $i$'th stratum, $i = 1, ..., k$.
  - common log odds ratio, $\gamma$.

- For a study designed to have 2 age strata, for example, investigators need to specify type I and type II error rates in addition to 6 proportions and a common log odds ratio!

- Difficulties in obtaining accurate values for all parameters often results in sample size being estimated ignoring the stratification.

- In any event, Breslow and Day (1987, page 306) claim that for studies with only two strata one would expect, in practice, an increase of no more than 10% in sample size.
10. Case-control and related designs

- Consider a case-control study with an equal number, \( n \), of cases and controls where \( \pi_1 \) and \( \pi_2 \) denote the true risk of exposure among cases and controls respectively.

- The required sample size may be estimated as...

\[
n = \frac{(1.96 + 0.841)^2 \left[ \pi_1(1 - \pi_1) + \pi_2(1 - \pi_2) \right]}{(\pi_1 - \pi_2)^2}
\]

assuming a two-tailed 5% type I error rate and 80% power.

- The formula can be extended to allow for unequal numbers of cases and controls using the approach described on page 8.

- The specified sample size formula corresponds to an odds ratio

\[
\psi = \frac{\pi_1(1 - \pi_2)}{\pi_2(1 - \pi_1)} \text{ so that }
\pi_1 = \psi \pi_2 / (1 - \pi_2 + \psi \pi_2).
\]

- The same formula may be used for nested case-control studies or for case-cohort studies however the resulting sample size may be conservatively large since analyses will likely then be based on extensions of the Cox model (Barlow et al. (1999), Langholz and Thomas, (1990)).
11. Three or more independent groups

- Suppose an investigator is interested in designing a cohort study comparing the risk of having a low weight baby among \( n \) mothers classified as never, current and former smokers respectively denoted \( \pi_1, \pi_2 \) and \( \pi_3 \) and assuming that under the null hypothesis \( H_0 : \pi_1 = \pi_2 = \pi_3 \)

- Power can be obtained as described by Lachin (1977) using a Pearson chi-square test with 2 degrees of freedom and effect size

\[
ES = \frac{\sum_{i=1}^{3} Q_i(\pi_i - \pi)^2}{\pi(1 - \pi)} \text{ where }
\]

\[
\pi = \sum_{i=1}^{3} Q_i \pi_i
\]

and the number of subjects in the \( j \)'th risk group is equal to \( n_i = nQ_i, \)
\( i = 1, 2, 3, \) such that \( Q_1 + Q_2 + Q_3 = 1 \) and \( n \times ES \) follows a non-central chi-square distribution.

- A non-central chi-square random variable can be evaluated using the SAS function

\[
\text{NC}=n*ES;
\]

\[
\text{chi2}=\text{cinv}(0.95, \text{df});
\]

\[
\text{power}=1-\text{probchi(chi2, \text{df}, \text{NC})};
\]

- Lachin’s (1977) method may also be extended allowing for multinomial outcomes.

- Any potential gain in scientific relevance obtained by comparing three (or more) independent groups needs to be tempered against the larger required sample size than would be the case for a two group comparison.

- One alternative would be to define two independent questions

(a) Is there any difference in the risk of having a low weight baby comparing mothers classified as ever vs. never smokers?

(b) Is there any difference in the risk of having a low weight baby among mothers classified as current or former smokers?

- How should power be determined when interventions (or exposure) differ only by dose (Breslow and Day (1987, Section 7.4), Wickramaratne (1995))?
12. Randomized Trials Using a Factorial Design
Piantadosi (1997, Chapter 15)

- Randomized trials using a factorial design offer a good example of the challenges posed when designing a study having more than two intervention groups.

- For example, the $\alpha$–tocopherol $\beta$–carotene Lung Cancer Prevention Trial used a 2x2 factorial design in which 29,133 male smokers in Finland were randomly assigned to one of four intervention groups; (i) $\alpha$–tocopherol; (ii) $\beta$–carotene; (iii) both $\alpha$–tocopherol and $\beta$–carotene, or (iv) placebo.

- In the absence of interaction the required sample size can be estimated in two steps:
  (a) Estimate the sample size as if conducting two separate randomized trials of lung cancer incidence comparing $\alpha$–tocopherol vs. placebo and comparing $\beta$–carotene vs. placebo.
  (b) The larger of the two sample size estimates would be the sample size for the factorial design.

- In the absence of interaction effects it would be inefficient to design a factorial design as if one were comparing outcomes among four independent groups.

- Most factorial designs are designed to have little power to detect interaction effects.

- One can show that the variance for a test of interaction is four times that of the variance for a test of a main effect assuming that the outcome is normally distributed and has a fixed variance resulting in a four-fold increase in sample size as compared to a study designed to detect a main effect.

- A similar increase in sample size is required to detect effects of interaction for case-control studies (see, e.g. Smith and Day, 1984).
13. Matched Case-Control Studies

- If McNemar’s test for correlated proportions is used to test the hypothesis $H_0: \psi = 1$ vs. $H_A: \psi \neq 1$ for the specific alternative $\psi_A$ the required number of matched pairs is given by (Rosner (1995, page 387))

$$n = \left(1.96 + \left[2 \times 0.841 \sqrt{\pi_A(1 - \pi_A)} \right]\right)^2$$

where $\pi_A = \psi_A / (1 + \psi_A)$ is the proportion of discordant pairs for which the case is exposed and the control is unexposed, while $\pi_D$ denotes the proportion of discordant pairs among all pairs and assuming a two-tailed 5% type I error rate and 80% power.

- Often there are a limited number of cases.

- Matching $k$ controls per case can result in a reduction in the number of required cases.

- Let $n'$ be the number of cases in the study and $kn'$ the number of controls.

- The required sample size in each group can be estimated in two steps:
  - Calculate the value of $n$ needed assuming pair-matched data.
  - The number of cases required assuming a $k : 1$ match results in

  $$\text{number of cases} = \frac{1}{2} n (1 + (1/k))$$

  $$\text{number of controls} = \frac{1}{2} n (1 + k)$$

- Or equivalently for a fixed number of available cases some gains in power can be obtained by increasing the number of controls matched to each case.

- There tends to be little gains in power achieved after matching approximately 4 or so controls per case (Julious et al., 1999) although if one anticipates more missing data among controls one might select a higher matching ratio.
14. Cross-over Trials

- Cross-over trials are a second example of a matched design.

- A two-period crossover design involves a single group of patients, each of whom serves as his own control in the comparison of two treatments.

- Each patient is given both interventions in random order.

- Sample size is typically estimated under the stringent assumption that there is no treatment by time period interaction (i.e. no carryover effect).

- For binary outcome data sample size may then be estimated using the methods described for matched case-control studies.

- A simpler approach is applicable when outcomes are normally distributed and have a fixed variance.

- The required number of subjects is then given by \( n(1 - r) \) where \( n \) denotes the number of subjects required for a comparison of two independent groups and \( r \) denotes the correlation between responses in a single subject on the two occasions (Donner, 1984).
15. **Time to event data**

(Lachin, 1981)

- Analyses of time to event data are most often constructed using procedures such as the log-rank test or the Cox model neither of which makes any parametric assumptions about the distribution of failure times.

- For purposes of sample size estimation, however, parametric assumptions are needed.

- Consider a randomized trial enrolling \( n \) subjects per intervention group and assuming that failure times follow the exponential distribution with hazard rate \( \lambda_i, i = 1, 2 \).

- The required number of subjects per group may be estimated by

\[
 n = \frac{(1.96 + 0.841)^2 (\lambda_1 + \lambda_2)^2}{2(\lambda_1 - \lambda_2)^2}
\]

assuming a two-tailed 5% type I error rate and 80% power and supposing that there are no censored observations and that all subjects start the study at the same point in time.

- In practice many subjects will have censored observations because (i) it is not practical to follow all subjects until they have an event and (ii) subjects may drop out or die before they have the event of interest.

- Furthermore it will not be practical to enroll all of the required subjects at the same time so that some subjects will be followed much longer than other subjects.

- Therefore power calculations for time to event outcomes need to make assumptions regarding

  - the rate at which patients are recruited over time,
  - accrual time,
  - follow-up time,
  - censoring rate due to competing events

further complicating the task which may be simplified somewhat using available software (Iwane et al., 1997) or the approximations provided by Lachin (1981) and Lachin and Foulkes (1986).
16. **Count data**

- Consider a cohort study comparing mortality rates among equal numbers of exposed and unexposed subjects such that each group is followed for a total of $P$ person-years.

- The required number of person-years of follow-up per exposure group may be given by

$$P = (1.96 + 0.841)^2 \frac{\lambda_E + \lambda_U}{(\lambda_E - \lambda_U)^2}$$

where $\lambda_E$ and $\lambda_U$ respectively denote mortality rates of exposed and unexposed subjects and assuming a two-tailed 5% Type I error rate and 80% power.

- Sample size formulae given by Breslow and Day (1987, page 282) for the required number of events, $O_+$, may be shown to be equivalent since $O_+ \approx P(\lambda_E + \lambda_U)$.

- Breslow and Day (1987, Chapter 7) also provide sample size formula for cohort studies having unequal numbers of subjects per exposure group and for studies comparing mortality or incidence rates with an external standard (e.g. SMR type studies).

- How might one determine sample size for a smoking cessation trial comparing the number of cigarettes smoked per day among experimental and control subjects?
17. **Multiple regression**

- Peduzzi et al. (1996) have popularized the ‘rule’ that there be at least 10 events and 10 nonevents for each covariate included in a logistic regression model.

- This ‘rule’ may ensure the validity of statistical inferences but offers absolutely no guarantee that the study will be adequately powered.

- Consider a study designed to compare binomial outcomes from two independent groups.

- Cohen (1992) provides an effect size approach for estimating sample size applicable to tests that $R^2$ from a multiple linear regression model is equal to zero. Suggest a limitation of this approach.

- Hsieh et al. (1998) provide a simple method of extending comparisons of two independent groups accounting for confounding variables:
  - Determine sample size per group omitting consideration for confounders;
  - Increase the sample size dividing by $1 - R^2$ where $R^2$ denotes the true multiple correlation coefficient between the exposure variable and the confounding variables.

- What would $R^2$ be for a randomized trial comparing intervention groups and adjusting for baseline predictors?
18. **Effect of missing data**

- Subjects who provide all of their requested data tend to be unusual.

- It is also difficult in advance of a study to obtain accurate estimates for the rates of missingness even for the primary study outcome or for the main exposure variable.

- For instance in the context of a prospective cohort study attention will usually be limited to accounting for loss to follow-up under the unrealistic assumptions that (i) such loss will occur equally for exposed and unexposed subjects and (ii) be independent of the outcome.

- Let...
  - $N=$ total number of subjects required not accounting for loss to follow-up
  - $N'=\text{total number of subjects taking into account loss to follow-up}$

  \[
  N' = \frac{N}{1 - L}
  \]
  where $L$ denotes the anticipated loss to follow-up.

- A separate discussion could be included in the protocol based on Greenland and Finkle (1995), for example, describing how missing data will be accounted for in the data analysis.
19. **Effect of errors in measurement or incomplete compliance with intervention**

- Patients do not always receive treatments as specified by protocol while exposures may be incorrectly measured in observational studies.

- Suppose that \( r_i \) patients in the \( i \)'th intervention group of a clinical trial, \( i = 1, 2 \) are noncompliant ending up with the intervention offered in the alternative arm and that \( 0 < r_i < 1 \).

- After adjusting for noncompliance the required sample size per intervention group is given by \( n/(1 - r_1 - r_2)^2 \) (Lachin and Foulkes (1986), Rosner (1995, Section 10.7.3)).

- Thus the required sample size would need to be increased by a factor of \( 1.2 = 1/(1 - 0.05 - 0.05)^2 \) if, as few as 5% of subjects per intervention group are noncompliant.

- The same procedure can account for *nondifferential* misclassification of a binary exposure.
20. **Equivalence trials**  
(Jones et al. (1996), Donner (1984))

- Most clinical trials are designed to detect a clinically meaningful minimum difference across intervention groups.

- Equivalence trials, however, are designed to demonstrate the equivalence of two (or more) interventions and are most often of interest when a new intervention is designed to match the efficacy of the standard intervention but has advantages in safety, convenience or cost.

- Absolute equivalence can never be demonstrated.

- In its place investigators may specify an interval of equivalence.

- Analyses and sample size can be based on confidence interval construction.

- Sample size may also be estimated by specifying Type I and Type II errors but revising the null and alternative hypotheses.

- Sample size for a one-sided test of $H_0: \mu_1 - \mu_2 \geq \Delta$ (not equivalent) vs. $H_A: \mu_1 - \mu_2 < \Delta$ (equivalent) assuming a normally distributed outcome with variance $\sigma^2$ is given by...

\[
n = \frac{(1.96 + 0.841)^2 \, 2\sigma^2}{\Delta^2}
\]

where $\Delta$ denotes the boundary for equivalence and assuming a one-sided 2.5% Type I error rate and 80% power.

- Intention to treat analyses are of lesser importance for equivalence trials.
Correlated Outcomes Data: Accounting for the Effects of Clustering
(Donner and Klar, 2000)

- Correlated outcomes data may arise in a number of study designs including complex surveys when sampling subjects by household, in longitudinal studies when repeated assessments are obtained for each subject over time, or in cluster randomization trials when randomizing families, doctor’s practices, schools or communities.

- Sample size for a study comparing outcomes from two independent groups may be estimated in two steps assuming that the exposure (or intervention) is measured at the cluster level:
  (a) estimate sample size assuming that outcomes are independent;
  (b) multiply the resulting sample size by the variance inflation factor
      \[ 1 + (\bar{n} - 1)\rho \]
      where \( \bar{n} \) denotes the average cluster size, and \( \rho \) denotes the intracluster correlation coefficient for the study outcome.

- Note that even very small degrees of intracluster correlation can dramatically inflate the variance.

- For example, in the context of a school-based cluster randomized trial Donner and Klar (2000) report a variance inflation factor of
  \[ 1 + (\bar{n} - 1)\rho \approx 1 + (100 - 1)0.01 = 1.99 \]
  doubling the required sample size.

- If exposure is measured at the individual level \( \rho \) needs to be replaced by \( \rho_e \) where \( \rho_e \) denotes the intracluster correlation coefficient for the exposure variable.

- Pair-matched case-control studies and 2x2 cross-over trials are extreme examples of clustered data where \( \rho_e = -1 \) so that the variance inflation factor is equal to \( (1 - \rho) \) implying that for such studies clustering increases rather than decreasing power!
22. Software

- Unfortunately none of the available software packages can be used for all problems.

- There are a number of free, downloadable sample size calculators of varying quality. I have heard good things about...

**PS: Power and Sample Size**
by William D. Dupont and Walton D. Plummer, Jr.
[http://www.mc.vanderbilt.edu/prevm/ps.htm](http://www.mc.vanderbilt.edu/prevm/ps.htm)

This freeware provides sample size estimates and power calculations for comparison of two independent groups having binary, continuous or time to event outcomes.

- Major statistical packages still offer only simple methods. My favourite is...

**STATA 7**

- StatXact is speciality software for exact analysis of categorical outcome data and includes power calculations when the outcome is multinomial or ordinal


- EAST is speciality software for the design and analysis of randomized trials with normal, binomial or time to event outcomes with planned interim analyses


- ACLUSTER provides sample size estimates for cluster randomized trials based on the text by Donner and Klar (2000). Additional information is available at

References


Daly LE. Confidence intervals and sample sizes; don’t throw out all your old sample size tables. BMJ 1991; 302:333-336.


