Module 5 - Design and Monitoring of Group Sequential Trials
University of Washington - Summer Institute in Biostatistics

Session 2 Take-Home Problems

1. Exploring the fixed sample design: For this problem, use `RCTdesign` to further explore potential changes to the statistical design for the Hodgkin's case study.

   (a) Compare the resulting power curves for the original design with 196 events to the modified design with 121 events. For what alternative (hazard ratio) would each study have 80%, 90%, 95%, and 97.5% power?

   (b) Again consider the original design with 196 events to the modified design with 121 events. Suppose that the true hazard ratio comparing treatment to control were .83. What is the power of each design?

   (c) Suppose the sponsor felt that they may be able to accrue 60 patients uniformly per year. What would be the total number of expected events if accrual lasted 3 years with an additional year of follow-up, assuming median survival (exponentially distributed) in the control arm was 9 months? Compare the relevant operating characteristics of this design to the 196 and 121 event designs. For this design, what would be the resulting 95% confidence interval for the true treatment effect (hazard ratio) if the observed test statistic were exactly equal to critical value for the test?

   (d) Suppose that the true median survival in the control arm were 1 year. How would this effect the expected length of the trial and/or power?

   (e) Assuming patients do accrue uniformly over 3 years at a rate of 60 per year, suppose that we wish to implement an interim analysis at 50% of the maximal information obtained in the study. How far after the start of accrual would this analysis likely take place?

2. Monte Carlo simulation for group sequential study designs: (**Note: This problem does require some familiarity with writing simulations in R, but hints are given throughout.**) A primary focus of our class is the design and implementation of group sequential stopping rules for monitoring a clinical trial. Due to efficiency and ethical (in the case of clinical trials) concerns, it is often advantageous to intermittently test data as it is accumulated in an experiment. The decision as to whether or not the experiment should continue after an interim analysis is formalized via a *stopping rule*. In the general case, a stopping rule is defined for a schedule of analyses occurring at times t_1, t_2, ..., t_J. Often, the analysis times are in turn defined according to the statistical information available at each analysis. Because many statistical models have statistical information proportional to the sample size accrued to the study, such an approach is equivalent to defining the sample sizes N_1, N_2, ..., N_J at which the analyses will be performed. For $j = 1, \ldots, J$, we calculate a specified test statistic T_j based on observations available at time t_j. The outcome space for T_j is then partitioned into stopping set $\mathcal{S}_j$ and continuation set C_j. Starting with $j = 1$, the experiment proceeds by computing T_j, and if $T_j \in \mathcal{S}_j$, the trial is stopped. Otherwise, T_j is in the continuation set C_j, and the trial gathers additional observations until time t_{j+1}. By choosing $C_J = \emptyset$, the empty set, the experiment must stop at or before the J-th analysis.

   Clearly, there are an infinite number of possible stopping rules that one could choose. Noting that each rule carries particular scientific and statistical implications, it is imperative that one consider the operating characteristics of any proposed rule. In this problem, we will use simulation to evaluate the operating characteristics of selected group sequential designs (rules). We will do this in the context of testing the mean from a single group consisting of normally distributed observations and a study design that has a total of 4 analyses. Thus we will assume that $X_i \sim \mathcal{N}(\mu, \sigma^2)$ and we wish to test $H_0 : \mu = \mu_0$ vs. $H_1 : \mu \neq \mu_0 = 0$. Further we will begin by considering a class of stopping rules such that at the j-th analysis, $j = 1, 2, 3, 4$ we will stop the trial if $|T_j| > q_j$ where

$$T_j \equiv \frac{\bar{X}_j}{\sqrt{s_j^2/N_j}},$$

with $\bar{X}_j$ denoting the sample mean of the first N_j observations and s_j^2 denoting the sample variance based upon the first N_j observations. Throughout this exercise, set `set.seed(123456)`.

(a) Simulate 100,000 experiments with $\mu = 0$, $\sigma^2 = 2$, $\alpha = .05$, $N_J = 100$ and analyses occurring after 25, 50, 75, and 100 observations have been accrued. For each of the simulated experiments, calculate $T_j, j = 1, 2, 3, 4$. Store these results in a 4 × 100,000 matrix called `interimStat`.

(b) A natural first choice for a stopping rule is use a standard fixed sample critical value at all analyses. That is, at the j-th analysis we will stop the trial if $|T_j| > z_{1-\alpha/2}$ where z_q denotes the q-th quantile of the standard normal distribution. Investigate the type I error of this proposed stopping rule by using your simulated experiments and completing the table below (*Note: To obtain the q$-th$ quantile of the standard normal distribution in R, type `qnorm( q )`.).

Significant at	Proportion Significant		Number Significant	Proportion Significant
Analysis 1			Exactly 1	
Analysis 2			Exactly 2	
Analysis 3			Exactly 3	
Analysis 4			All 4	
			Any	

(c) Plot and comment on the shape of the density of the observed test statistic after the above stopping rule is applied. To do this, you will need to obtain the statistic that is observed the first time the stopping boundary is crossed for each simulation. This can be done with following code (read it carefully and look at the `ifelse` help file to understand what I'm doing):

```
finalStat <- rep( NA, 100000 )
for( i in 1:4 ){
     finalStat <- ifelse( is.na(finalStat) & (abs(interimStat[i,]) > rep(qnorm(.975),4)),
                          interimStat[i,], finalStat )
}
finalStat <- ifelse( is.na(finalStat), interimStat[4,], finalStat )
```

To plot the estimated density of `finalStat` type:

```
plot( density( finalStat, bw=.2 ) )
```

(*Note: the option `bw` stands for bandwidth and controls the 'smoothness' of the density estimate.)

(d) Consider a generic stopping rule such that at the the j-th analysis we will stop the experiment if $|T_j| > q$ (ie. hold the stopping boundary constant at all analyses). Perform a search to find the value q such that the overall type I error rate is .05 (this is called a level .05 Pocock design). For each proposal in your search, you should estimate the type I error rate based upon 100,000 simulations.

(e) Now consider a generic stopping rule such that at the the j-th analysis we will stop the experiment if $|S_j| > q^*$ where $S_j \equiv \sum_{i=1}^{N_j} X_i$ is the partial sum of the first N_j observations. Perform a search to find the value q^* such that the overall type I error rate is .05 (this is called a level .05 O'Brien-Fleming design). For each proposal in your search, you should estimate the type I error rate based upon 100,000 simulations.

(f) Standardize q^* at each analysis by dividing by $\sqrt{\text{Var}[S_j]}$. Plot $q^*/\sqrt{\text{Var}[S_j]}$ vs. the sample size at each analysis and compare this to a plot of q vs. sample size from part (d).

(g) In the current context, statistical power is defined as $\Pr[\text{Reject}H_0|\mu]$. Based upon 100,000 simulated experiments, estimate the power of the level .05 Pocock and O'Brien-Fleming designs that you found in (d) and (e) at alternatives μ=0.2, 0.4, 0.6, 0.8, and 1. Based upon these 5 points, how do the power curves compare? How do they compare with the power curve of a fixed sample level .05 design where we only perform 1 analysis at 100 observations?

(h) As noted above, efficiency is a primary motivation for the use of group sequential methods. To examine this, estimate the *expected sample size* corresponding to the level .05 Pocock and O'Brien-Fleming designs that you found in (d) and (e) at alternatives μ=0.2, 0.4, 0.6, 0.8, and 0.15. How do they compare (discuss this in relation to your result from (g)) *Hint: You'll need to find the analysis at which experiment stopped. It might be easiest to adapt the code in part (c) to do this.