Sample-size versus information-based monitoring

Example: Mean comparison with the O'Brien-Fleming method

Consider the following example:

We would like to compare the mean response $delta = \mu_E - \mu_C$ between two treatment groups with $\alpha = 0.05$ (two-sided) and 90% power.

Suppose further that the common standard deviation is $\sigma = 15$. The required sample size is

$$N \geq \frac{2\sigma^2(z_{1-\alpha/2} + z_{1-\beta})^2}{\delta^2}$$

so that $N = 190$ patients per group.

The O'Brien-Fleming method
Sample-based monitoring

If we add four interim analyses to this study (i.e., $k = 5$) and chose the O'Brien-Fleming method for monitoring, the bounds would be as follows:

```
Symmetric two-sided group sequential design with
90 % power and 2.5 % Type I Error.
Spending computations assume trial stops
if a bound is crossed.

  Analysis  N    Z   Nominal p  Spend
         1  40 4.56     0.0000 0.0000
         2  79 3.23     0.0006 0.0006
         3 118 2.63     0.0042 0.0038
         4 157 2.28     0.0113 0.0083
         5 196 2.04     0.0207 0.0122
     Total                     0.0250

++ alpha spending:
 O'Brien-Fleming boundary.
```

Comments

From the previous output we see that there is a small inflation of the required sample size from $N = 190$ per group to $N' = 196$ per group.

The first interim analysis will occur when $n_1 = 40$ patients per group have been recruited, the second when $n_2 = 79$ patients have been recruited, and so on.

The O'Brien-Fleming method
Information-based monitoring

The same interim analysis can be done considering the *information fraction* of the study, so that the first interim analysis happens at $\tau_1 = 0.20$, the second at $\tau_2 = 0.40$ and so on:

```
Symmetric two-sided group sequential design with
90 % power and 2.5 % Type I Error.
Spending computations assume trial stops
if a bound is crossed.

           Sample
            Size
  Analysis Ratio*  Z   Nominal p  Spend
         1  0.205 4.56    0.0000 0.0000
         2  0.411 3.23    0.0006 0.0006
         3  0.616 2.63    0.0042 0.0038
         4  0.821 2.28    0.0113 0.0083
         5  1.026 2.04    0.0207 0.0122
     Total                       0.0250

++ alpha spending:
 O'Brien-Fleming boundary.
* Sample size ratio compared to fixed design with no interim
```

Comments

This is identical to the previous output but it does not deal with the specific sample size of the particular study.

An added advantage of this approach is that it is general enough for both mean and response-comparison studies (where information is proportional to the sample size) and time-to-event studies (where information is a function of the total number of events).

Inference after monitoring

Inference in study monitoring

By "inference" we are talking about two major areas:

- Determining p-values
- Estimation of treatment effect
- Constructing confidence intervals

Inference is affected by the fact that monitoring of the study has taken place.

The concept of the "p value"
Fixed sample design

The definition of a p-value is the probability under the null hypothesis of observing a test statistic as extreme or more extreme than what was observed.

In a fixed-sample design $|Z_1| < |Z_2|$ implies that Z_2 is more extreme Z_1.

In plain English, this means that the study that resulted in Z_1 produced more extreme evidence than the study that produced Z_2.

The concept of the "p value"
Group sequential design

This is not so clear in the group sequential context.

For example, if $Z_1(\tau_i) > Z_2(\tau_i)$, i.e., the test statistics of two identical studies at the same interim analysis τ_i, it may be clear that $Z_1(\tau_i)$ provides more extreme evidence than $Z_2(\tau_i)$, written formally as

$$(\tau_i, Z_1) \succ (\tau_i, Z_2)$$

However, the following is not as clear: Is $Z(\tau_i) = 3.50$ after stage τ_i which, say, did not result in the interruption of the study, more or less extreme than $Z(\tau_j) = 3.50$ after stage $\tau_j > \tau_i$ which, in this hypothetical experiment, resulted in the interruption of the study?

Governing assumption
Stage-wise ordering

The major assumption that will be made for the following discussion is that the evidence leading to the stopping of the study is *at least as extreme* in stage j as it was in stage $i < j$.

In other words, we assume that a study that was stopped in stage j generated more extreme evidence (larger deviations from the null) than was observed in stage i (where the study was not stopped). This is called "stage-wise ordering" (of the sample space)[1].

It only matters that you reached the kth stopping time. How you reached it is irrelevant.

Stage-wise ordering
Formal definition

If a_i and b_i, where $i = 1, \cdots, k$ are, respectively, futility and efficacy bounds, then $(\tau_i, Z(\tau_i)) \succ (\tau_j, Z(\tau_j))$ if any of the following conditions is satisfied:

1. If $\tau_i = \tau_j$ and $Z(\tau_i) \geq Z(\tau_j)$
2. If $\tau_i < \tau_j$ and $Z(\tau_i) \geq b_i$, i.e., if the study was stopped at an earlier stage
3. If $\tau_i > \tau_j$ and $Z(\tau_i) \leq a_i$, i.e., if the study was not stopped at an earlier stage

Note that this not simply mathematical minutia. Ordering of the sample space aids in determining the level of evidence produced by a study.

Definition of the p-value
One-sided case

With the previous assumption in mind, the p-value in the one-sided case is

$$p = \underbrace{\Pr\left(\cup_{i=1}^{j-1} Z(t_i) > c_i\right)}_{\text{Trial stops at } i<j} + \underbrace{\Pr\left(\cap_{i=1}^{j-1} Z(t_i) \leq c_i, Z(t_j) > z_j\right)}_{\text{Trial did not stop at } i \text{ stops at } j}$$

OR

where z_j is the observed z-score at the final stage.

Example: Two-stage interim analysis design

For example (Proshan, Lan & Wittes, 2006), with $c_1 = c_2 = 2.18$ and $Z(\tau_1) = 2.30$ (i.e., the study did not stop at the first analysis at $\tau_1 = 0.5$ but the null hypothesis was rejected at the second and final analysis at $\tau_2 = 1$).

The p-value is

$$p = \Pr\left(Z(0.5) \geq 2.18 \cup Z(1) \geq 2.30\right) = 0.0218$$

Note. When the study is stopped at the first interim analysis, the p-value is the same as if there had been no monitoring.

Definition of the p-value
Two-sided case

For two-tailed p-values, suppose that the study was stopped at stage τ_j and that the observed z-score was $Z(\tau_j) = z_j$.

If the boundaries are symmetric around zero, then the two-sided p-value is the one-sided p-value applied to $|Z(\tau_i)|$ for $i = 1, \cdots, j$ such that

$$p = \text{Pr}\left(\cup_{i=1}^{j-1} |Z(t_i)| \geq c_i\right) + \text{Pr}\left(\cap_{i=1}^{j-1} |Z(t_i)| < c_i, |Z(\tau_j)| \geq z_j\right)$$

Otherwise, the two-sided p-value is

$$p = 2\min(p_L, p_U)$$

where p_L and p_U are the one-sided p-values of crossing the lower and the upper boundaryrespectively.

Note: In the case where the lower boundaries are simply advisory, it is best to calculate the one-sided p-value p_U.

Confidence intervals
No monitoring

Recall that, without monitoring, we observe $\hat{\delta}$ and $z_{\text{obs}} = \hat{\delta}/\sqrt{\text{var}(\hat{\delta})}$, where $\hat{\delta}$ is $\bar{X}_1 - \bar{X}_2$, $p_C - p_T$ or λ_1/λ_2 in two-sided tests of means, proportions or the hazard ratio in studies of time to an event and z_{obs} is the observed z-score for testing the null hypothesis $H_0 : \delta = 0$.

The statistical test rejects the null hypothesis if

$$p = \Pr\left(|Z| \geq z_{\text{obs}}\right) \leq \alpha$$

Confidence intervals can be constructed by "inverting" this hypothesis test.

Confidence intervals
Definition in terms of p values

Inversion of the hypothesis test means that, using the above probability statement, we calculate lower and upper bounds δ_L and δ_U respectively such that

$$\Pr(\delta_L \geq \delta) \quad = \quad \Pr\left(\frac{\delta_L - \hat{\delta}}{\sqrt{\text{var}(\hat{\delta})}} \geq \frac{\delta - \hat{\delta}}{\sqrt{\text{var}(\hat{\delta})}}\right) = \Pr\left(\frac{\hat{\delta} - \delta_L}{\sqrt{\text{var}(\hat{\delta})}} \geq z_{\text{obs}}\right) \leq \alpha/2$$

and similarly for the upper bound δ_U.

Confidence intervals (cont'd)

This is equivalent to using a z-score with δ_L as the mean of the distribution so that

$$p_{\delta_L} = \mathsf{Pr}(Z > z_{\mathsf{obs}} | \delta = \delta_L) \leq \alpha/2$$

Similarly, the upper bound δ_U is calculated by considering

$$p_{\delta_U} = \mathsf{Pr}(Z < z_{\mathsf{obs}} | \delta = \delta_U) \leq \alpha/2$$

Confidence intervals resulting from stage-wise ordering

Confidence intervals calculated after a study which includes interim monitoring should have the following properties:

1. The confidence interval should be a (contiguous) interval
2. It should agree with the original test. In other words, if the test rejected H_0, then the value of δ under the null should not be contained within the interval.
3. The confidence interval should contain the MLE $\hat{\delta} = Z(t)/\sqrt{\nu_T}$
4. A narrower confidence interval is to be preferred to a wider one

All of these properties hold under the stage-wise ordering espoused in these notes.

Example: Diet trial (Proshan, Lan & Wittes, 2006)

To work through this, consider the following example:

In a clinical trial of 200 participants per arm the primary endpoint was weight change over 3 months. An O'Brien-Fleming spending function was used and four analyses of the data were planned.

With this situation, the O'Brien-Fleming bounds would be

Information time (t)	O-F boundary
0.22	$\pm$4.64
0.55	$\pm$2.81
0.74	$\pm$2.39
1.00	$\pm$2.01

Diet study example
Interim analyses

The first two analyses occurred at times $\tau_1 = 0.22$ and $\tau_2 = 0.55$.

The third analysis occurred after $n_T = 152$ and $n_C = 144$ subjects had been accrued in the treatment and control arms respectively, that is, at (information) time $\tau_3 = 0.74^*$.

*Since the total information is $I_{\max} = 200/2\sigma^2$ and the information at the third interim analysis is $I_3 = \left[\sigma^2 \left(1/152 + 1/144\right)\right]^{-1}$ then, the information fraction is $\tau_3 = \frac{2\sigma^2}{200}/\sigma^2 \left(\frac{1}{152} + \frac{1}{144}\right) \approx 0.74$

Diet study example
Early stopping

The z-score at the third interim analysis was

$$Z(0.74) = \frac{\bar{X}_1 - \bar{X}_2}{\sqrt{(4.8)^2(1/152 + 1/144)}} = 3.76$$

reflecting a sample standard deviation $s = 4.8$ and $\hat{\delta}(\tau_3) = 2.099$.

Diet study example (continued)
P-value

Note our contention before that we don't care how we got to this point and we certainly do not care what might have happened afterwards.

The output from R is as follows:

```
Boundary crossing probabilities and expected sample size assume
any cross stops the trial

Upper boundary (power or Type I Error)
          Analysis
  Theta 1      2 3  Total E{N}
      0 0 0.0025 0 0.0025  0.7

Lower boundary (futility or Type II Error)
          Analysis
  Theta 1      2 3  Total
      0 0 0.0025 0 0.0025
```

Comments

We have $p_U = 0.0025$ and $p_L = 0.0025$ so that the two-sided p-value is

$$p = 2\min(p_L, p_U) = 0.005$$

As $Z(0.74) = 3.76 > 2.39$ the study stops with a cumulative two-sided ("exit") probability=0.005).

Constructing the confidence interval
Effect size

The two-sided confidence interval in terms of the effect size is going to be

$$(\theta_L, \theta_U) = (1.1394, 6.2139)$$

Recall that the effect size is

$$\theta = \frac{\delta}{\sqrt{\text{var}(\delta)}} = \frac{\delta}{\sqrt{2\sigma^2/N}}$$

where N is the sample size for each group at the completion of the study.

Constructing the confidence interval
In terms of the effect of the itnervention

Thus $\delta = \theta\sqrt{2\sigma^2/N}$ and the relevant 95% confidence interval in the scale of δ is

$$\begin{aligned}(\delta_L, \delta_U) &= (\theta_L\sqrt{2\sigma^2/N}, \theta_U\sqrt{2\sigma^2/N}) \\ &= (1.1394\sqrt{2(4.8)^2/200}, 6.2139\sqrt{2(4.8)^2/200}) = (0.544, 2.982)\end{aligned}$$

This means that the experimental treatment reduces weight by between half and three kilograms.

Adaptive designs

Adaptive designs

Adaptive designs refer to designs which adapt various parameters of the initial study design during the study's implementation.

According to the FDA, and *adaptive design clinical study* "is defined as a study that includes a prospectively planned opportunity for modification of one or more specified aspects of the study design and hypotheses based on analysis of data (usually interim data) from subjects in the study."

The FDA Guidance for Industry continues: "Revisions not previously planned and made or proposed after an unblinded interim analysis raise major concerns about study integrity (i.e., potential introduction of bias)."

Thus, the term applies to *a priori* established changes to the study which modify its design.

Introduction

Several assumptions that enter the calculation of the sample size involve ancillary ("nuisance") quantities that are not related to the study question but, if incorrect, may severely impact the study.

Usually, this parameter is the variance σ^2 in analyses involving comparisons of means or the average, or the null proportions $p = \frac{p_C + p_T}{2}$ and p_C in analyses of proportions.

It would useful if the nuisance parameter could be estimated from the data during the study.

Types of adaptive designs

Types of adaptive designs include but are not limited to:

1. Adaptation of study eligibility criteria based on analyses of pretreatment (baseline) data
2. Adaptations to maintain study power based on blinded interim analyses
3. Adaptations based on interim results of an outcome unrelated to efficacy
4. Adaptations using group sequential methods for early study termination for lack of benefit or demonstrated efficacy
5. Adaptive randomization based on relative treatment Group responses
6. Adaptation of sample size based on interim-effect size estimates
7. Adaptation for endpoint selection based on interim estimate of treatment effect

We will concentrate on items 2 and 6 above.

Two-stage designs

First we consider adaptive designs which protect the power of the study.

To protect the power we need to have a good estimate of the variance. It would useful if the variance could be estimated from the data during the study. This leads to the following two-stage design:

- Stage 1: Enroll n_1 subjects per arm, compute additional sample size n_2 per arm based on an interim estimation of the nuisance parameter σ
- Stage 2: Enroll additional n_2 subjects per arm based on the estimate of the variability in the first stage.

Note that the remainder sample size n_2 is now random (i.e., it cannot be predicted *a priori*).

Stein's design
First stage

Stein's two-stage design (Stein, Ann Math Stat, 1945) involves recruiting $2n_1$ subjects and obtaining the pooled sample variance

$$s_T^2 = \frac{(n_1 - 1)s_1^2 + (n_1 - 1)s_2^2}{2(n_1 - 1)}$$

where $s_1^2 = \frac{1}{n_1 - 1} \sum_{i=1}^{n_1} \left(X_{Ti} - \bar{X}_{T_1}\right)^2$ and $\bar{X}_{T_1}$ are the sample variance and mean of the first n_1 observations in the experimental treatment arm collected at the end of the first stage. The sample variance $s_{C_1}^2$ in the control arm is calculated similarly.

Stein's design
Second stage

We use this pooled estimate of the variance to recalculate the sample size by recomputing the maximum required N as follows:

$$N' = \max\left\{n_1; 2s_1^2\left(t_{\alpha/2}; 2(n_1-1) + t_\beta, 2(n_1-1)\right)^2 \Big/ \delta^2\right\}$$

then recruit additional $n_2 = N' - n_1$ patients per arm in the second stage.

Stein's two-stage method
Test statistic

At the end, the appropriate test statistic will be

$$t_S = \frac{\bar{X}_T - \bar{X}_C}{\sqrt{2s_1^2/N}} \sim t_{2(n_1-1)}$$

The beauty of this approach is that the power is $1-\beta$ *regardless of the true variance* σ^2!

Example: Diet and diastolic blood pressure*

Consider a study that compares two diets with respect to diastolic blood pressure (DBP) from baseline to 6 weeks.

The estimated standard deviation is expected to be $\sigma = 5$ mmHg. The sample size needed to detect a difference of $\delta = 3$ mmHg with 85% power is

$$N = \frac{2\sigma^2 \left(z_{1-\alpha/2} + z_{1-\beta}\right)^2}{\delta^2} = \frac{2(5)^2(1.96 + 1.04)^2}{(3)^2} = 50$$

subjects per arm.

*Proshan, *J Biopharm Stat*, 2005

Diet and diastolic blood pressure
Stage 1

Now suppose that after 25 and 27 patients enrolled in the treatment and control arm respectively and $s_1 = 6$ mmHg.

Using Stein's method as mentioned above

$$\begin{aligned} N' &= \max\{52, 2\left(2(6)^2(t_{0.025,50} + t_{0.15,50})^2/(3)^2\right)\}^* \\ &= \max\{52, 2\left(2(6)^2(2.009 + 1.047)^2/(3)^2\right)\} = 150 \text{ total subjects} \end{aligned}$$

*Note here that, since the two groups had different number of patients, we expressed the numbers of patients as total sample sizes and not as sample size per arm.

Diet and diastolic blood pressure
Stage 2

At the end of the study, we observe $\hat{\delta} = 2.40$ with 74 and 78 control and treatment participants respectively, with a pooled estimate of the variance $s_p^2 = 40.1$.

The test statistic is

$$t_S = \frac{\delta}{s_1\sqrt{\frac{1}{n_1} + \frac{1}{n_2}}} = \frac{2.40}{(6)\sqrt{1/74 + 1/78}} = 2.465$$

Comments

- Even though the pooled estimate of the variance is $s_p^2 = 40.1$ the sample variance from the first stage, $s_1^2 = 36$ was used. This is critical, because the t_S statistic above should be compared to a t distribution with (25+27-2)=50 and not (74+78-2)=150 degrees of freedom! The p-value associated with this statistic is thus, $p = 0.017$.
- Had the pooled estimate of the variance been used, the resulting t statistic would have been 2.336. The p-value compared to a t statistic with 150 degrees of freedom would have been 0.021. This is the "naïve t-test" procedure of Wittes & Brittain (*Stat Med*, 1990)*.
- Had we not used Stein's method and performed a study with 50 patients per arm (and, further, assuming that the pooled estimate of the variance were $s_p^2 = 40.10$, and the difference still $\delta = 2.4$ mmHg, the t statistic would have been 1.894 which, compared to a t distribution with $2 \times 50 - 2 = 98$ degrees of freedom produces a p-value 0.061 (i.e., not statistically significant at the 5% level).

* "Naïve" because it does not acknowledge that N' is random.

Concerns with two-stage designs

A big concern about using these designs is the possibility of unblinding the investigator to the treatment effect during the study.

This could happen if one had access to both the "lumped" estimate of the variance

$$s^2_{L1} = \frac{1}{2n_1 - 1} \sum (X_i - \bar{X})^2$$

the variance that would be estimated if treatment and control subject data were lumped together, as well as the pooled estimate of the variance

$$s^2_p = \frac{(n_1 - 1)s^2_1 + (n_1 - 1)s^2_2}{2(n_1 - 1)}$$

where $s^2_i = \frac{1}{n_1 - 1} \sum (X_{i1} - \bar{X}_1)^2$ for $i = 1, 2$.

Concerns with two-stage designs (cont'd)

This is because of the variance decomposition formula that states

$$(2n_1 - 1)s_{L1}^2 = n_1\hat{\delta}_1^2/2 + 2(n_1 - 1)s_1^2$$

(e.g., Proshan, *J Biopharm Stat*, 2005) thus providing an estimate of the interim treatment effect!

A partial solution might be to use the lumped instead of the pooled variance (Gould & Shih, *Comm Stat (A)*, 1992, Gould, *Stat Med*, 1995).

While using the lumped estimate of the variance overestimates the sample size, the inflation is small in most cases (generally in the area of 3%).

Analysis of proportions

Gould (*Stat Med*, 1992) presented the following adaptive procedure for the case of the comparison between two proportions:

- **Stage 1:** Recruit $2n_1$ patients and compute the overall proportion $\hat{p}_1$ of stage-1 patients with an event.
  - Treating $\hat{p}_1 \approx \frac{p_T + p_C}{2}$ and, assuming the originally hypothesized relative risk $R = p_T/p_C$, solve for p_T and p_C using the two equations

$$p = \frac{p_C + p_T}{2} \text{ and } \frac{p_T}{p_C} = R$$

    which results in estimates of the two proportions of

$$\hat{p}_{C1} = \frac{2\hat{p}_1}{1+R} \text{ and } \hat{p}_{T1} = \frac{2R\hat{p}_1}{1+R}$$

  - Plug the two estimates $\hat{p}_{T1}$ and $\hat{p}_{C1}$ in sample size formula for proportions.

Gould's method (continued)

- **Stage 2:** The sample size n_2 for stage 2 will be $\max(n_1, N' - n_1)$, where

$$N' = \frac{\left[z_{1-a/2}\sqrt{2\hat{p}_1(1-\hat{p}_1)} + z_\beta\sqrt{\hat{p}_{T1}(1-\hat{p}_{T1}) + \hat{p}_{C1}(1-\hat{p}_{C1})}\right]^2}{(\hat{p}_{T1} - \hat{p}_{C1})^2}$$

where $\hat{p}_1 = (\hat{p}_{T1} + \hat{p}_{C1})/2$. At the completion of the stage 2 the appropriate test statistic is

$$Z = \frac{\hat{p}_C - \hat{p}_T}{\sqrt{2\hat{p}(1-\hat{p})/N}} \sim N(0,1) \text{ under } H_0$$

where $\hat{p} = (\hat{p}_T + \hat{p}_C)/2$.

Note. This procedure maintains the blind since only the overall event rate is calculated.

Example: Tumor progression*

Consider a study that compares the rate of tumor progression over a three-month period. The standard treatment is expected to have a $p_C = 0.50$ progression rate, and we would like to detect a 30% decline with 90% power.

This means that $p_T = (1 - 0.3)p_C = 0.35$ and $p = (p_T + p_C)/2 = 0.425$. The sample size is (per arm)

$$N = \frac{\left[(1.96)\sqrt{2(0.425)(0.575)} + 1.282\sqrt{(0.5)(0.5) + (0.35)(0.65)}\right]^2}{(0.15)^2} \approx 227$$

*Proshan, *J Biopharm Stat*, 2005

Example: Tumor progression

Stage 1

Now suppose that, after 200 patients have been evaluated, 58 have progressed. Then, the overall event rate is $\hat{p}_1 = 58/200 = 0.290$. Following the procedure as described previously, we get

$$\hat{p}_{C1} = 0.290/0.85 = 0.341 \text{ and } \hat{p}_{T1} = 0.239.$$

Substituting these estimates in the previous equation, produces a revised sample size estimate (per arm)

$$N' = \frac{\left[(1.96)\sqrt{2(0.29)(0.71)} + 1.282\sqrt{(0.341)(0.659) + (0.239)(0.761)}\right]^2}{(0.102)^2} \approx 414$$

The total sample size is 828 subjects, a huge increase over the original estimate of 454 subjects!

Example: Tumor progression
Stage 2: Constant treatment effect

Now suppose that, after the study is done (with all 828 patients accrued) the observed event rate in the control arm is $\hat{p}_C = 0.4$. This is less than 50% as thought at analysis time and implies that it might be this reduced rate in the control that was partially responsible for the 29% interim overall event rate.

If the relative risk is still $R = p_T/p_C = 0.7$ then $\hat{p}_T = 0.28$ and $\hat{p} = 0.34$.

Then the test statistic will be

$$Z = \frac{0.120}{\sqrt{2(0.34)(0.66)/414}} = 3.645$$

This is highly significant (i.e., we overshot the sample size recalculation by quite a bit).

Example: Tumor progression
Stage 2: Constant event rate average

Now if the event rate average is kept constant, with $p_C = 0.40$ we will have and $p_T = 0.26$, then $\hat{p} = 0.29$ as before, but the risk ratio is now $R' = 0.32/0.23 = 0.813$, (i.e., only a 18.7% reduction).

In this case, the z-score will be

$$Z = \frac{0.09}{\sqrt{2(0.29)(0.71)/414}} = 2.854$$

which is associated with a p-value 0.001. In this case, recalculation of the sample size has protected the study.

References

Proschan, M.A., Lan K.K.G. and Wittes J.T. (2006). *Statistical Monitoring of Clinical Trials: A unified approach*. Springer, New York, NY.