Lecture 6

Power and Sample Size in Linear Mixed Effects Models

Ziad Taib
Biostatistics, AZ
MV, CTH

May 2009

Outline of lecture 6

1. Generalities
2. Power and sample size under linear mixed model assumption
3. Example 1: The rat data
4. Estimating sample size using simulations
5. Example 2: COPD

- **"The number of subjects in a clinical study should always be large enough to provide a reliable answer to the question(s) addressed."**
- **"The sample size is usually determined by the primary objective of the trial."**
- **" Sample size calculation should be explicitly mentioned in the protocol ."**

(from ICH-E9)

Power and sample size

- Suppose we want to test if a drug is better than a placebo, or if a higher dose is better than a lower dose.
- Sample size: How many patients should we include in our clinical trial, to give ourselves a good chance of detecting any effects of the drug?
- Power: Assuming that the drug has an effect, what is the probability that our clinical trial will give a significant result?

Sample Size and Power

- **Sample size is contingent on design, analysis plan, and outcome**
- **With the wrong sample size, you will either**
  - Not be able to make conclusions because the study is "underpowered"
  - Waste time and money because your study is larger than it needed to be to answer the question of interest

Sample Size and Power

- **With wrong sample size, you might have problems interpreting your result:**
  - Did I not find a significant result because the treatment does not work, or because my sample size is too small?
  - Did the treatment REALLY work, or is the effect I saw too small to warrant further consideration of this treatment?
  - Issue of CLINICAL versus STATISTICAL significance

Sample Size and Power

- **Sample size ALWAYS requires the scientist/investigator to make some assumptions**
  - How much better *do you expect* the experimental therapy group to perform than the standard therapy groups?
  - How much variability *do we expect* in measurements?
  - What *would be* a clinically relevant improvement?
- **The statistician alone CANNOT tell what these numbers should be. It is the responsibility of the scientist/clinical investigators to help in defining these parameters.**

AstraZeneca

Errors with Hypothesis Testing

		Accept H_0	Reject H_A
Ho true	→ E = C	OK	Type I error
Ho false	→ E ≠ C	Type II error	OK

Ho: E=C (ineffective) H1: E≠C (effective)

AstraZeneca

Type I Error

Concluding for alternative hypothesis while null-hypothesis is true (false positive)

- **Probability of Type I error = *significance level* or α *level***
- **Value needs to be pre-specified in the protocol and should be small**
- **Explicit guidance from regulatory authorities, e.g. *0.05***
- **One-sided or two-sided ?**
- **H_0: drug ineffective vs. H_A: drug effective**

AstraZeneca

Type II Error

Concluding for null-hypothesis while alternative hypothesis is true (false negative)

- **Probability of Type II error = β *level***
- **Value to be chosen by the sponsor, and should be small**
- **No explicit guidance from regulatory authorities, e.g. 0.20, or 0.10**
- **Power = 1 - Type II error rate. Typical values 0.80 or 0.90**

AstraZeneca

Sample Size Calculation

- Following items should be specified
  - a primary variable
  - the statistical test method
  - the null hypothesis; the alternative hypothesis; the study design
  - the Type I error
  - the Type II error
  - way how to deal with treatment withdrawals

AstraZeneca

Power calculations in practice

- In the planning and development stage of an experiment, a sample size calculation is a critical step. In controlled clinical trials, sample size calculation is required to maintain specific statistical power (i.e. 80% power) for the study. Not surprisingly, there are many software packages (i.e. nQuery, PASS, StudySize etc) which perform sample size calculation for certain statistical tests. However, such software packages are not available for sample size calculations for complex designs and complex statistical analysis methods.

AstraZeneca

Power Calculations Under Linear Mixed Models

- In this lecture, we will reflect on the design of longitudinal studies.
  1. We will briefly discuss how power calculations can be performed based on linear mixed models.
  2. In practice longitudinal experiments often do not yield the amount of information hoped for at the design stage, due to dropout. This results in realized experiments with (possibly much) less power than originally planned.
  3. We will discuss how expected dropout can be taken into account in sample-size calculations. The basic idea behind this is that two designs with equal power under the absence of dropout are not necessarily equally likely to yield realized experiments with high power.
  4. The main question then is how to design experiments with minimal risk of huge losses in efficiency due to dropout.
  5. The above is illustrated in the context of the rat experiment.

Power Calculations Under Linear Mixed Models

- We have discussed inference for the marginal linear mixed model. Several testing procedures were discussed, including
  - approximate Wald tests,
  - approximate t-tests,
  - approximate F -tests,
  - and likelihood ratio tests (based on ML as well as REML estimation), for the fixed effects as well as for the variance components in the model.
- Obviously, any of these testing procedures can be used in power calculations.

.

Power Calculations Under Linear Mixed Models

- **Unfortunately**, the distribution of many of the corresponding test statistics is only known under the null hypothesis.
- In practice, this means that if such tests are to be used in sample-size calculations, extensive simulations would be required.
- One then would have to
  - sample data sets under the alternative hypothesis of interest,
  - analyze each of them using the selected testing procedure,
  - and estimate the probability of correctly rejecting the null hypothesis.
  - Finally, this whole procedure would have to be repeated for every new design under consideration

Power Calculations Under Linear Mixed Models

- Assume we are interested in a general linear hypothesis of the forms

$$\boldsymbol{H}_0 : \boldsymbol{\xi} = \boldsymbol{L}\boldsymbol{\beta} - \boldsymbol{\xi}_0 = 0, \text{versus } \boldsymbol{H}_A : \boldsymbol{\xi} \neq 0$$

- Then we can use the (Under H_0) approximately F-distributed statistic

$$\boldsymbol{F} = \hat{\boldsymbol{\xi}}' \left[\boldsymbol{L} \left(\sum_i \boldsymbol{X}_i' \boldsymbol{V}_i^{-1} \boldsymbol{X}_i \right)^{-1} \boldsymbol{L}' \right] \hat{\boldsymbol{\xi}} / (\boldsymbol{rank}(\boldsymbol{L}))$$

Power Calculations Under Linear Mixed Models

- Helmert (1992) reports that under the alternative hypothesis H_A , the distribution of F can also be approximated by an F-distribution, now with rank(L) and $\sum_i n_i$ - rank[X|Z] degrees of freedom and with non-centrality parameter:

$$\delta = \xi' \left[L \left(\sum_i X_i' V_i^{-1} X_i \right)^{-1} L' \right] \xi$$

- With notation as in previous lectures

The rat data

- The hypothesis of primary interest is H 0 : no effect, which turns out to be non-significant using an approximate Wald statistic (p=0.0987). A similar result (p=0.1010) is obtained using an approximate F -test, with Satterthwaite approximation for the denominator degrees of freedom. We conclude from this that there is little evidence for any treatment effects. However, the power for detecting the observed differences at the 5% level of significance and calculated using the F -approximation described in the previous section is as low as 56%.

$$\boldsymbol{Y}_{ij} = \begin{cases} \boldsymbol{\beta}_0 + \boldsymbol{b}_i + \boldsymbol{\beta}_1 \boldsymbol{t}_{ij} + \boldsymbol{\varepsilon}_{ij} & \text{if low dose} \\ \boldsymbol{\beta}_0 + \boldsymbol{b}_i + \boldsymbol{\beta}_2 \boldsymbol{t}_{ij} + \boldsymbol{\varepsilon}_{ij} & \text{if high dose} \\ \boldsymbol{\beta}_0 + \boldsymbol{b}_i + \boldsymbol{\beta}_2 \boldsymbol{t}_{ij} + \boldsymbol{\varepsilon}_{ij} & \text{if control dose} \end{cases}$$

- Note that, this rat experiment suffers from a severe degree of dropout, since many rats do not survive anesthesia needed to measure the outcome. Indeed, although 50 rats have been randomized at the start of the experiment, only 22 of them survived the 6 first measurements, so measurements on only 22 rats are available in the way anticipated at the design stage. For example, at the second occasion (age = 60 days), only 46 rats were available, implying that for 4 rats, only 1 measurement has been recorded. As can be expected, this high dropout rate inevitably leads to severe losses in efficiency of the statistical inferential procedures. Indeed, if no dropout had occurred (i.e., if all 50 rats would have withstood the 7 measurements), the power for detecting the observed differences at the 5% level of significance would have been 74%, rather than the 56% previously reported for the realized experiment.

Conclusion

- In the rat example, dropout was not entirely unexpected since it is inherently related to the way the response of interest is actually measured (anesthesia cannot be avoided) and should therefore have been taken into account at the design stage. Therefore we need methods for the design of longitudinal experiments, when dropout is to be expected. We will discuss such an approach and apply it to the rat data.

Power Calculations When Dropout Is to Be Expected

- In order to fully understand how the dropout process can be taken into account at the design stage, we first investigate how it affects the power of a realized experiment. Note that the power of the above F -test not only depends on the true parameter values β, D, and σ^2 (or, more generally, Σ_i) but also on the covariates X_i and Z_i. Usually, in designed experiments, many subjects will have the same covariates, such that there are only a small number of different sets (X_i ,Z_i).

- For the rat data, all 15 rats in the control group have X_i and Z_i equal to

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & 0 & 0 & \ln\left[1+\frac{(50-45)}{10}\right] \\ 1 & 0 & 0 & \ln\left[1+\frac{(60-45)}{10}\right] \\ . & . & . & . \\ . & . & . & . \\ . & 0 & 0 & . \\ 1 & 0 & 0 & \ln\left[1+\frac{(110-45)}{10}\right] \end{pmatrix}, \quad Z_i = \begin{pmatrix} 1 \\ 1 \\ . \\ . \\ . \\ 1 \end{pmatrix}$$

- However, due to the dropout mechanism, the above matrices have been realized for only four of them. Indeed, for a rat that drops out early, say at the kth occasion, the realized design matrices equal the first k rows of the above planned matrices; that is,

$$\boldsymbol{X}_i = \begin{pmatrix} 1 & 0 & 0 & \ln\left[1+\frac{(50-45)}{10}\right] \\ 1 & 0 & 0 & \ln\left[1+\frac{(60-45)}{10}\right] \\ . & . & . & . \\ . & . & . & . \\ . & 0 & 0 & . \\ 1 & 0 & 0 & \ln\left[1+\frac{(40+\boldsymbol{k}\times 10-45)}{10}\right] \end{pmatrix}, \quad Z_i = \begin{pmatrix} 1 \\ 1 \\ . \\ . \\ . \\ 1 \end{pmatrix}$$

- Note that the number of rats that drop out at each occasion is a realization of the stochastic dropout process, from which it follows that the power of the realized experiment is also a realization of a random variable, the distribution of which depends on the planned design and on the dropout process. From now on, we will denote this random power function by *P*.
- Since, in the presence of dropout, the power P becomes a stochastic variable, it is not obvious how two different designs with two different associated power functions P_1 and P_2 should be compared in practice. Several criteria can be used, such as the average power, $E(P_1)$, the median power, median(P_1), the risk of having a final analysis with power less than for example 70%, P (P < 70%), and so forth.

- Note that all of the above criteria are based on only one specific aspect of the distribution of P_1. A criterion which takes into account the full distribution selects the second design over the first one if P_1 is stochastically smaller than P_2 , which is defined as (Lehmann and D'Abrera 1975, p. 66)

P_1 is stochastically smaller than P_2

$\Leftrightarrow$

$P(P_1 \leq p) \geq P(P_1 < p)$ for all p

This means that, for any power p, the risk of ending up with a final analysis with power less than p is smaller for the second design than for the first design.

- Obviously, if the above criterion is to be used, one needs to assess the complete power distribution function for all designs which are to be compared. We propose doing this via sampling methods in which, for each design under consideration, a large number of realized values p_s , s=1,...,S, are sampled from P and used to construct the empirical distribution function below where $I[A]$ equals one if A is true and zero otherwise.

$$\hat{P}(P \leq p) = \frac{1}{S}\sum_{s=1}^{S} I[p_s \leq p]$$

- As indicated above, sampling from P actually comes down to sampling realized values for all $M_{j,k}$, k=1,...,n_j , j=1,...,M , and constructing all necessary realized matrices $X_j^{[k]}$ and $Z_j^{[k]}$. One then can easily calculate the implied non-centrality parameter δ and the appropriate numbers of degrees of freedom for the F-statistic, from which a realized power follows.
- It should be emphasized that the above approach is not restricted to any particular statistical test. The idea of sampling designs under specific dropout patterns is applicable for any testing procedure, as long as it remains possible to evaluate the power associated to each realized design.

- Note also that the only additional information needed, in comparison to classical power analyses, are the vectors p_j of marginal dropout probabilities $p_{j,k}$. This does not require full knowledge of the underlying dropout process.
- We only need to make assumptions about the dropout rate at each occasion where observations are designed to be taken.
- For example, we do not need to know whether the dropout mechanism is "completely at random" or "at random".

- Still, we have to assume that dropout is "not informative" in the sense that it does not depend on the response values which would have been recorded if no dropout had occurred, since otherwise our final analysis based on the linear mixed model would not yield valid results (see Section 15.8 and Chapter 21).
- Finally, the proposed method can be used in combination with techniques, such as those proposed by Helms (1992), which would allow the costs of performing the designs under consideration to be taken into account. This could yield less costly experiments with minimal risk of large efficiency losses due to dropout. This will not be explored any further here.

The rat data

- Observed conditional dropout rates at each occasion, for all treatment groups simultaneously.
- Age (days):
- Observed rate:

50	60	70	80	90	100
0.08	0.07	0.12	0.24	0.17	0.08

- Based on the data we assume that each time a rat is anesthetized, there is about 12% chance that the rat will not survive anesthesia, independent of the treatment.
- All calculations are done under the assumption that the true parameter values are given by earlier estimates and all simulated power distributions are based on 1000 draws from the correct distribution.

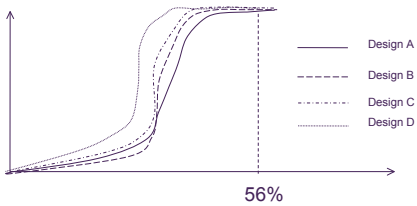
- Since, at each occasion, rats may die, it seems natural to reduce the number of occasions at which measurements are taken. We have therefore simulated the power distribution of four designs in which the number of rats assigned to each treatment group is the same as in the original experiment, but the planned number of measurements per subject is seven, four, three, and two, respectively.
- These are the designs A to I in the table below. Note that design A is the design used in the original rat experiment. The simulated power distributions are shown in the figure.

Rat Data. Summary of the designs compared in the simulation study when varying group sizes

Design	Occasions Age (days)	Number of subjects (M ,M ,M)	Power if no dropout
A	50-60-70-80-90-100-110	(15, 18, 17)	0.74
B	50-70-90-110	(15, 18, 17)	0.63
C	50-80-110	(15, 18, 17)	0.59
D	50-110	(15, 18, 17)	0.53
E	50-70-90-110	(22, 22, 22)	0.74
F	50-80-110	(24, 24, 24)	0.74
G	50-110	(27, 27, 27)	0.75
H	50-60-110	(26, 26, 26)	0.74
I	50-100-110	(20, 20, 20)	0.73

A: Original design

- Comparison of the simulated power distributions for designs with seven, four, three, or two measurements per rat, with equal number of rats in each design (designs A, B, C, and D, respectively), under the assumption of constant dropout rate equal to 12%. The vertical dashed line corresponds to the power which was realized in the original rat experiment (56%).



- First, note that the solid line is an estimate for the power function of the originally designed rat experiment under the assumption of constant dropout probability equal to 12%.
- It shows that there was more than 80% chance for the final analysis to have realized power less than the 56% which was observed in the actual experiment.
- Comparing the four designs under consideration, we observe that the risk of high power losses increases as the planned number of measurements per subject decreases.
- On the other hand, it should be emphasized that the four designs are, strictly speaking, not comparable in the sense that, in the absence of dropout, they have very different powers ranging from 74% for design A to only 53% for design D.

- Designs E, F, and G are the same as designs B, C, and D, but with sample sizes such that their power is approximately the same as the power of design A, in the absence of dropout.
- The simulated power distributions are shown in Figure 23.2 in the book. The figure suggests that P > A > E > F > G , from which it follows that, in practice, the design in which subjects are measured only at the beginning and at the end of the study is to be preferred, under the assumed dropout process.
- The above can be explained by the fact that the probability for surviving up to the age of 110 days is almost twice as high for design G (88%) as for the original design (46%).
- Note also that the parameters of interest [β_1 , β_2 , and β_3] are slopes in a linear model such that two measurements are sufficient for the parameters to be estimable. On the other hand, design G does not allow testing for possible nonlinearities in the average evolutions.

Comments

- Note that the above results fully rely on the assumed linear mixed model. For example, the simulation results show that design G, with only two observations per subject, is to be preferred over designs A, E, and F, with more than two observations scheduled for each subject.
- Obviously, the assumption of linearity is crucial here, and design G will not allow testing for nonlinearities. Hence, when interest would be in providing support for the used model, more simulations would be needed comparing the behavior of different designs under different models for the outcome under consideration, and design G should no longer be taken into account.
- As for any sample-size calculation, it would be advisable to perform some informal sensitivity analysis to investigate the impact of model assumptions and imputed parameter values on the final results.

Example: Estimating the sample size needed in a trial for chronic pulmonary diseases

- Chronic pulmonary diseases (such as Chronic Obstructive Pulmonary Disease – COPD) concern the development of emphysema. It is a slow progression over many years and the assessment of drug efficacy requires the observation of ***large numbers of patients for a long period of time***. Recently, lung densitometry (measuring the lung density through CT scan) considered for assessing the lung tissue loss over time in patients with emphysema.
- A clinical trial with lung densitometry as an endpoint is typically designed as a **longitudinal study with repeated measurements at fixed time intervals**. Since lung density measurements are closely correlated with lung volume (inspiration level), it is important to include lung volume measurements in statistical analyses as a longitudinal covariate. Lung volume is normally measured at the same time as the lung density is measured.

- The clinical efficacy can be assessed by comparing the progression of lung density loss between two treatment groups using a random coefficient model – a longitudinal linear mixed model with a random intercept and slope. *In planning the clinical trial with such complex statistical analyses, the calculation of the sample size required to achieve a given power to detect a specified treatment difference is an important, often complex issue.*
- In this example, an empirical approach is used to calculate the sample size by simulating trajectories of lung density and lung volume using SAS. We present step-by-step details for sample size calculation through simulation, and discuss the pros and cons of this approach.

$$Y_{ij} = (\beta_0 + b_0) + \beta_1 * TRT + (\beta_2 + b_2) * TIME + \beta_3 * COV_{ij} + \beta_4 * TRT * TIME + \varepsilon_{ij} \quad (1)$$

- Here Y_{ij} is the efficacy endpoint (i.e. lung density) measurement for subject *i = 1, 2,…, n, at fixed time point j = 1, 2, …, K.*
- *TRT is an indicator of subject i's treatment group (i.e. TRT=1 for active drug; TRT=0 for placebo).*
- *COV_{ij} is a longitudinal covariate (i.e. logarithm of lung volume) for subject i = 1, 2,…, n, at fixed time point j = 1, 2, …, K.*
- *Here b_0 and b_2 are subject-specific random effects for the intercept and slope, respectively, which are from a normal distribution with mean 0 and variance σ_0^2 and σ_0^2, respectively.*
- *ε_{ij} is the random error from a normal distribution with mean 0 and variance σ^2.*
- *The regression parameters β_0, β_1, β_2, β_3, and β_4 are the fixed effects for intercept, treatment, time, covariate and interaction of treatment and time respectively.*
- Here we assume that the benefits can be assessed quantitatively by comparing the slopes of lung density trajectories for the two treatment groups. This quantity is captured by β_4.

Sample Size Estimation Using Simulations

- In the model, β_4 is typically our interest, which is the difference in slope of time between two treatment groups (active vs. placebo). There is no direct mathematical formula to calculate the sample size for a given statistical power (i.e. 80%) to test the null hypothesis: $\beta_4=0$ with a specified type I error (i.e. $\alpha=0.05$). One approach to calculate the sample size for a given power is through the simulation.

- Assume we know the parameters (*β_0, β_1, β_2, β_3, and β_4* , and *σ_0^2 and σ_0^2*) from either history data, previous clinical trials or meaningful clinical differences we want to test, the study design in terms of number of time points (*K*) and fixed time intervals (*TIME*), and the longitudinal covariate *COV_{ij}*. *For a fixed equal sample size n for each treatment, the trajectories of efficacy measurement Y_{ij} (i.e. lung density) for the n subjects can be simulated through the model for each treatment group.*
- *Then, perform a statistical test on $\beta_4 =0$ by using the SAS Proc MIXED on the simulated data set, and record whether the p-value < 0.05.*

- The sample code to perform the test is as follow:

```
proc mixed data = data;
class id trt;
model y = trt time trt*time cov / solution;
random intercept time/ subject = id type = un;
run;
```

- For the fixed sample size *n per treatment group, simulate M (i.e. M=1000) times and the proportion of significance test on $\beta_4 =0$ among the total M simulations is the statistical power for the sample size n per treatment group. Then, adjust the sample size n to achieve desirable statistical power.*

Simulating the response

- In order to simulate the trajectories of Y_{ij}, it is necessary to simulate the trajectories of longitudinal covariate COV_{ij}. Similarly, assume COV_{ij} is from a linear model regressing against time with a random intercept

$$COV_{ij} = (\gamma_0 + r_0) + \gamma_1 TIME + \varepsilon_{ij} \qquad (2)$$

- Where γ_0 and γ_1 are the fixed intercept and slope respectively; r_0 and ε_{ij} are from a normal distribution with mean 0 and variance δ_1^2 and δ_2^2, respectively. If we know the parameters (γ_0, γ_1 , δ_1^2 and δ_2^2) from history data or previous clinical trials for the study population, it will be simple to simulate the trajectories of the longitudinal covariate COV_{ij} by using SAS random generating functions.

- In detail, a sample size can be determined for the models above through the following steps:

1. Obtain the pre-specified parameters through either history data, previous clinical trials or meaningful clinical difference to be tested from clinicians
2. Specify a desired statistical power (i.e. 80%) and a type-1 error rate (i.e. 5%)
3. Simulate trajectories of efficacy measurement (i.e. lung density) and longitudinal covariate (i.e. logarithm of lung volume) for a fixed sample size (*n) of subjects within each treatment arm*
   - A. Trajectories of longitudinal covariate (i.e. logarithm of lung volume) are simulated through model (2)
   - B. Trajectories of efficacy measurement (i.e. lung density) are simulated through model (1)

4. Perform the statistical test on β4=0 based on the simulated data set. Record whether a p-value < 0.05 was obtained
5. Repeat steps 3 and 4 M (i.e. M=1000) times and calculate the statistical power for the fixed sample size
6. Repeat steps 3 - 5 for various values of *n. Stop when desired statistical power is obtained*

Results: Example of a Simulation

- Assume there are two treatment groups (active vs. placebo) in a study design. The efficacy endpoint along with the longitudinal covariate will be measured at *K=4 time points at baseline, 1 year, 2 years and 3 years. All corresponding parameters specified in model (1) and (2) could be obtained either through history data, previous clinical trials or meaningful clinical difference to be tested from clinicians. For purpose of simulation, they are randomly selected and specified as below:*

$$\beta_0 = 150, \beta_1 = 5, \beta_2 = -1.8, \beta_3 = -57, \beta_4 = 0.7, \text{ and } \sigma_0^2 = 280, \sigma_2^2 = 0.4, \sigma^2 = 5;$$
$$\gamma_0 = 2, \gamma_1 = 0.0007, \delta_0^2 = 0.05, \delta_1^2 = 0.0016.$$

- The summary of statistical power for a given sample size per treatment based on M = 1000 simulated data sets is listed below:

N per treatment	Statistical Power (%)
30	62.4
40	76.9
45	**79.9**
50	84.4
60	91.3

- Therefore, a sample size 45 per treatment arm has an estimated statistical 80% power to detect the treatment slope difference of 0.7 in a random coefficient model for the study design above.

Conclusions and Discussion

- As described above, it is possible to perform sample size calculations for a random coefficient model using simulation techniques and SAS. It is also straightforward to extend the simulation frame to other linear mixed models (LMM) or generalized linear mixed models (GLMM). Other extensions to settings involving multiple treatment groups (i.e. treatment groups greater than 2), unequal sample size among treatment groups (i.e. 2:1 for active vs. placebo) can be implemented. For an active-controlled trial, it is usually interest to test non-inferiority of test drug compared to active-control. The simulation frame can be applied for such non-inferiority test by calculating the confidence interval for the parameter tested in the statistical model and comparing the lower limit or upper limit of the confidence interval to the pre-specified equivalence margin in the Step 4.

- Other study design parameters such as number of repeated measurements (*K*) *of efficacy endpoint and the duration of the fixed time intervals (time) also affect the sample size estimation. Greater number of repeated measurements of efficacy endpoint for the fixed study duration will increase the statistical power. However, it might increase the difficulty and cost of the study depending on the efficacy endpoint. The number of repeated measurement of efficacy endpoint and duration of the fixed time intervals should be determined within the clinical research team upon the constraints such as the difficulty of efficacy endpoint measurement, cost and duration of the clinical trial.*

- In practice, it is rarely the case that all subjects have the complete data for all visits in the study because of missing certain study visits, drop out or other reasons. Since our simulation framework assumes there are no missing observations, we recommend that the implemented sample size for the designed trial include more subjects than the number estimated from the simulation. In most cases an increase of 5% or 10% should suffice, but depending on the characteristics of the designed trial such as the study population, difficulty of study procedure, difficulty of study measurement etc to cause the subject's drop out or missing of study measurements. The appropriate percentage could vary.

Post-Hoc Power (also known as observed power or retrospective power)

You have collected the data, ran an appropriate statistical analysis, and did not observe statistical significance as indicated by a relatively "large" p-value. So you decide to compute **post-hoc power** to see how powerful the test was, which, by itself is essentially an empty, meaningless result. Of course the statistical test wasn't powerful enough -- that's why the p-value isn't significant. Post-hoc power is merely a one-to-one transformation of the p-value (based on the F-statistic and degrees of freedom as illustrated above). In this situation power was computed based only on what this particular sample data showed: the observed difference in means, the computed standard error, and the actual sample sizes of the groups all contributed to the observed "power" exactly as they did to the p-value.

Post hoc power also assumes the observed results establish the minimum effect size that you would like to detect; that is, the minimum observed difference in means is now dictated by the data and is not based on your knowledge of the subject matter as to what difference would be meaningful in relation to the objectives of the study. Observed results may help you interpret the sources of variability better, but if you now compute power with different group sizes or if you want to detect a different minimum effect size, the question immediately becomes prospective. What were formerly sample statistics are elevated to the status of population parameters. So, power calculations can only be considered as a prospective or an "a priori" concept. Power calculations should be directed towards planning a study, not an after-theexperiment review of the results.

None of the SAS statistical procedures (e.g., PROCs REG, TTEST, GLM, or MIXED and others) provide retrospective (post hoc) power calculations. (However, through saving results from PROC MIXED with the ODS and following through with a few basic SAS functions, it is quite simple to compute them in a DATA step or with the inputs to PROC POWER or PROC GLMPOWER.) SAS developers know these computations produce misleading and biased results and thus won't automatically do it for you (although they are commonly found in the output from other statistical procedures and all-too-often are requested by some journals and their reviewers). See Hoenig and Heisey, 2001, for reasons behind this fallacious thinking.

References

1. Hoenig, John M. and Heisey, Dennis M. (2001), "The Abuse of Power: The Pervasive Fallacy of Power Calculations for Data Analysis," The American Statistician, 55, 19-24.
2. Lenth, R. V. (2001), "Some Practical Guidelines for Effective Sample Size Determination," The American Statistician, 55, 187-193
3. Lenth, R. V. (2006) Java Applets for Power and Sample Size (Computer Software). Retrieved 08/15/2007 from http://www.math.uiowa.edu/~rlenth/Power/
4. Littell, Ramon C., George A. Milliken, Walter W. Stroup, Russell D. Wolfinger, and Oliver Schabenberger. 2006. SAS@ for Mixed Models, Second Edition. Cary, NC: SAS Institute.