DIGITALCOMMONS @WAYNESTATE

Journal of Modern Applied Statistical Methods

Volume 4 | Issue 2 Article 9

11-1-2005

Sample Size Calculation and Power Analysis of Time-Averaged Difference

Honghu Liu
UCLA, hhliu@ucla.edu

Tongtong Wu
UCLA, tongtong@ucla.edu

Journal of Modern Applied Statistical Methods
November, 2005, Vol. 4, No. 2, 434-445

1538 – 9472/05/$95.00

Sample Size Calculation and Power Analysis of Time-Averaged Difference

Honghu Liu
David Geffen School of Medicine
UCLA

Tongtong Wu
Department of Biostatistics
UCLA

Little research has been done on sample size and power analysis under repeated measures design. With detailed derivation, we have shown sample size calculation and power analysis equations for time-averaged difference to allow unequal sample sizes between two groups for both continuous and binary measures and explored the relative importance of number of unique subjects and number of repeated measurements within each subject on statistical power through simulation.

Key words: sample size calculation; power analysis; repeated measures design; time-averaged difference

Introduction

Sample size calculation and power analysis are essentials of a statistical design in studies. As statistical significance is likely the desired results of investigators, proper sample size and sufficient statistical power are of primary importance of a study design (Cohen, 1988). Although a larger sample size yields higher power, one cannot have as large a sample size as one wants, since sample subjects are not free and the resources to recruit subjects are always limited. As a result, a good statistical design that can estimate the needed sample size to detect a desired effect size with sufficient power will be critical for the success of a study.

Some research has been done for sample size calculation and power analysis regarding different designs with cross-sectional data, such as difference between correlations, sign-test (Dixon & Massey, 1969), difference between means with two group t-test or analysis of variance (ANOVA) (Machin, Campbell, Fayers, & Pinol, 1997), contingence tables (Agresti, 1996), difference of proportions between two groups, F-test (Scheffé, 1959), multiple regressions and logistic regressions (Whittemore, 1981; Hsieh et al., 1998).

However, little research has been done about sample size calculation and power analysis with repeated measures design, especially for unbalanced designs, which is widely used in biological, medical, health services research and other fields. For example, in research for diseases with low incidence and prevalence; designs where the non-diseased group is much larger than the diseased group to ensure a sufficient large sample size for multivariate modeling.

Unbalanced repeated measures situations also emerge in cluster randomized trials (Eldridge et al., 2001). Diggle et al. (1994) proposed a basic sample size calculation formula for time-averaged difference (TAD) with both continuous and binary outcome measures for the situation only with equal sample size in each group. Fitzmaurice et al. (2004) proposed a two-stage approach for sample size and power analyses of change in mean response over time for both continuous and binary outcomes.

Statistical software and routines have made sample size calculation and power analysis process much easier and flexible for researchers. With statistical software, one can efficiently

Dr. Honghu Liu Professor of Medicine in the Division of General Internal Medicine & Health Services Research of the David Geffen School of Medicine at UCLA. 911 Broxton Plaza, Los Angeles, CA 90095-1736. Email: hhliu@ucla.edu. Tongtong Wu is a Ph.D. Candidate in the Department of Biostatistics in the UCLA School of Public Health. 911 Broxton Plaza, Los Angeles, CA, 90095-1736. Email: tongtong@ucla.edu

examine designs with different parameters and select the best design to fit the need of a research project. Currently, there are many types of statistical software that can conduct sample size and power analyses. These include the general purpose software which contain power analysis routines such as: NCSS (NCSS, 2002), SPSS (SPSS Inc., 1999), and STATA (STATA Press, 2003); general purpose software that can be used to calculate power (i.e., contain non-central distribution or simulation purpose) such as: SAS (SAS Institute Inc., 1999), S-Plus (MathSoft, 1999), and XLISP-STAT (Wiley, 1990); and stand-alone power analysis software such as: NCSS-PASS 2002 (NCSS, 2002), nQuery advisor (Statistical Solutions, 2000), and PowerPack (Length, 1987). A comprehensive list of sample size and power analysis software can be found at http://www.insp.mx/dinf/stat_list.html.

Although a lot of software can conduct sample size and power analyses, they are basically all for data with different cross-sectional designs. The only software that can conduct sample size and power analyses with repeated measures design is NCC-PASS 2002, which handles power analysis for repeated measures ANOVA design. There is, however, no software available for TAD with repeated measures design.

In this article, a formula has been developed for sample size calculation and power analysis of TAD for both continuous and binary measures to allow unequal sample size between groups. In addition, the relative impact and equivalence of number of subjects and the number of repeated measures from each subject on statistical power was examined. Finally, a unique statistical software for conducting sample size and power analysis for TAD was created.

Methodology

Sample size Calculation and Power Analysis

Sample size calculation and power analysis are usually done through statistical testing of the difference under a specific design when the null or alternative hypothesis is true. Although there are many factors that influence sample size and power of a design, the essential factors that have direct impact on sample size and statistical power are type I error (H_0 may be rejected when it is true and its probability is denoted by α), type II error (H_0 may be accepted when it is false and its probability is denoted by β), effect size (difference to be tested and it is usually denoted by Δ) and variation of the outcome measure of each group (for example, standard deviation σ). Sample size and power are functions of these factors. Sample size and power analysis formulas link all of them together. For example, the sample size calculation formula for a two group mean comparison can be written as a function of the above factors:

$$n_2 = ((z_{1-\beta} + z_{1-\alpha/2})/(\Delta/S))^2/(1+1/r),$$

where n_2 is the sample size for group2, S is the common standard deviation of the two groups, r $0 < r \le 1$ is a parameter that controls the ratio between the sample sizes of group 1 and group 2 (i.e., $n_1 = n_2/r$). $z_{1-\beta}$ is the normal deviate for the desired power, $z_{1-\alpha/2}$ is the normal deviate for the significance level (two-sided test) and Δ is the difference to be detected.

For given levels of a type I error, a type II error and an effect size, sample size and statistical power are positively related: the larger the sample size, the higher the statistical power. Type I error is negatively related to sample size: the smaller Type I error, the larger sample size that is required to detect the effect size for a given statistical power. The larger type II error, the smaller power and thus one will need smaller sample size to detect a given effect size.

Repeated Measures Design

Time-Averaged Difference (TAD)

In many biomedical or clinical studies, researchers use the experimental design that takes multiple measurements on the same subjects over time or under different conditions. By using this kind of repeated measures design, treatment effects can be measured on “units” that are similar and precision can be determined by variation within same subject. Although the analyses become more complicated because

measurements from the same individual are no longer independent, the repeated measures design can avoid the bias from a single snapshot and is very popular in biological and medical research.

Suppose there are two groups, group 1 and group 2, and one would like to compare the means of an outcome, which could vary from time to time or under different situations between the two groups. With cross-sectional design, one will directly compare the means of the outcome between the groups with one single measure from each subject, which may not reflect the true value of the individual.

For example, it is known that an individual's blood pressure is sensitive to many temporary factors, such as mood, the amount of time slept the night before and the degree of physical exercise/movement right before taking the measurement. This is why the mean blood pressure of a patient is always examined from multiple measurements to determine his/her true blood pressure level. If only a single blood measurement is taken from each individual, then comparing mean blood pressure between two groups could be invalid as there is large variation among the individual measures for a given patient. To increase precision, the best way to conduct this is to obtain multiple measurements from each individual and to compare the time-averaged difference between the two groups (Diggle, 1994).

Notations

Suppose that there is a measurement for each individual $y_{g(ij)}$, where $g = 1,2$ indicating which group, $i = 1,...,m_k$ (with $k = 1,2$) indicating the number of individuals in each group, and $j = 1,...,n$ indicating the number of repeated measures from each individual subject. Then TAD will be defined as:

$$d = ((\sum_{i=1}^{m_1}\sum_{j=1}^{n} y_{1(ij)})/n*m_1)) - ((\sum_{i=1}^{m_2}\sum_{j=1}^{n} y_{2(ij)})/n*m_2).$$

The following notations will be used to define the different quantities used in sample size calculation and power analysis for TAD:

1. α: Type I error rate
2. β: Type II error rate
3. *d*: Smallest meaningful TAD difference to be detected
4. σ: Measurement deviation (assume to be equal for the two groups)
5. *n*: Number of repeated observations per subject
6. ρ: Correlation between measures within an individual
7. m_1, m_2: Number of subjects in group 1 and group 2, respectively
8. $M = m_1 + m_2$: Total number of subjects in the design
9. $\pi = m_1 / M$: Proportion of number of subjects within group 1 ($\pi = 0.5$ gives equal sample size. $m_1 = \pi M, m_2 = (1-\pi)M$)

Using the above notations, the next two sections will derive the sample size calculation formula for TAD between two groups with the flexibility of possible unequal sample size from each group for continuous and binary measures, respectively.

Continuous responses

Consider the problem of comparing the time-averaged difference of a continuous response between two groups. Supposed the model is of the following form:

$$Y_{ij} = \beta_0 + \beta_1 x + \varepsilon_{ij}, \quad i = 1,\cdots,M;\ j = 1,\cdots n$$

where *x* indicates the treatment assignment, $x = 1$ for group 1 and $x = 0$ for group 2. To test if the time-averaged difference is zero is equivalent to test $H_0 : \beta_1 = 0$ vs. $H_1 : \beta_1 \neq 0$. Without showing details of derivation, Diggle et al. (1994) have shown the sample size in the situation when group 1 and group 2 have the same sample size. With step by step derivation, here it is shown generally to the cases that the sample sizes of two groups could be unequal. Assume that the within subject correlation

$$Corr(y_{ij}, y_{ik}) = \rho \quad \text{for any} \quad j \neq k$$

and

$$Var\ (y_{ij}) = \sigma^2 .$$

Without lost generality, it is assumed that the smallest meaningful difference $d > 0$, and let the power of the test be $1-\beta$. Under H_0:

$$z = \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \rightarrow N(0,1)$$

The above model can be written in matrix form:

$$Y_i = X'\beta + \varepsilon$$

where

$$X_i = \begin{pmatrix} 1 & 1 \\ 1 & 1 \\ \vdots & \vdots \\ 1 & 1 \end{pmatrix} \text{ for group 1}$$

or

$$X_i = \begin{pmatrix} 1 & 0 \\ 1 & 0 \\ \vdots & \vdots \\ 1 & 0 \end{pmatrix} \text{ for group 2}$$

and

$$Y_i = \begin{pmatrix} y_{i1} \\ y_{i2} \\ \vdots \\ y_{in} \end{pmatrix}$$

The variance-covariance matrix (compound symmetry) can be written as

$$\Sigma = \sigma^2 \begin{bmatrix} 1 & \rho & \cdots & \rho \\ \rho & 1 & \cdots & \rho \\ \vdots & & \ddots & \vdots \\ \rho & \rho & \cdots & 1 \end{bmatrix}.$$

The estimates of regression coefficients of such a model are

$$\hat{\beta} = \left(\sum_i X_i' \Sigma^{-1} X_i \right)^{-1} \left(\sum_i X_i' \Sigma^{-1} Y_i \right),$$

and the estimates of variance estimate are

$$\begin{aligned} \text{var}(\hat{\beta}) &= \sigma^2 \left(\sum_i X_i' \Sigma^{-1} X_i \right)^{-1} \\ &= \frac{\sigma^2[1+(n-1)\rho]}{n[(m_1+m_2)m_2 - m_1^2]} \begin{bmatrix} m_2 & -m_1 \\ -m_1 & m_1+m_2 \end{bmatrix} \end{aligned}$$

By definition, it is known that

$$\begin{aligned} \text{Power} &= 1-\beta \\ &= \Pr(\text{rejecting } H_0 \mid H_1) = \Pr(|z| > z_{1-\alpha/2} \mid H_1) \end{aligned}$$

so,

$$\begin{aligned} &\text{Power} \\ &= \Pr\left(\left| \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \right| > z_{1-\alpha/2} \mid H_1 \right) \\ &= \Pr\left(\frac{\hat{\beta}_1}{se(\hat{\beta}_1)} > z_{1-\alpha/2} \mid H_1 \right) + \Pr\left(\frac{\hat{\beta}_1}{se(\hat{\beta}_1)} < -z_{1-\alpha/2} \mid H_1 \right) \\ &\approx \Pr\left(\frac{\hat{\beta}_1}{se(\hat{\beta}_1)} > z_{1-\alpha/2} \mid H_1 \right) \end{aligned}$$

it is assumed that $d > 0$, therefore, the second term can be ingored

$$= \Pr\left(\frac{\hat{\beta}_1 - d}{se(\hat{\beta}_1)} > z_{1-\alpha/2} - \frac{d}{se(\hat{\beta}_1)} \mid H_1 \right)$$

Therefore,

$$-z_{1-\beta} = z_{1-\alpha/2} - \frac{d}{se(\hat{\beta}_1)},$$

or

$$(z_{1-\alpha/2} + z_{1-\beta})^2 = \frac{d^2}{\text{var}(\hat{\beta}_1)} = \frac{n[(m_1+m_2)m_2 - m_1^2]d^2}{\sigma^2[1+(n-1)\rho](m_1+m_2)}$$

In other words, given power $1-\beta$, the total sample size needed to detect the smallest meaningful difference $d > 0$ is

$$M = \frac{(z_{1-\alpha/2} + z_{1-\beta})^2[1+(n-1)\rho]s^2}{n(1-\pi-\pi^2)d^2}, \qquad (1)$$

where *s* is the estimate of standard deviation. When $m_1 = m_2 = m$, the above formula becomes the same as that shown in Diggle et al. (1994) for balanced design:

$$m = \frac{2(z_{1-\alpha/2} + z_{1-\beta})^2[1+(n-1)\rho]s^2}{nd^2}. \qquad (2)$$

Given sample size,

$$z_{1-\beta} = -z_{1-\alpha/2} + \frac{d}{se(\hat{\beta}_1)} = -z_{1-\alpha/2} + \frac{\sqrt{nM(1-\pi-\pi^2)}\cdot d}{\sqrt{1+(n-1)\rho}\cdot s}$$

Therefore, the power of the test can be written as:

$$\text{Power} = 1-\beta = 1-\Phi\left(z_{1-\alpha/2} - \frac{\sqrt{nM(1-\pi-\pi^2)}\cdot d}{\sqrt{1+(n-1)\rho}\cdot s}\right) \qquad (3)$$

Binary responses

Suppose a binary response variable is to be compared between group 1and group 2. Assume

$$\Pr(Y_{ij}=1) = \begin{cases} p_1 \text{ in group 1} \\ p_2 \text{ in group 2} \end{cases}$$

To test if the proportions of responses being 1of the two groups are equal, the following model is considered

$$E(Y_{ij} \mid x_{ij}) = \Pr(Y_{ij}=1 \mid x_{ij}) = \beta_0 + \beta_1 x_{ij},$$
$$i = 1, \cdots, M;\ j = 1, \cdots n$$

where *x* indicates the treatment assignment, $x=1$ for group 1 and $x=0$ for group 2. this test will be equivalent to test $H_0: \beta_1 = 0$ vs. $H_1: \beta_1 \neq 0$. Without showing the details, Diggle et al (1994) have shown the sample size in the situation when group 1 and group 2 have the same sample size. With step by step derivation, here it is generalized to the case that the sample size could be different between the two groups.

Suppose $d = p_1 - p_2 > 0$ and the power of the test is $1-\beta$. Under H_0, the estimate of σ^2 is

$$\hat{\sigma}_0^2 = \frac{m_1 p_1 + m_2 p_2}{m_1+m_2}\cdot\left(1 - \frac{m_1 p_1 + m_2 p_2}{m_1+m_2}\right) = \frac{(m_1 p_1 + m_2 p_2)(m_1 q_1 + m_2 q_2)}{m_1+m_2}$$

where $q_1 = 1-p_1$ and $q_2 = 1-p_2$. Under H_1, the estimate of σ^2 is

$$\hat{\sigma}_1^2 = \frac{m_1}{m_1+m_2}p_1 q_1 + \frac{m_2}{m_1+m_2}p_2 q_2 = \frac{m_1 p_1 q_1 + m_2 p_2 q_2}{m_1+m_2}$$

The variance estimator of $\hat{\beta}_1$ is

$$\text{var}(\hat{\beta}_1) = \frac{\sigma^2 (m_1 + m_2)[1 + (n-1)\rho]}{n[(m_1 + m_2)m_2 - m_1^2]},$$

and it is denoted as $\hat{\sigma}_{\hat{\beta}_1, H_0}$ when replacing σ^2 by $\hat{\sigma}_0^2$, and $\hat{\sigma}_{\hat{\beta}_1, H_1}$ when replacing σ^2 by $\hat{\sigma}_1^2$.

The power of the test is:

$$\begin{aligned}
&\text{Power} \\
&= \Pr\left(\left| \frac{\hat{\beta}_1}{se(\hat{\beta}_1)} \right| > z_{1-\alpha/2} \mid H_1 \right) \\
&\approx \Pr\left(\frac{\hat{\beta}_1}{se(\hat{\beta}_1)} > z_{1-\alpha/2} \mid H_1 \right) \quad \text{because we assume } d > 0 \\
&= \Pr\left(\frac{\hat{\beta}_1 - d}{\hat{\sigma}_{\hat{\beta}_1, H_0}} > z_{1-\alpha/2} - \frac{d}{\hat{\sigma}_{\hat{\beta}_1, H_0}} \mid H_1 \right) \\
&= \Pr\left(\frac{\hat{\beta}_1 - d}{\hat{\sigma}_{\hat{\beta}_1, H_1}} \cdot \frac{\hat{\sigma}_{\hat{\beta}_1, H_1}}{\hat{\sigma}_{\hat{\beta}_1, H_0}} > z_{1-\alpha/2} - \frac{d}{\hat{\sigma}_{\hat{\beta}_1, H_0}} \mid H_1 \right) \\
&= \Pr\left(\frac{\hat{\beta}_1 - d}{\hat{\sigma}_{\hat{\beta}_1, H_1}} > \frac{\hat{\sigma}_{\hat{\beta}_1, H_0}}{\hat{\sigma}_{\hat{\beta}_1, H_1}} \cdot z_{1-\alpha/2} - \frac{d}{\hat{\sigma}_{\hat{\beta}_1, H_1}} \mid H_1 \right)
\end{aligned}$$

Therefore,

$$-z_{1-\beta} = \frac{\hat{\sigma}_{\hat{\beta}_1, H_0}}{\hat{\sigma}_{\hat{\beta}_1, H_1}} \cdot z_{1-\alpha/2} - \frac{d}{\hat{\sigma}_{\hat{\beta}_1, H_1}},$$

Or

$$\left(\frac{\hat{\sigma}_{\hat{\beta}_1, H_0}}{\hat{\sigma}_{\hat{\beta}_1, H_1}} \cdot z_{1-\alpha/2} + z_{1-\beta} \right)^2 = \frac{d^2}{\hat{\sigma}^2_{\hat{\beta}_1, H}}$$

i.e.,

$$\left(\sqrt{\frac{(m_1 p_1 + m_2 p_2)(m_1 q_1 + m_2 q_2)}{m_1 p_1 q_1 + m_2 p_2 q_2}} \cdot z_{1=\alpha/2} + z_{1-\beta} \right)^2 = \frac{nM(1-\pi-\pi^2)d^2}{[1+(n-1)\rho][\pi p_1 q_1 + (1-\pi) p_2 q_2]}$$

In other words, given power $1-\beta$, the total sample size needed to detect the smallest meaningful difference $d > 0$ is

$$M = \frac{\left(\sqrt{\frac{(m_1 p_1 + m_2 p_2)(m_1 q_1 + m_2 q_2)}{m_1 p_1 q_1 + m_2 p_2 q_2}} \cdot z_{1=\alpha/2} + z_{1-\beta} \right)^2 [1+(n-1)\rho][\pi p_1 q_1 + (1-\pi) p_2 q_2]}{n(1-\pi-\pi^2)d^2} \quad (4)$$

When $m_1 = m_2$, the above formula is the same as shown in Diggle et al. (1994) for balanced design. Given sample size, the power of the test can be calculated using the following equation:

$$\text{Power} = 1-\beta = \Phi\left(\frac{\hat{\sigma}_{\hat{\beta}_1, H_0}}{\hat{\sigma}_{\hat{\beta}_1, H_1}} \cdot z_{1-\alpha/2} - \frac{d}{\hat{\sigma}_{\hat{\beta}_1, H_1}} \right) \quad (5)$$

The Relative Impact of Number of Subjects and Number of Repeated Measures on Power

As the cost and the amount of effort to recruit subjects or to increase the number of repeated measurements for each participant is often different, it will be useful for investigators to know the relative impact of number of subjects and number of repeated measures on statistical power for testing TAD. The relative importance of the total number of subjects M and number of repeated measures n, which have nonlinear effects on the power, is now investigated. For easy derivation, let's examine the situation of continuous measure.

First, if the within subject correlation is $\rho = 0$, then it can be seen that the number of subjects *M* and number of repeated measures *n* will have exactly the same impact on statistical

power. Using formula (3) and plugging in $\rho = 0$, the power then becomes:

$$\text{Power} = 1-\beta = 1-\Phi\left(z_{1-\alpha/2} - \frac{\sqrt{nM(1-\pi-\pi^2)}\cdot d}{s}\right) \quad (6)$$

It can be explained that when $\rho = 0$ all the observations are independent and thus there is no distinction between the repeated measurements and different subjects. Second, when $\rho = 1$, the number of repeated measures has no more impact on power because it just repeats the same observations over again. This can be seen by plugging in $\rho = 1$ in formula (3):

$$\text{Power} = 1-\beta = 1-\Phi\left(z_{1-\alpha/2} - \frac{\sqrt{M(1-\pi-\pi^2)}\cdot d}{s}\right) \quad (7)$$

To examine the impacts of M and n on the power when $0 < \rho < 1$, the amounts that need to be increased on M and n to achieve the same power are calculated. With other factors fixed and for a given n and M, how much does n need to be increased to achieve the same impact on power when increasing *M* by 1? Recall the power function is

$$\text{Power} = 1-\beta = 1-\Phi\left(z_{1-\alpha/2} - \frac{\sqrt{nM(1-\pi-\pi^2)}\cdot d}{\sqrt{1+(n-1)\rho}\cdot s}\right)$$

.

With other factors fixed, all that is required is to make the term,

$$\frac{nM}{1+(n-1)\rho},$$

a constant to achieve the same power. Let n' be the new n that will have the same impact on power as M increased by 1. Then the following equation can be solved

$$\frac{n(M+1)}{1+(n-1)\rho} = \frac{n'M}{1+(n'-1)\rho},$$

and the following equation is obtained:

$$n' = \frac{n(M+1)(1-\rho)}{M-(M+n)\rho} \quad (8)$$

Thus increasing *n* by the amount,

$$n' - n = \frac{n(1-\rho+n\rho)}{M-(M+n)\rho} \quad (9)$$

is the same as increasing M by 1. This amount of increment depends on M, n and ρ. For example, if $\rho = 0.5$, then n needs to increase by $n(1+n)/(M-n)$; if $\rho = 0.05$ n needs to increase by $n(0.95+0.05n)/(0.95M-0.05n)$ in order to have the same impact on power as M increased by 1.

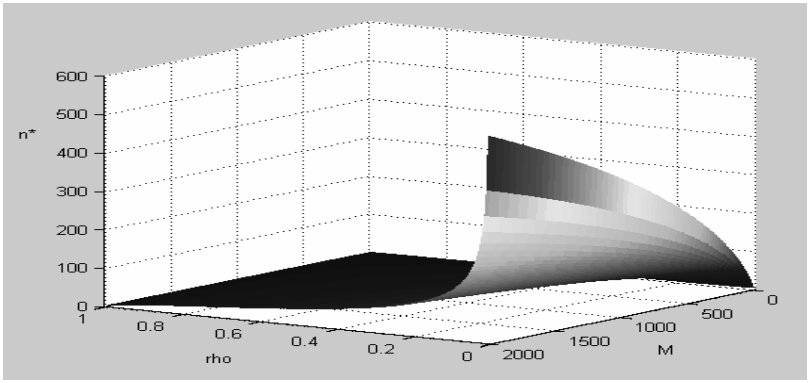
To examine which variable, M or n, has a larger impact on the power, it is required that one checks which variable needs to increase more to get the same power. The larger amount that needs to increase, the lower impact the variable has on statistical power. Set (9) equal to 1 and obtain the following equation.

$$\rho n^2 + n - (1-\rho)M = 0 \quad (10)$$

This is a quadratic function of n, and thus it has two roots

$$n^* = \frac{-1 \pm \sqrt{1+4\rho(1-\rho)M}}{2\rho} \quad (11)$$

Figure 1. The Relationship of n*, ρ and M.



Because n is always greater than 0, the positive root is taken. To say that the amount (9) is greater than 1, is equivalent to stating that equation (10) is greater than 0, or n is greater than n*, the root of (10). In other words, the impact of n on power is smaller than the impact of M when n is greater than n*. Based on (11), one can see that n* depends on both M and ρ nonlinearly. Figure 1 below shows the non-linear relationship among M, n and ρ.

This 3-D figure reveals that the threshold n* will increase when M increases but for a same M value, the threshold will be larger when ρ smaller. Figure 2 and Figure 3 are special slides of the 3-D figure of Figure 1. Figure 2 shows the relationship between the threshold n* and ρ for M=300 and Figure 3 shows the relationship between the threshold n* and M for ρ =0.4.

Figure 2. The Relationship of n* and ρ, with M = 300 fixed.

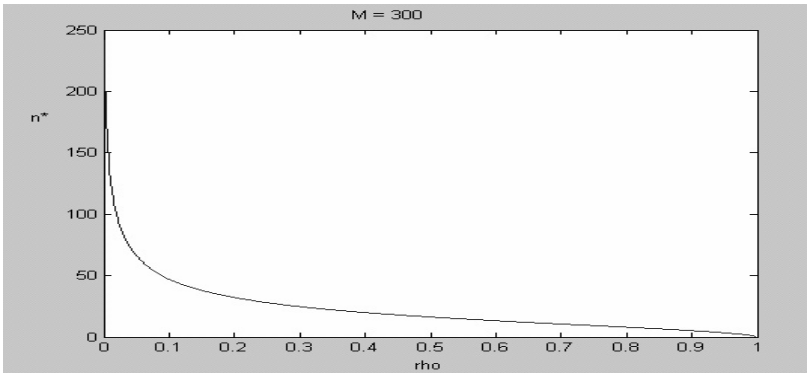
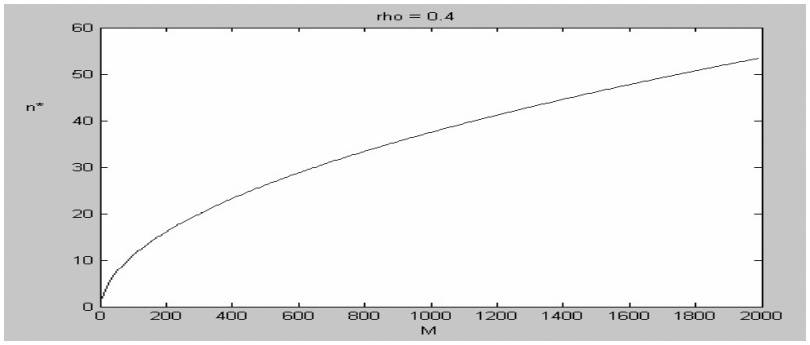


Figure 3. The Relationship of n* and M, with ρ = 0.4 fixed.



%SP_TAD Software, Syntax and Parameters

To implement the algorithm for calculating the sample sizes or power for time-averaged difference, we have written a statistical macro procedure %SP_TAD, where SP stands for sample and power, TAD stands for time averaged difference in SAS/MACRO.

The syntax of the macro is simple and straightforward. To use this macro, one simply needs to invoke the macro with specific values for the parameters required. Here is the list of parameters that need to be specified:

(1) type-------continuous (=1) or binary (=2) responses. This sets up the tone of the type of the outcome measure to be analyzed. The following parameters of (2) to (9) must be provided for continuous responses:
(2) alpha----Type I error rate
(3) beta----- Type II error rate
(4) d--------Smallest meaningful difference to be detected
(5) sigma----Measurement deviation (for continuous responses)
(6) n--------Number of repeated observations per subject
(7) rho------Correlation among each subject
(8) pi---------Proportion of number of subjects within group 1
(9) M--------Total number subjects

For binary outcome, sigma is not needed. Instead, two more parameters need to be provided:

(10) pa-------Pr(Y_ij=1) in group 1
(11) pb-------Pr(Y_ij=1) in group 2

To run the macro, one needs simply to issue:

%sp_tad(type=, alpha=, beta=, d=, sigma=, n=, rho=, pi=, pa=, pb=, M=);

where pa and pb should be left as blank for continuous outcome, and sigma should be left blank for binary outcome. Beta and M should not be provided at the same time. To calculate required sample size, beta must be provided. To calculate power, M must be provided. Type is 1 or 2, where 1 stands for continuous responses and 2 stands for binary responses. The software code is available upon request from the author.

Application

Repeated measures design has wide applications in social, biological, medical and health service research. To avoid possible bias from snapshot of data collection at one time point and to reduce the cost of collecting data from different subjects, repeated measures data are often collected. Through a real example, this section demonstrates the input, output and the functionality of the %SP_TAD software and how the procedure works with continuous outcome measures. For binary outcome measures, the process will be similar.

For continuous measures, an example of a patient's diastolic blood pressure between a treatment and control group is examined (generally, diastolic blood pressure below 85 is considered "normal"). The level of a person's blood pressure could be affected by many temporary factors, such as the amount of time that the person slept last night, the person's mood, physical activity right before taking blood pressure measurement, etc. Thus, a one time snapshot of blood pressure will likely not be accurate. To accurately estimate the level of blood pressure of a patient or a group of patients, means of multiple measurements of blood pressure from a patient are usually used.

Suppose that a design is required to examine the difference of diastolic blood pressure between the treatment and control groups. To avoid bias from one time snapshot, five repeated measures of blood readings were taken from each patient within a week (one reading each day). Based on previous studies, intra-class correlation at the level of 0.4, type I error 0.05 and type II error 0.15 and a common standard deviation of 15 was used. Assume that a difference in mean blood pressure as small as 10 points between the treatment and control groups is desired. Since the treatment is more expensive than the control and more controls than treatment participants is desired, with a ratio of 3:2. Using these parameters, the calculation with the following syntax can be established:

%***sp_tad***(type=**1**, alpha=**0.05**, beta=**0.15**, d=**10**, sigma=**15**, n=**5**, rho=**0.4**, pi=**0.6**, pa=, pb=, M=);

Execute the procedure and the answer is 158 in treatment group and 105 in control group. Assume that the control group had a mean diastolic blood pressure 88. Then, the given sample size of 158 in the treatment group and 105 in the control group with 5 repeated measurements from each patient will allow one to detect a mean diastolic blood pressure of the treatment as low as 78.

For the same question, assume 158 patients in treatment group and 105 patients in the control group with 5 repeated measures of blood pressure. With a type I error 0.05, what kind of power will is needed to detect a difference in mean blood pressure of as small as 10 points? Using the same procedure, these parameters can be instituted and the macro with the following syntax can be executed:

%***sp_tad***(type=**1**, alpha=**0.05**, beta=, d=**10**, sigma=**15**, n=**5**, rho=**0.4**, pi=**0.6**, pa=, pb=, M=263);

The answer for power will be 85%.

Conclusion

Time-averaged difference of repeated measures data has wide applications in many fields of research. TAD provides the opportunity to examine the difference in means between groups with higher precision using repeated measurements from each subject. This article deals with sample size and power analyses issues for time-averaged difference of repeated measures design. It presents the details of derivation of the general sample size calculation and power analysis formula for TAD with unequal sample size between two groups. Allowing unequal sample size will enable researchers to have the opportunity to choose an unbalanced design so that smaller number of

subjects could be used for the group that is either more expensive, hard to recruit or with limited number of available subjects.

Repeated measures data points also arise from cluster randomized trials, where it typically has repeated individuals within randomized clusters. There is growing literature on the topic starting with initial work involving balanced equally sized groups, but is now extending to more complex situations, of which unequal group sizes is also a possible scenario (Eldridge, 2001).

Repeated measures data has two dimensions of sample sizes: the number of different individuals and the number of repeated measurements from each individual. As shown in the article, because data from different individuals are independent, the number of different subjects seems to have a larger effect on power than the number of repeated measurements from the same subject. However, there is a threshold of the number of repeated measures, which will yield a larger impact by increasing the number of repeated measures than by increasing the number of subjects on statistical power. However, increasing the number of subjects by 1 means to increase the number of observations by n (the new subject gets n repeated measurements as others) and increasing the number of repeated measures by 1 means to increase the number of observations by M (every subject increases one repeated measurement). Thus, when ρ is very small (i.e. about zero), one will need a larger n to exceed n*, the threshold, in order to have a larger impact of increment of n than M on power.

In most of the situations, n is not large and much smaller than M, thus likely M will have larger impact than n. For the two extreme cases where $\rho = 0$ or $\rho = 1$, the impact of the increase of the number of repeated measures will be the same as the increase of the number of individuals in each group ($\rho = 0$) or there will be no impact of increasing the number of repeated measures ($\rho = 1$) on power.

The software created is easy to use and can handle both continuous outcome measure and dichotomous outcome measure by issuing a value of "1" or "0" for the parameter "type". For the same software, one can also estimate the underlying statistical power for a given sample size with a given type I error, type II error, variation and effect size.

References

Agresti, A. (1996). *An introduction tocategorical data analysis*. Wiley: New York.

Cohen, J. (1988). *Statistical poweranalysis for the behavioral sciences*. Second edition. Lawrence Erlbaum Associates: Hove and London.

Diggle, P. J., Liang, K. Y., & Zegger,S. L. (1994). *Analysis of longitudinal data*. Oxford University Press: Oxford.

Dixson, W. J., Massey, F. J. (1969).*Introduction to statistical analysis*. McGraw-Hill: New York.

Elashoff, J. D. (2000). *nQuery advisor* (Version 4.0.). Statistical Solutions: Cork, Ireland.

Lenth, R. V. (1987). *"PowerPack," Software for IBM PCs and compatibles. Provides an interactive environment for power and sample-size calculations and graphics.*

Eldridge, S., Cryer, C., Defer, G., & Underwood, M. (2001). Sample size calculation for intervention trials in primary care randomizing by primary care group: an empirical illustration from one proposed intervention trial. *Statistics in Medicine 20*(3), 367-376.

Fitzmaurice, G. M., Laird, N. M., & Ware, J. H. (2004). *Applied longitudinal analysis*. Wiley: Chichester.

Hsieh, F. Y., Block, D. A., & Larsen, M. D. (1998). A simple method of sample size calculation for linear and logistic regression. *Statistics in Medicine, 7*, 1623-1634.

Machin, D., Campbell, M., Fayers, P., &Pinol, A. (1997). *Sample size tables for clinical lstudies* (2nd ed.). London: Blackwell Science.

NCSS Statistical Software, NCSS: Kaysville, Utah, 2002.

SAS/IML, User's Guide, Version 8. SAS Institute Inc: Cary , NC, 1999.

SAS/STAT, User's Guide, Version 9. SAS Institute Inc: Cary, NC, 1999.

Scheffé, H. (1959). *The analysis ofvariance*. Wiley: New York.

S-PLUS 2000 User's Guide, MathSoft Data Analysis Products Division: Seattle, WA, 1999.

SPSS Base 10.0 for Windows User's Guide. SPSS Inc.: Chicago IL, 1999.

STATA, Version 8. STATA Press:Texas. 2003.

Tierney, L. (1990). *Lisp-Stat, an object-oriented environment for dynamic graphics*. Wiley: New York.

Whittemore, A. (1981). Sample size forlogistic regression with small response probability. *Journal of the American Statistical Association, 76,* 27-32.