So far...

- We have discussed the role of phase III clinical trials in drug development
- We discussed the goal and unique characteristics of phase III clinical trials as confirmatory trials
- We discussed several approaches to adaptive designs in phase III
  - sequential testing
  - adaptively incorporating historical information
  - Seamless phase II/III designs
Sample Size Calculations

- Phase III clinical trials serve as the final evaluation before the drug is approved.
- One of the most important steps in designing a clinical trial is calculating the sample size.
- We require a large enough sample size to fully evaluate our scientific question.
- Specifically, we desire a high probability of rejecting the null hypothesis if the alternative hypothesis is true.
The following information is needed to calculate the required sample size:

- null hypothesis and test being used
- desired type-I error rate
- desired power
- alternative hypothesis to base power calculations
- values of important nuisance parameters
• Clinical trials are designed to reject the null hypothesis with high power assuming that a pre-specified alternative hypothesis is true.
• The alternative hypothesis is usually the expected effect size for the experimental treatment.
• Problem: what if we are unsure about the expected effect size:
  • The study population has changed throughout drug development.
  • We have imprecise estimate of the effect size due to underpowered phase II trials.
  • etc.
One approach to overcoming this difference is to power a trial to detect a clinically meaningful difference.

In this case, we have acceptable power for any large difference and do not care about smaller differences.

Problem

- It is not easy to define a clinically meaningful difference.
- Furthermore, it is very difficult to achieve unanimity on what constitutes a clinically meaningful difference.
Nuisance Parameters

• In addition to the alternative hypothesis, we also require knowledge about specific nuisance parameters
  • variance for normally distributed data
  • baseline hazard in survival analysis
  • etc.

• We do not usually have a precise estimates of the parameter of interest

• We are even less likely to have a precise estimates of nuisance parameters!
External Pilot Data

- Ideally, estimates of the alternative or nuisance parameters are available from phase II.
- Alternately, we could complete an external pilot study to estimate these parameters.
- An external pilot study will provide preliminary estimates of the alternative hypothesis and nuisance parameters.
- In addition, external pilot data allows us to evaluate our study protocol to identify potential problems in advance of the larger trial.
Disadvantages to External Pilot Data

- External pilot studies are ideal because they are independent of the primary study but are not perfect.
- Running an external pilot study is expensive and delays the start of the trial.
- This is especially true because the protocol of an external pilot study still has to go through local review (IRB, etc.) before opening.
An alternate approach is to complete an internal pilot study.

The internal pilot data can be used to estimate the unknown parameters and estimate the required sample size.

The study continues to enroll and then a standard, fixed-sample test is completed at study completion.
Re-estimation: alternative hypothesis vs. nuisance parameters

- Internal pilot data can be used to estimate the true difference and the nuisance parameters.
- Re-estimation based on the nuisance parameters will not generally impact the type-I error rate.
- Re-estimation based on the outcome (i.e. the true difference) can increase the type-I error rate substantially.
- For now, we will only consider sample size re-estimation based on the nuisance parameters.
Consider a two-arm, randomized clinical trial

Our primary outcome follows a normal distribution with unknown variance

- $x_1, x_2, \ldots \sim N(\mu_x, \sigma^2)$
- $y_1, y_2, \ldots \sim N(\mu_y, \sigma^2)$

We will test the null hypothesis, $H_0 : \mu_x = \mu_y$, using the two-sample t-test
Sample Size Calculation

- We calculate the sample size using the following formula

\[ n = \frac{(\Phi(1 - \alpha/2) + \Phi(1 - \beta))^2 2\sigma^2}{\delta^2} \]

- where
  - \( \alpha \) is the two-sided type-I error rate
  - \( \beta \) is the type-II error rate
  - \( \delta = \mu_x - \mu_y \)
Initial Sample Size Calculation

- We begin by calculating an initial sample-size
- This requires us to specify $\alpha$, $\beta$ and $\delta$
- We make an initial guess at $\sigma_0^2$ and calculate the sample size based on $\sigma_0^2$, $n_0$
For the internal pilot phase, we randomize a small number of subjects to each group \( n_p \ll n_0 \).

Using these data, we estimate the variance, \( \hat{\sigma}^2 \).

Using the estimated variance, we are able to estimate the required sample size assuming \( \alpha, \beta \) and \( \delta \).
Altering the Sample Size

- Let $n_1$ be the required sample size estimated from the initial pilot data.
- We do not alter the sample size if $n_1 < n_0$.
- If $n_1 > n_0$, then $n_1$ replaces $n_0$ as the total sample size.
- Alternately, you may only re-estimate the sample size if $n_1$ is substantially larger than $n_0$ (i.e., 25% larger).
- The trial continues until the total sample size is reached and we compare the two groups using the t-test.
• Recall the Mr Fit study
• We designed a two-arm trial assuming the following parameters
  • \( \alpha = 0.05 \)
  • \( \beta = 0.10 \)
  • \( \delta = 5 \)
  • \( \sigma = 14 \)
• This requires a sample size of 165 subjects/group
Example

- Let’s assume that an internal pilot study was run and we estimate $\hat{\sigma} = 16$
- A study with 165 subjects/group would be underpowered
- Assuming $\sigma = 16$, we would require a sample size of 216 subjects/group
- This is a substantial increase over our initial sample size estimate (31% increase)
Consider this small simulation study to investigate the operating characteristics of this design.

The initial design is based on the following parameters:

- $\alpha = 0.05$
- $\beta = 0.10$
- $\delta = 0.5$
- $\sigma = 1$

This results in an initial sample size of 85 subjects/group.

The sample size of the internal pilot study is 9 subjects/group.
Consider the following true value of $\sigma$: 1.0, 1.5, 2.0, 2.5, 3.0
These values of $\sigma$ correspond to sample sizes of 85, 190, 337, 526 and 757 subjects/group
We will consider the type-I error rate and power with and without sample size re-estimation
Simulation Results: No Re-estimation

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>$\delta = 0$</th>
<th>$\delta = 0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.051</td>
<td>0.896</td>
</tr>
<tr>
<td>1.5</td>
<td>0.046</td>
<td>0.580</td>
</tr>
<tr>
<td>2.0</td>
<td>0.048</td>
<td>0.372</td>
</tr>
<tr>
<td>2.5</td>
<td>0.054</td>
<td>0.253</td>
</tr>
<tr>
<td>3.0</td>
<td>0.048</td>
<td>0.189</td>
</tr>
</tbody>
</table>
Simulation Results: with Re-estimation

<table>
<thead>
<tr>
<th>$\sigma$</th>
<th>$\delta = 0$</th>
<th>$\delta = 0.5$</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.0</td>
<td>0.051</td>
<td>0.919</td>
</tr>
<tr>
<td>1.5</td>
<td>0.055</td>
<td>0.859</td>
</tr>
<tr>
<td>2.0</td>
<td>0.049</td>
<td>0.861</td>
</tr>
<tr>
<td>2.5</td>
<td>0.047</td>
<td>0.860</td>
</tr>
<tr>
<td>3.0</td>
<td>0.050</td>
<td>0.867</td>
</tr>
</tbody>
</table>
Simulation Study: Conclusions

- Incorrectly specifying $\sigma$ leads to dramatically underpowered studies
- Re-estimating $\sigma$ in an internal pilot study results in power close to the desired rate in all cases
Consider the slightly different scenario of comparing two binomial proportions

- $x_1, x_2, \ldots \sim Bern(p_x)$
- $y_1, y_2, \ldots \sim Bern(p_y)$

We will test the null hypothesis, $H_0 : p_x = p_y$, using the two-sample t-test
Sample Size Calculation

- We calculate the sample size using the following formula:

$$n_0 = \frac{(\Phi (1 - \alpha/2) + \Phi (1 - \beta))^2}{\delta^2} 2p^* (1 - p^*)$$

- where
  - $\alpha$ is the two-sided type-I error rate
  - $\beta$ is the type-II error rate
  - $\delta = p_x - p_y$
  - $p^* = \frac{(p_x + p_y)}{2} = \frac{(2p_x + \delta)}{2} = p_x + \frac{\delta}{2}$
• We can again re-estimate the sample size using an internal pilot study
• We enroll $n_p << n$ subjects/group
• We re-estimate the sample size, replacing $p^*$ with either
  • $p^* = \hat{p}_x + \frac{\delta}{2}$
  • $p^* = (\hat{p}_x + \hat{p}_y) / 2$
• $n = max (n_1, n_0)$
Consider this small simulation study to investigate the operating characteristics of this design.

The initial design is based on the following parameters:

- \( \alpha = 0.05 \)
- \( \beta = 0.10 \)
- \( \delta = 0.1 \)
- \( p_x = 0.1 \)

This results in an initial sample size of 268 subjects/group.

The sample size of the internal pilot study is 14 subjects/group.
Simulation Study

- Consider the following true value of $p_x$: 0.1, 0.2, 0.3, 0.4, 0.5
- These values of $\sigma$ correspond to sample sizes of 268, 395, 479, 521, and 521 subjects/group
- We will consider the type-I error rate and power with and without sample size re-estimation
Simulation Results: No Re-estimation

<table>
<thead>
<tr>
<th>( p_0 )</th>
<th>( \delta = 0 )</th>
<th>( \delta = 0.1 )</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.048</td>
<td>0.904</td>
</tr>
<tr>
<td>0.2</td>
<td>0.053</td>
<td>0.755</td>
</tr>
<tr>
<td>0.3</td>
<td>0.045</td>
<td>0.671</td>
</tr>
<tr>
<td>0.4</td>
<td>0.049</td>
<td>0.631</td>
</tr>
<tr>
<td>0.5</td>
<td>0.052</td>
<td>0.633</td>
</tr>
</tbody>
</table>
Simulation Results: with Re-estimation

<table>
<thead>
<tr>
<th>$p_0$</th>
<th>$\delta = 0$</th>
<th>$\delta = 0.1$</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.1</td>
<td>0.047</td>
<td>0.923</td>
</tr>
<tr>
<td>0.2</td>
<td>0.050</td>
<td>0.882</td>
</tr>
<tr>
<td>0.3</td>
<td>0.052</td>
<td>0.885</td>
</tr>
<tr>
<td>0.4</td>
<td>0.049</td>
<td>0.891</td>
</tr>
<tr>
<td>0.5</td>
<td>0.052</td>
<td>0.889</td>
</tr>
</tbody>
</table>
Simulation Study: Conclusions

• We observe the same pattern as before:
  • Incorrectly specifying $p_x$ leads to dramatically underpowered studies
  • Re-estimating $p_x$ in an internal pilot study results in power close to the desired rate in all cases
Internal Pilot Study: Summary

- Incorrectly specifying nuisance parameters in a power calculation can have a dramatic impact on power.
- Internal pilot data can be used to estimate these nuisance parameters.
- We can re-estimate the sample size based on these nuisance parameters.
- This results in power closer to the desired rate and does not, generally, inflate the type-I error rate.
To this point, we have only considered sample size re-estimation based on nuisance parameters.
This has little, if any, impact on the type-I error rate.
Alternately, one could use internal pilot data to estimate $\delta$ and re-estimate the sample size based on the estimated $\delta$. 
Basic setup

- The set-up is the same as when we were only estimating nuisance parameters from internal pilot data.
- We estimate an initial sample size, $n_0$, based on initial guesses of $\delta$ and any important nuisance parameters.
- We enroll an initial pilot sample, $n_p << n_0$, and estimate $\delta$ and the nuisance parameters.
- We re-estimate the sample size using these estimates, $n_1$, and set $n = max(n_0, n_1)$.
- How do you test the null hypothesis?
Naive Approach

• The simplest approach is to test the null hypothesis using a standard Z statistic

\[ Z = \hat{\delta}_n \sqrt{I_n} \]

• This can dramatically inflate the type-I error rate (Cui, Hung and Wang, 1999)
I completed a simulation study where the initial sample size was estimated using the following:

- $\alpha = 0.05$
- $\beta = 0.10$
- $\delta = 0.5$
- $\sigma = 1$

This requires a sample size of 85 subjects/group.

I used an internal pilot study of 9 subjects/group to estimate $\delta$ and $\sigma$.

This increase the type-I error to 0.10.
Unweighted Z statistic

- The naive approach weights all data equally
- That is, $Z_{np}$, is weighted equally to $Z_{n-np}$

\[ Z = \sqrt{\frac{np}{n}} Z_{np} + \sqrt{1 - \frac{np}{n}} Z_{n-np} \]

- Where $np$ is the sample size for the pilot study and $n$ is the re-estimated sample size
An alternate approach is to weight the Z statistic:

\[ Z = \sqrt{tZ_{n_p}} + \sqrt{1 - tZ_{n-n_p}} \]

Where \( t = \frac{n_p}{n_0} \)

That is, \( t \) is the ratio of the pilot sample size to the initial sample size.
The weighted Z statistic is a weighted average of the Z statistic calculated from the pilot data and the Z statistic calculated using the remaining data.

This statistic controls the type-I error rate.

It accomplishes this goal by down-weighting the later data.

This violates the principal of one patient, one vote.
• The principal of one patient, one vote says that each subject should have equal weight in the final analysis
• The weighted Z statistic gives more weight to the initial pilot study
• Furthermore, Chen, DeMets and Lan (2004) argue that weighted Z statistic is unnecessary if the interim analysis is blinded
An alternate approach to sample size re-estimation proposed by Mehta and Pocock (2010) uses conditional power.

- Conditional power refers to the conditional probability of rejecting the null hypothesis given the current data.
- This is similar to the predictive probability approach in the Bayesian paradigm.
Predictive Power Approach

- At the interim analysis, we estimate $\hat{\delta}$ and determine the conditional power assuming $\delta = \hat{\delta}$
- Mehta and Pocock consider three zones
  - unfavorable
  - promising
  - favorable
• Unfavorable refers to the cases when there is very low conditional power
• In their example, Mehta and Pocock use 36% as the cut-off
• In this case, an unreasonably large increase in the sample size would be required
• Trial is stopped for futility
Favorable

- Favorable refers to the case when there is high conditional power
- In their example, Mehta and Pocock use 90% as the cut-off
- In this case, we are likely to reject with the initial sample size and no increase is required
- The trial continues as planned
Promising

- Promising refers to intermediate values of conditional power
- In their example, Mehta and Pocock consider conditional power between 36% and 90% as promising
- These trials are promising in the sense that there appears to be a difference between the two groups but we are underpowered to detect this difference
- They suggest increasing the sample size in this case
Mehta and Pocock Design

- Determine an initial sample size based on an initial guess at the treatment effect and a maximum sample size for re-estimation.
- At the interim analysis, calculate the conditional power.
- Evaluate the conditional power and re-estimate the sample size accordingly.
- Test the null hypothesis at study completion using a modified test statistic to control the type-I error rate.
- The key to this design is that a plan for re-estimation is pre-planned, which allows them to calculate the type-I error rate and power.
Sample size estimation allows us to salvage studies that are under powered due to poor initial guesses at the effect size.

The decision to re-estimate the sample size is based on intuitive criteria based on conditional power.

All decisions are pre-specified, which allows us to evaluate the type-I error rate and power of the design.
Disadvantages

- Statistical significant is closely tied to sample size
- Any difference will be statistically significant with a large enough sample size
- A promising conditional power may represent a non clinically significantly different result, rather than an important difference that we are underpowered to detect
Disadvantages

• This design essentially represents a group sequential design with a pre-specified maximum sample size.
• You can design a design with the same type-I error rate and power by specifying the appropriate group sequential design.
• In essence, this really isn’t new and potentially complicates the issue if not done correctly.
Sample Size Re-estimation Summary

- Sample size estimation depends on the correct specification of several important design parameters.
- Incorrectly specifying these parameters can potentially lead to clinical trials that are dramatically under powered.
- Internal pilot data can be used to estimate these unknown parameters.
- Sample size re-estimation based on nuisance parameters increases power when nuisance parameters are misspecified without increasing the type-I error rate.
- Sample size re-estimation based on the true difference can inflate the type-I error rate and the test statistic must be adjusted to control the type-I error rate.