Getting Started with RCTdesign

Scott S. Emerson, M.D., Ph.D.

RCTdesign.org

August 18, 2012

Abstract

In this tutorial, we demonstrate the general approach to clinical trial design using RCTdesign (or S+SeqTrial). For convenience, we use a hypothetical trial of an experimental treatment for hypertension (high blood pressure), and we imagine that it will be tested in a randomized clinical trial (RCT) against placebo. The major emphasis is the interpretation of the most commonly used output.

1 Introduction

1.1 Clinical Setting

We consider a hypothetical clinical trial of a new treatment for hypertension. In terms of the elements of a clinical indication, we assume

- *Disease*: Hypertension as defined by systolic blood pressure (SBP) greater than 150 mm Hg on three occasions a week apart.
- *Patient population*: Adults over 18 years, nonpregnant, no liver disease, not previously treated for hypertension.
- *Treatment*: Daily dosing with experimental therapy.
- *Outcome*: Decrease in systolic blood pressure relative to what might have been obtained in the absence of any treatment.

1.2 Clinical Trial Setting

We consider

- *RCT population* Patient volunteers seen at participating clinics with hypertension (SBP $>$ 150 mm Hg as defined for the indication) for whom medical treatment is being considered for the first time and who meet well-defined eligibility criteria.
- *Comparator treatments* Patients will receive either the experimental treatment or a placebo treatment for a duration of 6 months.
- *Assignment to treatment* Patients will be randomized in a 1:1 ratio to either the experimental treatment or placebo in a double blind fashion.

- *Clinical outcome* Systolic blood pressure measured 6 months after randomization.
- *Statistical summary measure* Difference in mean systolic blood pressure between the study arms after 6 months of treatment. The measure of treatment effect will be the mean SBP for patients randomized to the experimental treatment minus the mean SBP for patients randomized to receive the placebo.

1.3 Probability Model

In the most general case considered here, we use a distribution-free probability model in which we will compare mean SBP using a t test that allows for unequal variances. For a more detailed discussion of the theory and notation behind this probability model as it is implemented in RCTdesign, see the tutorial *Mean Probability Model*. Some general notation is provided below, however, as an aid for discussing program inputs.

- Y_{kit} represents the systolic blood pressure for the ith patient $i = 1, \ldots, n$ on study arm k ($k = 0$ for placebo, $k = 1$ for the experimental treatment) at time t ($t = 0$ at randomization, $t = 6$ at end of study.
- We presume $Y_{kit} \sim (\mu_{kt}, \sigma^2_{kt})$, signifying that the mean SBP of the patients treated on study arm k is μ_{kt} at time t, and the variance of the SBP for patients treated on study arm k is σ^2_{kt} at time t.
- We presume that all patients are independent of each other, and that patients receiving the experimental treatment are different than the patients receiving placebo.
- We presume that measurements made at different times on the same patient on study arm k have correlation ρ_k. Hence, $corr(Y_{ki0}, Y_{ki6}) = \rho_k$.
- By randomization, we know that measurements made at the time of randomization have the same distribution on both study arms. Hence, $\mu_{00} = \mu_{10}$ and $\sigma^2_{00} = \sigma^2_{10}$.
- Apart from the above assumptions, we do not otherwise characterize the distribution of Y_{kit}. In particular, we make no assumption that the distributions of Y_{ki0} and Y_{ki6} have the same shape, nor do we assume that the distributions of Y_{0i6} and Y_{1i6} have the same shape.

Now in the subsection on the Clinical Setting, we characterized the desired outcome for the treatment indication as a lower SBP than might have been obtained in the absence of any treatment. This would suggest on focusing on the distributions of constructed differences $D_{ki} \equiv Y_{ki6} - Y_{ki0}$. From standard theory of expectations, we find $D_{ki} \sim (\omega_k, \tau^2_k)$, where

$$\omega_k \equiv E[D_{ki}] = E[Y_{ki6} - Y_{ki0}] = E[Y_{ki6}] - E[Y_{ki0}] = \mu_{k6} - \mu_{k0}$$
$$\tau^2_k \equiv Var[D_{ki}] = Var[Y_{ki6} - Y_{ki0}] = Var[Y_{ki6}] + Var[Y_{ki0}] - 2Cov[Y_{ki6}, Y_{ki0}] = \sigma^2_{k6} + \sigma^2_{k0} - \rho_k \sigma_{k6} \sigma_{k0}.$$

The measure of treatment effect would logically then be defined as $\theta = \omega_1 - \omega_0$. However, by considering the definition of ω_k, we also note that the measure of treatment effect could be defined as $\theta = (\mu_{16} - \mu_{10}) - (\mu_{06} - \mu_{00}) = \mu_{16} - \mu_{06}$. So because of randomization, we can define the treatment effect as the difference in mean change in SBP across study arms (as suggested by our description of the Clinical Setting) or by the difference in mean final SBP across study arms (as suggested by our description of the Clinical Trial Setting). (In a later section we will discuss how the analysis of covariance (ANCOVA) model could be used with yet another formulation of treatment effect θ.)

Note: *In RCTdesign, the treatment effect θ is always called* `theta`. *In two sample comparisons, it will represent either a difference computed as some summary of the response distribution (e.g., mean, proportion)*

on the treatment arm minus the summary of response on the control arm or a ratio computed as a summary of the response distribution (e.g., odds, hazard) on the treatment arm divided by the summary of the response on the control arm.

1.4 Statistical Hypotheses

We are interested in testing the null hypothesis

$$H_0 : \theta = 0$$

against the one-sided alternative of a lesser hypothesis

$$H_0 : \theta < 0.$$

2 Design Parameters for Sample Size Computation

For purposes of study design, we must define hypothesis parameters:

- Null hypothesis: We hypothesize that if the treatment does not provide benefit, the average treatment effect will be $\theta_0 = 0$ mm Hg.
- Alternative hypothesize: We hypothesize that the treatment will result in a treatment effect of $\theta_1 = -10$ mm Hg.

We also must estimate the "nuisance" distributional parameters:

- the standard deviation of the 6 month change in SBP measurements on the treatment and control arms under the null hypothesis is estimated to be 30 mm Hg.
- we estimate that the the standard deviation of the change in SBP measurements is unaffected by the treatment, and thus estimate that the standard deviation on the treatment and control arms under the alternative hypothesis to be 30 mm Hg.

The design parameters related to the statistical criteria for evidence are:

- We desire to perform a one-sided hypothesis test of a lesser alternative (i.e., the magnitude of the treatment effect under the alternative hypothesis is numerically less than the treatment effect under the null hypothesis).
- We desire to have a type I statistical error of 0.025.
- We desire to have 95% statistical power to reject the null hypothesis when the design alternative hypothesis is true.

3 RCTdesign Code for Specifying the Clinical Trial Design

3.1 Preliminaries

Before we can use the R functions contained in the RCTdesign package, we must declare the use of that library. This is effected by the following R code:

```
> library(RCTdesign)
```

We note that when the RCTdesign package is loaded, we also load the `lattice` and `survival` R packages in order to be able to handle plotting and censored time to event data analysis, respectively.

Each RCTdesign function carries a version date. You can obtain printout of the version date for every function by executing `RCTversion()`. In that printout, the functions will be grouped in general categories according to their actions: "Utilities", "Model", "Scale", "Boundary", "Design", "Distribution", "OperatingChar", "Inference", "PHNSubjects", or "Monitor". You can obtain version dates for selected categories by supplying the category name to `RCTversion()`.

```
> RCTversion("Design")
```

```
$Design
                                Version
validSeqOptions                 20120731
all.equal.seqDesign             20120725
is.seqDesign                    20120725
as.seqDesign                    20120725
is.seqFixDesign                 20120725
is.seqHypothesis                20120725
seqDesign                       20120731
seqDesignKArms                  20120725
seqDesignCtoS                   20120725
update.seqDesign                20120725
seqFixDesign                    20120725
seqFixDesign.seqDesign          20120725
seqFixDesign.seqOperatingChar   20120725
seqFixDesign.seqOC              20120725
print.seqDesign                 20120725
print.seqHypothesis             20120725
print.seqParameters             20120725
plot.seqDesign                  20120725
```

Alternatively, you can obtain the version date for a specific function by calling the function with argument `version=TRUE`

```
> seqDesign(version=TRUE)
```

```
[1] "20120731"
```

3.2 Using `seqDesign()`

In RCTdesign and SeqTrial, clinical trial designs are created using the function `seqDesign()`. As there are many different types of clinical trial designs, there are also many different arguments that can be supplied to `seqDesign()`. The typical user will not use all of them. In the following, we present what are probably the most widely used arguments, noting, however, that the defaults are often adequate to handle the common settings.

The following R code will produce a `"seqDesign"` object named `dsnN` (where we append the "N" to help us remember that we asked seqDesign() to calculate the sample size).

```
> dsnN <- seqDesign(
+         prob.model= "mean",
+         arms= 2,
+         ratio= c(1,1),
+         null.hypothesis= 0,
+         alt.hypothesis= -10,
+         sd= c(30,30),
+         test.type= "less",
+         size= 0.025,
+         power= 0.90
+         )
```

We can obtain an abbreviated print out of the design by just typing the object name.

```
> dsnN

Call:
seqDesign(prob.model = "mean", arms = 2, null.hypothesis = 0,
    alt.hypothesis = -10, sd = c(30, 30), ratio = c(1, 1), test.type = "less",
    size = 0.025, power = 0.9)

PROBABILITY MODEL and HYPOTHESES:
   Theta is difference in means (Treatment - Comparison)
   One-sided hypothesis test of a lesser alternative:
           Null hypothesis : Theta >=   0    (size  = 0.025)
    Alternative hypothesis : Theta <= -10    (power = 0.900)
   (Fixed sample test)

STOPPING BOUNDARIES: Sample Mean scale
                       Efficacy Futility
    Time 1 (N= 378.27)  -6.0464  -6.0464
```

It is useful to belabor the elements of this printout in this simple case, just to draw attention to the key parts that will always displayed for every design.

- The R code that was used to generate the design is printed first. In this way you can be reminded of the key design parameters that describe the clinical and clinical trial setting. This call information is stored in the `"seqDesign"` object `dsnN`, and it can be retrieved directly as `dsnN$call`. (It should be noted that the information may not be exactly what was typed into R: abbreviated argument names will be replaced by the full name, and if the design was created by an `update()` command, the value of `$call` will reflect the entire updated call– see the later discussion of `update()`.)
- The printout then indicates that the probability model corresponds to a treatment effect `theta` that reflects the difference of means in a two arm study. It also reminds you that the difference is computed as the treatment arm mean minus the control arm mean. (Had we asked for a one arm study, `theta` would have been described as the mean response. And in other probability models, `theta` might correspond to geometric mean(s), binomial proportion(s), odds (or odds ratios), rates, or hazard ratios.)
- Next the printout indicates that the design is intended to test a one-sided hypothesis of a lesser alternative. (Other possibilities include a one-sided test of a greater alternative, a two-sided test, or some advanced designs that might correspond to a hybrid of, for instance, a superiority test and a noninferiority test.)

- The printout indicates the null hypothesis (and the test's type I error) and the design alternative hypothesis (and the test's power to detect that alternative).
- The printout then indicates that the design is a fixed sample test (i.e., it does not involve sequential testing and interim analyses). (When a sequential stopping boundary is being used, the printout will attempt to provide references to any manuscripts that first described the particular stopping boundary.)
- The printout provides an array of "stopping boundaries", that in this case are just the critical values for statistical analysis. The printout indicates that these critical values will be displayed on the "sample mean scale" by default. (The "stopping boundary" terminology is truly more appropriate if we were conducting a sequential clinical trial. That setting will be described in more detail later.)
- The critical values are displayed for each time that the data might be analyzed. In this example, we did not specify multiple analysis times, so there are only critical values specified for `Time 1`. The printout indicates that the number of observations that should be available at that analysis is 378.27, if we want to have 90% power to detect a treatment that causes a 10 mm Hg decrease in average blood pressure using a type I error of 0.025. The sample size reported is the total sample size, so we would anticipate having approximately 189 subjects on each study arm. (Note that it is of course absurd in this setting to have a fractional person. But because there are some advanced settings that the "sample size" is really corresponding to "statistical information", RCTdesign finds it convenient to not round the sample size at this stage. RCTdesign routines that simulate data will round these sample sizes as appropriate. As demonstrated later, the user may also specify an integer sample size.)
- The printout displays the critical values on the "sample mean scale" by default. RCTdesign is very much organized around giving information useful to a scientist or a clinician. While you could have requested it to tell you that the critical value for a Z statistic would be 1.96, or that you would reject if the p value were less than 0.025, RCTdesign encourages you to consider the estimated treatment effect that would be the threshold for statistical significance. In this case it indicates that statistical significance (at level 0.025) would be declared for results more extreme than an observed difference in final SBP corresponding to a mean SBP that was 6.046 mm Hg less on the treatment arm than on the control arm. Note that the lefthand critical value is labeled "Efficacy", telling us that estimated treatment effects lower than the critical value would be considered statistically significant evidence that the treatment was efficacious. On the other hand, the "Futility" label on the righthand column indicates that observations higher than the critical value would correspond to a failure to reject the null hypothesis.

Before proceeding with computation of more detailed operating characteristics of the design, it is useful to point out that it is possible to obtain this very same design with far less typing: many of the choices for this design are default values:

- If `prob.model` is not specified, the `"mean"` probability model is presumed.
- If `arms` is not specified, a two arm study is presumed.
- if `ratio` is not specified, 1:1 randomization is presumed for two arm studies.
- If `null.hypothesis` is not specified, a null hypothesis of $\theta_0 = 0$ is presumed for a `"mean"` probability model.
- If `test.type` is not specified, a one-sided hypothesis test of a less hypothesis is presumed when `alt.hypothesis` is less than `null.hypothesis`. (Otherwise, a one-sided test of a greater hypothesis would have been presumed.)
- If `size` is not specified, a level 0.025 test is presumed for a one-sided test. (For a two-sided test, a level 0.05 test would have been presumed.)

Furthermore, there is no need to always specify more than one value for some of the arguments:

- If only a single value is specified for `sd`, it is presumed that that same value applies to both the treatment and control arms.
- If only a single value is specified for `ratio`, it is presumed that the second value is 1.

When specifying the argument labels, you need only type enough letters to uniquely identify the intended value. Hence, `alt.hypothesis=-10` could have been abbreviated to `alt=-10`. (Before indiscriminately using abbreviated argument names, however, it is highly advisable that a user become familiar with all the arguments that might be supplied to `seqDesign()`, because there are many of them.)

Lastly, when specifying a value for `prob.model`, if that value is listed first, `seqDesign()` will presume that the first value corresponds to `prob.model`. (It is actually a little more complicated than this, and we recommend that only `prob.model` be specified without an argument name, and that it always be specified first.)

In light of the above, the exact same design could be specified as

```
> dsnA <- seqDesign(
+         alt.hypothesis= -10,
+         sd= 30,
+         power= 0.90
+         )
> dsnA

Call:
seqDesign(alt.hypothesis = -10, sd = 30, power = 0.9)

PROBABILITY MODEL and HYPOTHESES:
   Theta is difference in means (Treatment - Comparison)
   One-sided hypothesis test of a lesser alternative:
           Null hypothesis : Theta >=   0    (size  = 0.025)
    Alternative hypothesis : Theta <= -10    (power = 0.900)
   (Fixed sample test)

STOPPING BOUNDARIES: Sample Mean scale
                       Efficacy Futility
    Time 1 (N= 378.27)  -6.0464  -6.0464

> all.equal(dsnN,dsnA)

[1] TRUE
```

3.3 Computing Designs when Sample Size is Constrained

In the previous example, we specified a design alternative of $\theta_1 = -10$ mm Hg, and a desired statistical power of 90%. In computing the clinical trial design, `seqDesign()` returned an estimate for the sample size that should be accrued to the study.

Often, however, there are constraints on the sample size: either the sample size that is computed is too large for the available resources, or a larger sample size is needed to meet regulatory agency requirements for safety data. In those cases, we can specify a sample size and ask `seqDesign()` to either

- compute the alternative for which the clinical trial design has the desired power, or
- compute the power that the design provides to detect some specified alternative.

For instance, suppose we were still most interested in the alternative of $\theta_1 = -10$ mm Hg, but that we could only afford to accrue $N = 300$ subjects. We might then execute the following code to ask `seqDesign()` to calculate the power instead of the sample size. (Note that had we not specified `power="calculate"`, `seqDesign()` would have ignored the `sample.size` argument. This is because `power` has a default argument.).

```
> dsnPwr <- seqDesign(
+         sample.size= 300,
+         alt.hypothesis= -10,
+         sd= 30,
+         power= "calculate"
+         )
> dsnPwr

Call:
seqDesign(alt.hypothesis = -10, sd = 30, sample.size = 300, power = "calculate")

PROBABILITY MODEL and HYPOTHESES:
   Theta is difference in means (Treatment - Comparison)
   One-sided hypothesis test of a lesser alternative:
           Null hypothesis : Theta >=   0    (size  = 0.025)
    Alternative hypothesis : Theta <= -10    (power = 0.823)
   (Fixed sample test)

STOPPING BOUNDARIES: Sample Mean scale
                    Efficacy Futility
    Time 1 (N= 300)  -6.7895  -6.7895
```

Alternatively, we might want to continue to focus on having 90% power, and just want to find out the alternative for which 300 subjects would provide that power. Because we did not supply an alternative hypothesis, we will need to tell `seqDesign()` that we are interested in testing a lesser alternative hypothesis, or it will by default assume a one-sided test of a greater alternative. (Note that if we had specified an alternative hypothesis, `seqDesign()` would have ignored the `sample.size` argument, because the design was overspecified– power and sample size computations use any two of sample size, power, and alternative hypothesis to compute the third.)

```
> dsnAlt <- seqDesign(
+         sample.size= 300,
+         test.type= "less",
+         sd= 30,
+         power= 0.9
+         )
> dsnAlt

Call:
seqDesign(sd = 30, sample.size = 300, test.type = "less", power = 0.9)

PROBABILITY MODEL and HYPOTHESES:
```

```
   Theta is difference in means (Treatment - Comparison)
   One-sided hypothesis test of a lesser alternative:
           Null hypothesis : Theta >=   0.00    (size  = 0.025)
    Alternative hypothesis : Theta <= -11.23    (power = 0.900)
   (Fixed sample test)

STOPPING BOUNDARIES: Sample Mean scale
                     Efficacy Futility
    Time 1 (N= 300)  -6.7895  -6.7895
```

In fact, the two designs `dsnPwr` and `dsnAlt` are truly identical in the sense that their critical values and power curves are the same. The only difference in this case is the printout. To see this, we can use the function `seqOC()` to print the alternative having a specified power, or to print the power for a specified alternative. In the following code, you will see that the `dsnAlt` and `dsnPwr` have the same power to detect $\theta_1 = -10$ and have 90% to detect the same alternative.

```
> seqOC(dsnAlt, theta= -10)

### Asymptotic Operating Characteristics
Operating characteristics at theta= -10
ASN= 300
Expected theta= -10
Lower Power= 0.823

Stopping Probabilities:
                Lower Null Upper Total
Analysis time 1 0.823    0 0.177     1

> seqOC(dsnPwr, power= 0.9)

### Asymptotic Operating Characteristics
Operating characteristics at theta= -11.2289
ASN= 300
Expected theta= -11.2289
Lower Power= 0.9

Stopping Probabilities:
                Lower Null Upper Total
Analysis time 1   0.9    0   0.1     1
```

3.4 Modifying a Clinical Trial Design

Design of a randomized clinical trial in inevitably an iterative process in which an initial candidate design is computed, and its operating characteristics are discussed among the collaborators. Based on those discussions, the design is modified in some way. The process is repeated until the best balance is found for the often competing concerns of the collaborating basic scientists, clinical scientists, statisticians, practicing clinicians, ethicists, financial directors, study coordinators, and regulators.

Computing such modifications is made easy in RCTdesign by using the `update()` function.

To illustrate this use, we consider the possibility that the clinical trial collaborators might want to better address study efficiency and ethics through the use of a group sequential stopping rule. We imagine the

collaborators consider conducting the RCT with a maximum of either 2 or 4 equally spaced analyses (1 or 3 interim analyses, and a final analysis). Furthermore, the collaborators adopt a stopping boundary that corresponds to a one-sided symmetric boundary with O'Brien-Fleming boundary relationships (see Emerson and Fleming, *Biometrics*, 1989).

Because equally spaced analyses are the default schedule of analyses in `seqDesign()` and O'Brien-Fleming boundary relationships are the default for stopping boundaries, we can easily specify two additional designs for this clinical trial setting: using the following code everything previously specified will stay the same, except there will be either 2 or 4 analyses (the default value for number of analyses is to use a "fixed sample design", i.e., a design with no interim analyses). In order to help us remember what each design represents, we choose to name the designs by indicating that maximum number of analyses is $J = 2$ or $J = 4$.

```
> dsnJ2 <- update(dsnN,
+         nbr.analyses= 2
+         )
> dsnJ2

Call:
seqDesign(prob.model = "mean", arms = 2, null.hypothesis = 0,
    alt.hypothesis = -10, sd = c(30, 30), ratio = c(1, 1), nbr.analyses = 2,
    test.type = "less", size = 0.025, power = 0.9)

PROBABILITY MODEL and HYPOTHESES:
   Theta is difference in means (Treatment - Comparison)
   One-sided hypothesis test of a lesser alternative:
           Null hypothesis : Theta >=   0    (size  = 0.025)
    Alternative hypothesis : Theta <= -10    (power = 0.900)
   (Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                       Efficacy Futility
    Time 1 (N= 191.44) -12.0972   0.0000
    Time 2 (N= 382.89)  -6.0486  -6.0486

> dsnJ4 <- update(dsnN,
+         nbr.analyses= 4
+         )
> dsnJ4

Call:
seqDesign(prob.model = "mean", arms = 2, null.hypothesis = 0,
    alt.hypothesis = -10, sd = c(30, 30), ratio = c(1, 1), nbr.analyses = 4,
    test.type = "less", size = 0.025, power = 0.9)

PROBABILITY MODEL and HYPOTHESES:
   Theta is difference in means (Treatment - Comparison)
   One-sided hypothesis test of a lesser alternative:
           Null hypothesis : Theta >=   0    (size  = 0.025)
    Alternative hypothesis : Theta <= -10    (power = 0.900)
   (Emerson & Fleming (1989) symmetric test)
```

```
STOPPING BOUNDARIES: Sample Mean scale
                       Efficacy Futility
    Time 1 (N=  98.65) -24.2030  12.1015
    Time 2 (N= 197.29) -12.1015   0.0000
    Time 3 (N= 295.94)  -8.0677  -4.0338
    Time 4 (N= 394.59)  -6.0508  -6.0508
```

Note that in order to maintain the specified power to detect the design alternative, the maximal sample size increases with additional analyses. (Later we shall see, however, that the average sample size requirements decrease with additional analyses.)

4 RCTdesign Code for Evaluation of a Design

As noted above, clinical trial design is typically a careful process involving detailed comparisons of operating characteristics for multiple candidate designs. S+SeqTrial was originally developed because we found that the commercially available software at the time did not facilitate such comparisons (see Emerson, *The American Statistician*, 1996). We continue to regard that the major purpose of SeqTrial and RCTdesign is to facilitate such evaluation and comparisons. To that end, we include a broad spectrum of group sequential design families and provide a full complement of evaluative techniques and functions.

Randomized clinical trials represent experimentation in human volunteers. Hence, there is much more to a clinical trial design than the sample size. In our two *Statistics in Medicine* Tutorials on frequentist and Bayesian evaluation of group sequential designs (see Emerson, Kittelson, and Gillen, *Statistics in Medicine*, 2007) we discuss the evaluation of operating characteristics that include

- The power function describing the statistical power to reject the null hypothesis over the full range of plausible alternatives, including the type I error (power at the null hypothesis) and the power at selected alternatives that might have been used to compute the maximal sample size or are otherwise of particular clinical interest.
- The alternatives that can be rejected with high statistical power, including the alternative that is rejected with statistical power of 97.5%, because that is the alternative that will be perfectly discriminated from the null hypothesis by a 95% confidence interval.
- The worst case sample size that will need to be accrued if the study continues to the final analysis.
- The expected sample size (average sample N or ASN) that would be accrued as a function of the various hypothesized treatment effects.
- The probability of stopping at each analysis as a function of the various hypothesized treatment effects, both overall and according to whether the null hypothesis would or would not be rejected.
- The stopping boundaries at each analysis, especially when expressed on the relevant scale that corresponds to the estimated treatment effect, but optionally also on scales corresponding to a Z statistic and/or a fixed sample P value.
- The statistical inference possible at each of the analyses expressed as adjusted point estimates, confidence intervals, and p values (where the adjustment accounts for the use of a sequential sampling plan).
- The Bayesian statistical inference that would be made at each stopping boundary under a spectrum of assumed prior distributions.

- The Bayesian predictive probability or freqentist conditional probability of a trial continuing to the final analysis and attaining a statistically significant result (however, see Emerson, Kittelson, and Gillen, *BEPress*, 2007 for a discussion of the difficulties with interpreting such "stochastic curtailment" measures).

4.1 Using `seqEvaluate()`

For a single design, most of the above operating characteristics can be obtained using `seqEvaluate()`.

```
> dsnJ4.eval <- seqEvaluate(dsnJ4)
> names(dsnJ4.eval)
```

```
[1] "bndTable"       "pwrTable"       "altTable"       "inferenceTable"
[5] "design"         "designName"
```

If we now just type the name of the `"seqEvaluate"` object, it would print several tables and several plots. For the purpose of this tutorial, we will print the parts of the evaluation separately in order to facilitate discussion of their contents.

A `"seqEvaluate"` object contains an element `$bndTable` that provides the stopping boundaries at each analysis on several display scales (these boundaries are computed when `seqEvaluate()` calls `changeSeqScale()`).

```
> dsnJ4.eval$bndTable
```

```
    Anlys  SampSize      CrudeEst             Z         FxdP      Hnoninf
Eff     1  98.64708 -2.420304e+01 -4.006459e+00 3.081788e-05 0.9997394756
Eff     2 197.29416 -1.210152e+01 -2.832994e+00 2.305709e-03 0.9774236753
Eff     3 295.94125 -8.067679e+00 -2.313130e+00 1.035774e-02 0.8762749666
Eff     4 394.58833 -6.050759e+00 -2.003230e+00 2.257632e-02           NA
Fut     1  98.64708  1.210152e+01  2.003230e+00 9.774237e-01 0.0002605244
Fut     2 197.29416  4.733810e-10  1.108196e-10 5.000000e-01 0.0225763247
Fut     3 295.94125 -4.033839e+00 -1.156565e+00 1.237250e-01 0.1237250334
Fut     4 394.58833 -6.050759e+00 -2.003230e+00 2.257632e-02           NA
```

The columns included by default include

- A label indicating whether the row corresponds to an efficacy ("Eff") or a futility ("Fut") boundary.
- "Anlys"= The analysis index.
- "SampSize"= The total sample size contributing data to the analysis.
- "CrudeEst"= The crude (not adjusted for the bias introduced by sequential sampling) estimate of treatment effect that corresponds to a threshold for stopping.
- "Z"= The Z statistic corresponding to the threshold for stopping.
- "FxdP"= The fixed sample p value that corresponds to the Z statistic (it should be noted that this is not a true p value for a sequential sampling plan).

- "Hnoninf"= The Bayesian predictive power based on a noninformative (improper) prior for the treatment effect (this estimates the probability of achieving a statistically significant result at the final analysis assuming that all possible alternative were equally likely *a priori*).
- (Not shown: you can optionally specify additional boundary scales to be printed, see the help file for `seqEvaluate()`)

A `"seqEvaluate"` object contains an element `$pwrTable` that provides the alternative hypotheses corresponding to specified power levels, as well as the average sample size (ASN) and cumulative stopping probabilities for each analysis (these tables are computed when `seqEvaluate()` calls `seqOC()`).

```
> dsnJ4.eval$pwrTable
```

```
 Power    TrueEff AvgSampSiz CumStpPrb 1 CumStpPrb 2 CumStpPrb 3 CumStpPrb 4
 0.975 -12.101518   255.0144 0.022607140   0.5025733   0.8897007           1
 0.950 -11.123843   269.8266 0.015251553   0.4143725   0.8351041           1
 0.900 -10.000000   286.9618 0.009485676   0.3212784   0.7602621           1
 0.800  -8.642483   305.4480 0.005297619   0.2308532   0.6674781           1
```

(In creating this particular `"seqEvaluate"` object, we did not request a `$altTable` providing similar output for specified alternatives, but we could have.)

A `"seqEvaluate"` object contains an element `$inferenceTable` that provides the statistical inference that would be made if the study stopped with results corresponding exactly to the stopping thresholds (these tables are computed when `seqEvaluate()` calls `seqInference()`.

```
> dsnJ4.eval$inferenceTable
```

```
    Anlys  SampSize         BAM      CIlo.m        CIhi.m       Pval.m
Eff     1  98.64708 -22.9997622 -31.7432188 -1.225737e+01 3.082106e-05
Eff     2 197.29416 -11.2581350 -18.3532751 -3.541377e+00 2.414848e-03
Eff     3 295.94125  -7.6835655 -13.5446916 -9.675509e-01 1.233929e-02
Eff     4 394.58833  -6.0507591 -12.1015182 -1.653907e-09 2.500000e-02
Fut     1  98.64708  10.8982440   0.1558498  1.964170e+01 9.765257e-01
Fut     2 197.29416  -0.8433832  -8.5601416  6.251757e+00 4.011230e-01
Fut     3 295.94125  -4.4179527 -11.1339673  1.443173e+00 6.715299e-02
Fut     4 394.58833  -6.0507591 -12.1015182 -1.653907e-09 2.500000e-02
```

The columns included by default include

- A label indicating whether the row corresponds to an efficacy ("Eff") or a futility ("Fut") boundary.
- "Anlys"= The analysis index.
- "SampSize"= The total sample size contributing data to the analysis.
- "BAM"= The "bias adjusted mean" point estimate that would be reported for stopping exactly at the boundary.
- "CIlo.m", "CIhi.m"= The confidence interval bounds for the true treatment effect, where the bounds are calculated using the MLE ordering (alternatively or in addition, we could have used the stagewise (or analysis time) ordering or the likelihood ratio ordering).

- "Pval.m"= The true p value adjusted for the sequential sampling plan as calculated under the MLE (or mean) ordering (again, stagewise or likelihood ration orderings could have been requested).

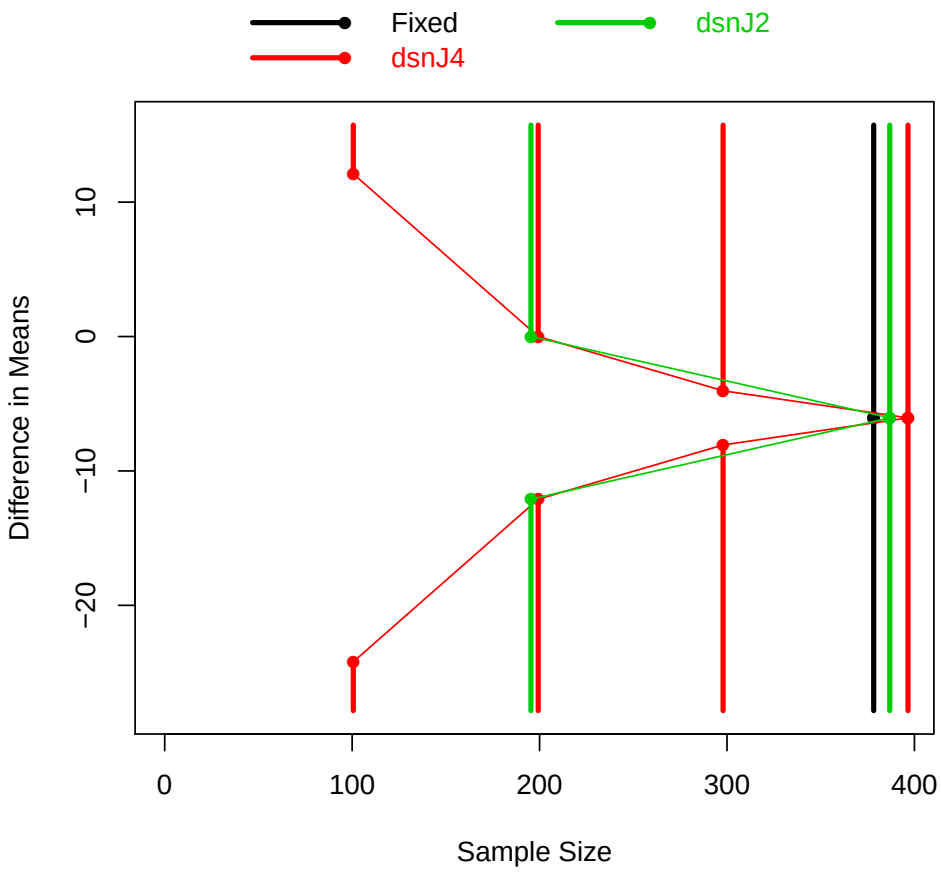
When printing a `"seqEvaluate"` object, plots of the stopping boundaries, the statistical inference, the power curve, the ASN curve, and the stopping probabilities are displayed by default. Rather than show these plots for a single design, we illustrate the use of those functions when comparing one or more designs.

4.2 Comparing Designs Using Plots

We can use `plot()` (which just calls `seqPlotBoundary()`) to overlay several stopping boundaries.

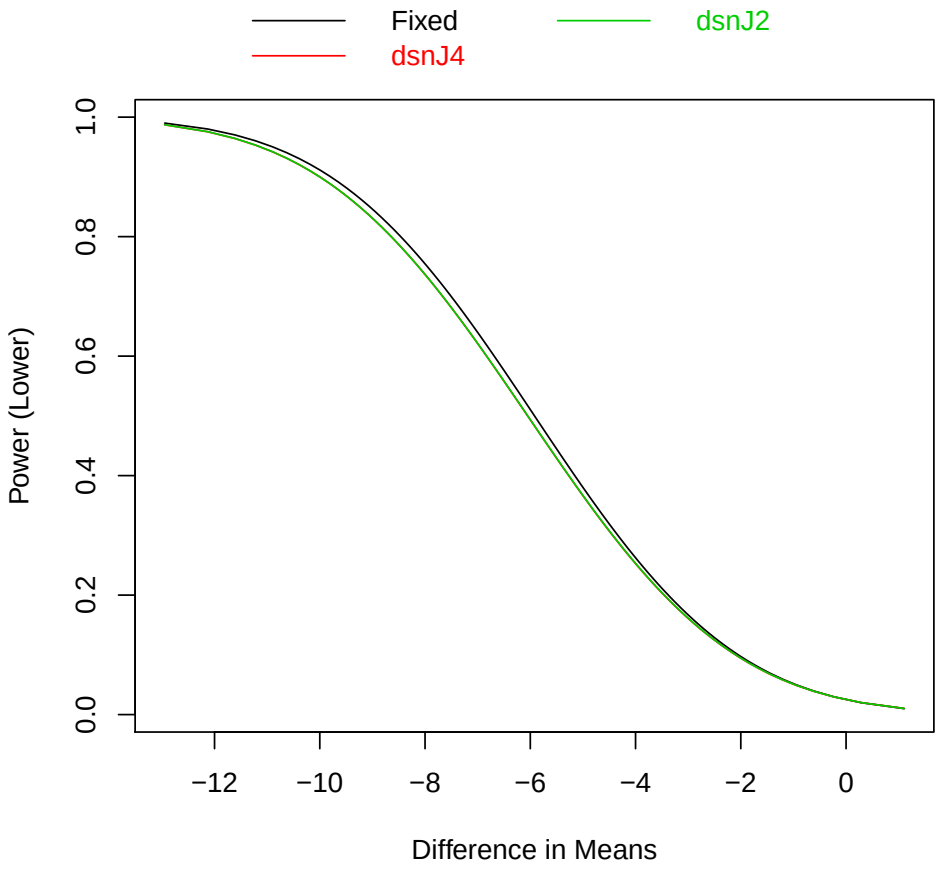
- By default, a fixed sample design having the same power as the first listed design's design alternative is also plotted.
- The x-axis of the boundary plots is the sample size, thereby allowing you to assess any increased sample size requirements across designs.
- The y-axis is the stopping threshold on some scale. By default, RCTdesign encourages the use of the MLE or `"X"` scale, rather than, say, the `"Z"` scale used by some other programs. The reasoning is that considerations about study ethics relate more to the magnitude of a treatment effect that would be associated with stopping (though statistical credibility also matters). Increases in sample size will not generally affect the boundary on the `"Z"` scale, though there would be a large difference in the magnitude of the boundary on the `"X"` scale. (See tutorials on *Scientific versus Statistical Scales*.)
- The true stopping regions are represented by the vertical lines. Test statistics that are observed on one of those vertical lines would correspond to a recommendation to stop. The horizontal lines are just to aid the visualization of the stopping boundary relationships across analyses.

```
> plot(dsnJ4, dsnJ2)
```

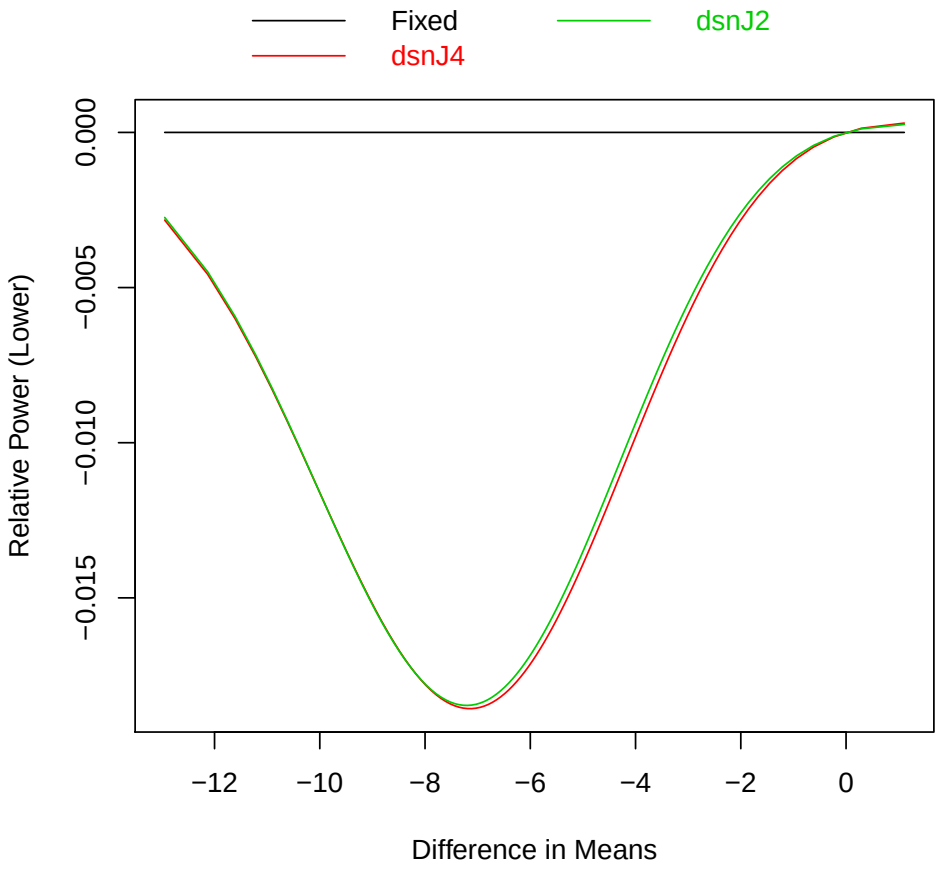


We can use `seqPlotPower()` to compare power curves. The choice `reference=TRUE` or `reference=`*someDesign* will plot the change in power relative to a fixed sample design with the same maximal sample size or some specified design.

```
> seqPlotPower(dsnJ4, dsnJ2)
```

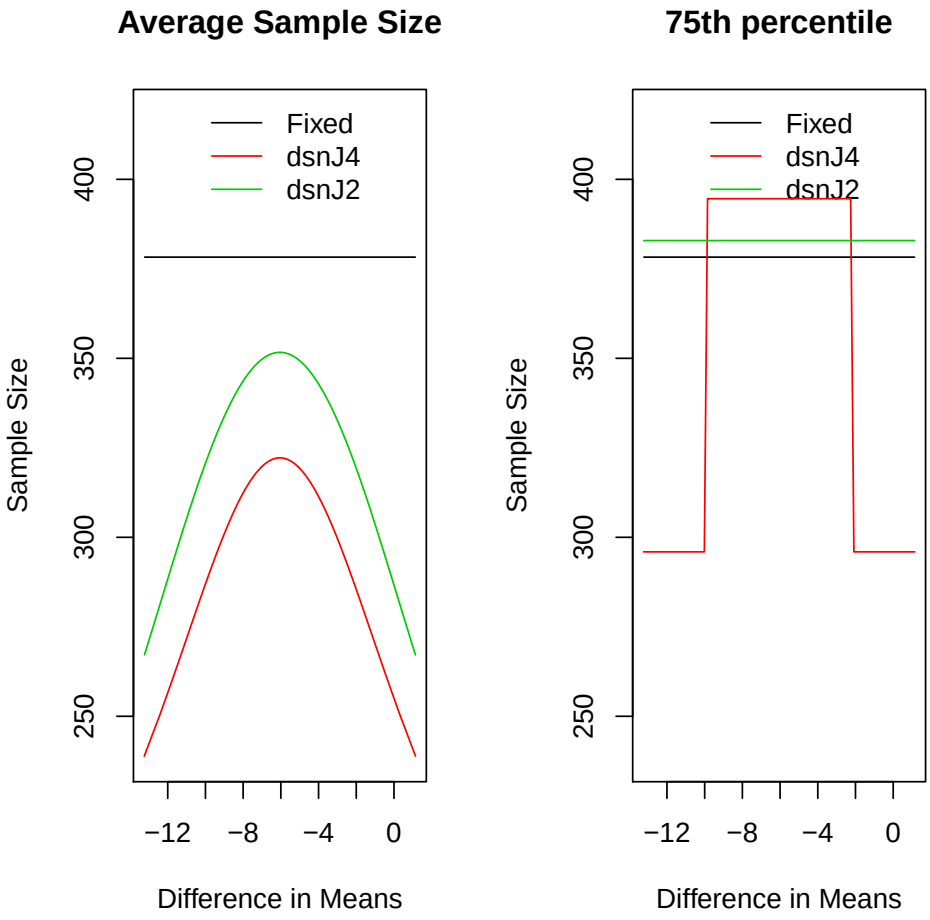



```
> seqPlotPower(dsnJ4, dsnJ2, reference=T)
```



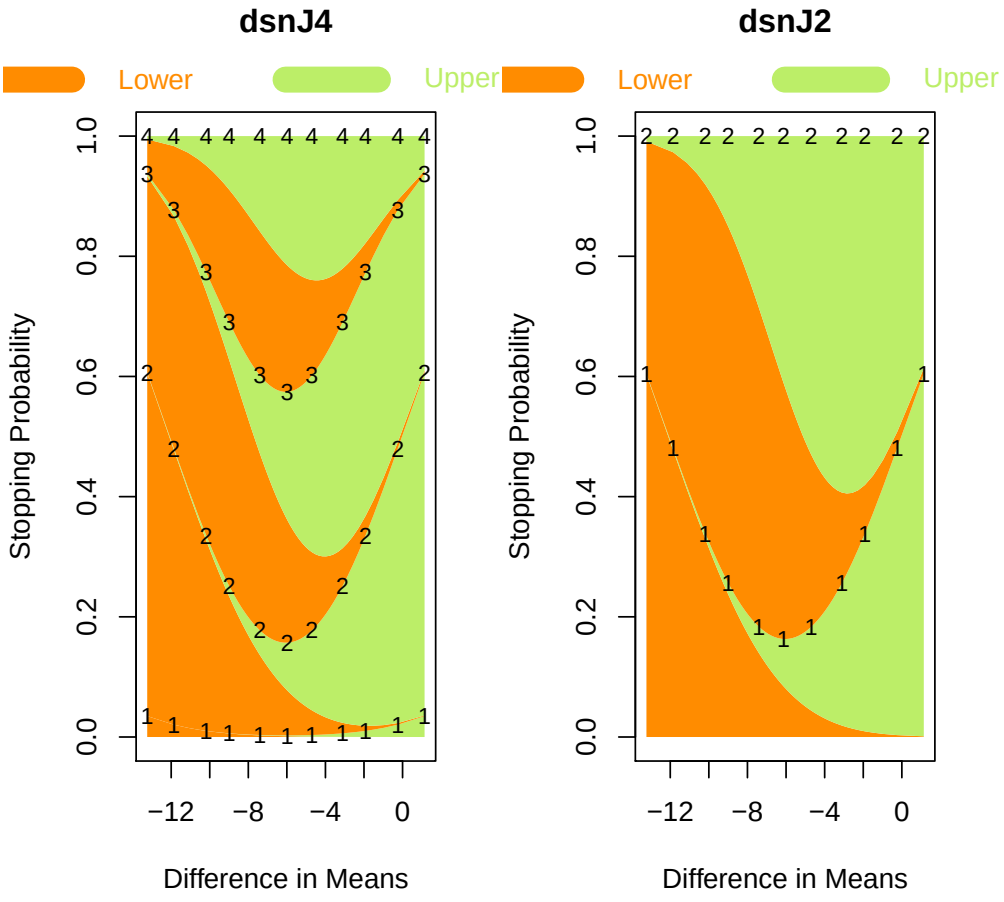
We can use `seqPlotASN()` to compare ASN curves displaying the expected sample size as a function of the true treatment effect. Quantiles (75th percentile by default) of the sample size distribution are also displayed. By default the fixed sample size test having the same power at the design alternative as the first listed design is included.

```
> seqPlotASN(dsnJ4, dsnJ2)
```



We can use `seqPlotStopProb()` to compare stopping probabilities as a function of the true treatment effect. By default these curves are not superposed. The lines indicated by numbers indicate the total cumulative stopping probability at each analysis, while the color coding indicates the decisions that would be made. It should be remembered that sample sizes at each analysis time may not be equal across designs.

```
> seqPlotStopProb(dsnJ4, dsnJ2)
```



5 Final Comments

The major purpose of this tutorial was just to illustrate the use of the most commonly used RCTdesign functions in the setting of a typical probability model. While some comments were made about the behavior of group sequential stopping rules, we do not pretend that we covered the theory and practice of sequential test design in this document. We do encourage you to browse through other tutorials (both on the Software pages and the Learning pages) to gain greater familiarity with those topics you find most useful in your area of clinical trial design.