"Mean" Probability Model

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

RCTdesign.org

August 21, 2012

Abstract

In this tutorial, we provide details about the "mean" probability model in RCTdesign. In the first section we provide the theory behind the test statistics, and then in the second section describe a hypothetical clinical trial setting that will be used in section 3 to provide examples of the specification of these models using `seqDesign()`.

1 Notation and General Distributional Theory

Notationally, we consider a setting in which we have n_k totally independent subjects on the kth study arm ($k = 0$ for the control arm and $k = 1$ for the arm receiving the experimental treatment).

We let Y_{ki} be the clinical outcome measurement on the ith individual on arm k, and we presume that the true average outcome is $E[Y_{ki}] = \mu_k$, and the true variance of response measurements is $Var(Y_{ki}) = \sigma_k^2$. (We write this $Y_{ki} \sim (\mu_k, \sigma_k^2)$.)

We make no other assumptions about the shape of the distribution on either study arm, including the fact that we make no assumption about any similarity or lack of similarity between the outcome probability distributions for the two study arms.

1.1 One Arm Studies

We suppose that all subjects with available data are assigned to the experimental treatment arm (so $n_0 = 0$).

- Our target of inference is $\theta = \mu_1$.
- We are interested in testing the null hypothesis $H_0 : \theta = \theta_0$.
- Because we make no assumptions about the shape of the outcome probability distribution, we use the distribution free estimate of the population mean, the sample mean $\overline{Y}$, as the basis for our inference:

$$\overline{Y} = \frac{1}{n_1} \sum_{i=1}^{n_1} Y_{1i}.$$

- In a fixed sample study (i.e., with no interim analyses), the central limit theorem provides that in moderate to large samples

$$\overline{Y} \dot{\sim} \mathcal{N}\left(\theta, \frac{\sigma_1^2}{n_1}\right),$$

hence under the null hypothesis we would have

$$Z = \sqrt{n_1}\frac{(\overline{Y} - \theta_0)}{\sigma_1} \doteq \mathcal{N}(0,1).$$

- We do not usually know σ_1^2, so we estimate it using the sample variance

$$s^2 = \frac{1}{n_1 - 1}\sum_{i=1}^{n_1}(Y_{1i} - \overline{Y})^2.$$

The sample variance s^2 is unbiased for σ_1^2 in a fixed sample setting, regardless of the underlying distribution. Furthermore, in that setting the sample variance s^2 is also consistent for the true variance σ_1^2, meaning that in large samples the probability that s^2 is different from σ_1^2 goes to 0. Hence, in hypothesis testing we typically use

$$T = \sqrt{n_1}\frac{(\overline{Y} - \theta_0)}{s},$$

which, under the null distribution, also has the approximate distribution $T \doteq \mathcal{N}(0,1)$. This approximate normal distribution (no matter the underlying probability distribution for clinical outcomes) holds in large fixed sample RCTs due to Slutsky's theorem, because σ_k/s tends toward 1 in large samples.

- In practice, we generally assume a t distribution (with $n-1$ degrees of freedom) for the distribution of T when testing the null hypothesis. This can be thought of as a "small sample adjustment" for the approximate distribution. This is exactly the correct thing to do when the distribution of the clinical outcome Y_{1i} is normally distributed, and it has become the accepted standard to use in other situations when Y_{1i} is a continuous random variable. Such a "small sample adjustment" is valid statistically in large samples, because the t distribution with $n-1$ degrees of freedom tends toward the standard normal distribution as n becomes large.

In sequential sampling, we continue to focus on the Z statistic as the basis for clinical trial design, and use the T statistic when analyzing simulated or actual RCT data.

- Our sequential sampling scheme involves performing analyses after accrual of $N_1, N_2, \ldots, N_J = n_1$ observations, and computing the statistics $\overline{Y}_j$, s_j^2, and T_j (computed using the first N_j observations) to stopping boundaries $a_j \le b_j \le c_j \le d_j$. The group sequential test statistic (M, T) is defined as
  - $M = \min\{1 \le j \le J : T_j \notin (a_j, b_j) \cup (c_j, d_j)\}$, and
  - $T = T_M$.

  Note that because the Jth analysis is the last analysis, we must have $a_J = b_J$ and $c_J = d_J$. For convenience, we assume that for $j \ne J$, $(a_j, b_j) \cup (c_j, d_j)$ is a nonempty proper subset of $(-\infty, \infty)$. (It truly poses no theoretical difficulty on derivation of the group sequential sampling distribution if this last assumption is violated.)

- At the time of planning our study, we have to use an estimated σ_1^2. As is usual, we pretend that value is the true value, and we derive stopping bounds that are truly appropriate for a Z_j statistic. We use the approximate joint distribution of $(Z_1, Z_2, \ldots, Z_J)$ based on a multivariate normal distribution to find boundaries that will have a desired experimentwise type I error. That multivariate distribution is known and computationally tractable, because our study structure and test statistic satisfies the "independent increment" structure that allows the methods of Armitage, McPherson, and Rowe (1969) to be applied.

- When simulating operating characteristics, the RCTdesign function `rSeq()` computes the T_j statistic on the simulated data. It then computes the fixed sample P value using the t distribution with $N_j - 1$ degrees of freedom. These p values are then compared to the stopping boundaries when the stopping boundaries are expressed on the fixed sample P value boundary scale. Pocock (1977) found that such a procedure (i.e., designing stopping rules based on multivariate normal statistics, but then implementing the boundaries based on fixed sample P values derived in other settings) was robust.
- When actually monitoring a clinical trial, the RCTdesign function `seqMonitor()` also converts stopping boundaries to the `"P"` boundary scale, and uses the fixed sample P values from a one-sample t test to make recommendations about continuing or terminating sampling.

1.2 Two Arm Studies

We suppose that subjects with available data are randomized in an $r : 1$ ratio to the experimental treatment and control arms (so $n_1 = rn_0$).

- Our target of inference is $\theta = \mu_1 - \mu_0$.
- We are interested in testing the null hypothesis $H_0 : \theta = \theta_0$.
- Because we make no assumptions about the shape of the outcome probability distribution, we use the distribution free estimate of the population means, the sample mean $\overline{Y}_k$, as the basis for our inference:

$$\overline{Y}_k = \frac{1}{n_k} \sum_{i=1}^{n_k} Y_{ki}.$$

- In a fixed sample study (i.e., with no interim analyses), the central limit theorem provides that in moderate to large samples

$$\overline{Y}_k \dot{\sim} \mathcal{N}\left(\mu_k, \frac{\sigma_k^2}{n_k}\right),$$

and by independence of the subjects on the two arms, we have

$$\hat{\theta} \equiv \overline{Y}_1 - \overline{Y}_0 \dot{\sim} \mathcal{N}\left(\theta = \mu_1 - \mu_0, \frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}\right).$$

Hence under the null hypothesis we would have

$$Z = \frac{(\hat{\theta} - \theta_0)}{\sqrt{\frac{\sigma_1^2}{n_1} + \frac{\sigma_0^2}{n_0}}} \dot{\sim} \mathcal{N}(0, 1).$$

- In a fixed sample study, we do not usually know the σ_k^2's, so we estimate them using the sample variance

$$s_k^2 = \frac{1}{n_k - 1} \sum_{i=1}^{n_k} (Y_{ki} - \overline{Y}_k)^2.$$

The sample variance s_k^2 is unbiased for σ_k^2 in a fixed sample setting, regardless of the underlying distribution. Furthermore, the sample variance s_k^2 is consistent for the true variance σ_k^2, meaning that in large samples the probability that s_k^2 is different from σ_k^2 goes to 0. Hence, in hypothesis testing we typically use

$$T = \frac{(\hat{\theta} - \theta_0)}{\sqrt{\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}}},$$

which, under the null distribution, also has the approximate distribution $T \dot{\sim} \mathcal{N}(0,1)$. The approximate normal distribution (no matter the underlying probability distribution for clinical outcomes) holds in large samples due to Slutsky's theorem, because σ_k/s_k tends toward 1 in large samples.

- In practice, we generally assume a t distribution with d degrees of freedom for the distribution of T when testing the null hypothesis, where d is typically computed using the Satterthwaite approximation

$$d = \frac{\left(\frac{s_1^2}{n_1} + \frac{s_0^2}{n_0}\right)^2}{\frac{s_1^4}{n_1^2(n_1-1)} + \frac{s_0^4}{n_0^2(n_0-1)}}.$$

  This test (often referred to as "the t test that allows for unequal variances") can be thought of as a "small sample adjustment" for the approximate distribution. This has been found to be a very good approximation (even in small samples) when the distribution of the clinical outcome Y_{1i} is normally distributed, and it has become the accepted standard to use in other situations when Y_{ki} is a continuous random variable. Such a "small sample adjustment" is valid statistically in large samples, because the t distribution with d degrees of freedom tends toward the standard normal distribution as d becomes large. Note, however, that for the large sample approximation to be valid, we need $\min(n_0, n_1)$ to become large.

- An alternative (but we think less desirable) approach is to estimate a "pooled sample variance" s_p^2

$$s_p^2 = \frac{(n_1-1)s_1^2 + (n_0-1)s_0^2}{n_1 + n_0 - 2},$$

  and then to compute a test statistic

$$T_{eq} = \frac{(\hat{\theta} - \theta_0)}{\sqrt{\frac{s_p^2}{n_1} + \frac{s_p^2}{n_0}}}.$$

  Under the null hypothesis, and using the consistency of s_k^2 for σ^2 and the $r:1$ randomization ratio, the approximate distribution for T_{eq} under the null hypothesis in moderate to large fixed sample RCTs can be shown to be

$$T_{eq} \dot{\sim} \mathcal{N}\left(0, \frac{\sigma_1^2 + r\sigma_0^2}{r\sigma_1^2 + \sigma_0^2}\right).$$

  Note that for this asymptotic (large sample) distribution:

  - The variance of this asymptotic distribution is 1 if $r = 1$ or if $\sigma_1^2 = \sigma_0^2$.
  - The variance is greater than 1 if $r > 1$ and $\sigma_1^2 < \sigma_0^2$ or if $r < 1$ and $\sigma_1^2 > \sigma_0^2$.
  - The variance is less than 1 if $r > 1$ and $\sigma_1^2 > \sigma_0^2$ or if $r < 1$ and $\sigma_1^2 < \sigma_0^2$.

- In practice, it is usually presumed that T_{eq} has a t distribution with $d = n_1 + n_0 - 2$ degrees of freedom. In the setting of $r = 1$ or equal variances (i.e., $\sigma_1^2 = \sigma_0^2$), this "t test that presumes equal variances" can be thought of as a "small sample adjustment" that is exactly the correct distribution (even in small samples) when the clinical outcomes Y_{ki} are normally distributed with $\sigma_1^2 = \sigma_0^2$. Some important caveats:

  - Again, the fact that a t distribution with d distribution tends toward a standard normal distribution as d gets large, this test is statistically valid in large samples when the variances are equal or $r = 1$.
  - However, the test is not a statistically valid test of the means when the randomization ratio is not 1 and the variances are unequal.

  - In the latter setting, the use of T_{eq} will tend toward anti-conservative inference (i.e., p values tending to be too small under the null hypothesis and confidence intervals too narrow) when the larger sample size is on the arm with the smaller variance.
  - The use of T_{eq} will tend toward conservative inference (i.e., p values tending to be too large under the null hypothesis and confidence intervals too wide) when the larger sample size is on the arm with the larger variance.

In sequential sampling, we continue to focus on the Z statistic as the basis for clinical trial design, and use the T statistic (or if you really want it, the T_{eq} statistic) when analyzing simulated or actual RCT data.

- Our sequential sampling scheme involves performing analyses after accrual of $N_1, N_2, \ldots, N_J = n_1 + n_0$ observations, and computing the statistics $\overline{Y}_{kj}$, s^2_{kj}, and T_j (computed using the first N_j observations, where N_j represents the total number accrued to both arms combined at the time of the jth analysis) to stopping boundaries $a_j \leq b_j \leq c_j \leq d_j$. The group sequential test statistic (M, T) is defined as
  - $M = \min\{1 \leq j \leq J : T_j \notin (a_j, b_j) \cup (c_j, d_j)\}$, and
  - $T = T_M$.

  Note that because the Jth analysis is the last analysis, we must have $a_J = b_J$ and $c_J = d_J$. For convenience, we assume that for $j \neq J$, $(a_j, b_j) \cup (c_j, d_j)$ is a nonempty proper subset of $(-\infty, \infty)$. (It truly poses no theoretical difficulty on derivation of the group sequential sampling distribution if this last assumption is violated.)
- At the time of planning our study, we have to use an estimated σ^2_k. As is usual, we pretend that value is the true value, and we derive stopping bounds that are truly appropriate for a Z_j statistic. We use the approximate joint distribution of $(Z_1, Z_2, \ldots, Z_J)$ based on a multivariate normal distribution to find boundaries that will have a desired experimentwise type I error.
- When simulating operating characteristics, the RCTdesign function `rSeq()` computes the T_j statistic on the simulated data. It then computes the fixed sample P value using the t distribution with d degrees of freedom using the Satterthwaite approximation. The user can optionally choose to use T_{eq} instead. These p values are then compared to the stopping boundaries when the stopping boundaries are expressed on the fixed sample P value boundary scale. Pocock (1977) found that such a procedure (i.e., designing stopping rules based on multivariate normal statistics, but then implementing the boundaries based on fixed sample P values derived in other settings) was robust.
- When actually monitoring a clinical trial, the RCTdesign function `seqMonitor()` also converts stopping boundaries to the `"P"` boundary scale, and uses the fixed sample P values from a one-sample t test to make recommendations about continuing or terminating sampling.

2 Description of Setting for Illustrative Examples

We illustrate the specification of the mean probability model in the context of a hypothetical clinical trial. As the focus here is how the inputs to the `seqDesign()` function might differ in a number of scenarios, we do not alter other aspects of the design across those scenarios. Hence, we will arbitrarily consider one-sided symmetric group sequential designs (Emerson & Fleming, 1989) having a maximum of $J = 4$ analyses. O'Brien-Fleming boundary relationships will be used, and the type I error will be set at $\alpha = 0.025$ (one-sided), and the statistical power to detect the design alternative will be set at 90%.

2.1 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.

2.2 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.
- *Clinical outcome* Systolic blood pressure measured 6 months after randomization.
- *Study structure* For the purposes of our illustration, we consider both a one arm study (uncontrolled or using historical controls), a two arm study (concurrent controls ideally randomized in a double blind fashion to receive the experimental treatment or placebo), and a randomized crossover study (each patient receives both the experimental treatment and placebo in random order).

2.3 Notation for Available Data and Probability Model

We first define a notation that can be used to compare and contrast a variety of RCT designs that might be used in this setting. *(We note that for our hypothetical setting, we would not actually recommend all of the study designs presented here. But we do find it useful to illustrate the general use of RCTdesign in these varied settings, while at the same time raising some of the scientific and statistical issues involved in choosing one particular approach from among the alternatives.)*

In the most general case considered here, we use a distribution-free probability model in which

- X_{kit} represents the systolic blood pressure for the ith patient $i = 1, \ldots, n_k$ 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 $X_{kit} \sim (\mu_{kt}, \tau^2_{kt})$, 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 τ^2_{kt} at time t.
- We presume that all patients are independent of each other, and that patients receiving the experimental treatment in two arm studies are different from the patients receiving placebo in those studies. In the crossover study, the ith patient during the experimental treatment period will be the same person as the ith patient during the placebo period.

- 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 (and under the presumption of no carryover effects in the crossover design), we know that measurements made at the start of treatment have the same distribution on both study arms. Hence, $\mu_{00} = \mu_{10}$ and $\tau^2_{00} = \tau^2_{10}$.
- Apart from the above assumptions, we do not otherwise characterize the distribution of X_{kit}. In particular, we make no assumption that the distributions of Y_{ki0} and X_{ki6} have the same shape, nor do we assume that the distributions of X_{0i6} and X_{1i6} have the same shape.

2.4 Presumed Baseline Distribution and Hypothesized Treatment Effect

In the absence of treatment:

- *SBP at Time of Study Accrual*: We presume that the average SBP at the time of study accrual is 160 mmHg, with a standard deviation of 27.386 mmHg.
- *Average Trend over 6 Months in the absence of treatment*: We presume that in the absence of treatment, subjects' average a 3 mmHg increase in SBP. We further presume no time effect on the variability of measurements. Hence, for subjects receiving placebo during the first 6 months following study accrual, we presume the average SBP would be 163 mmHg, and the standard deviation would be 27.386 mmHg.
- *Correlation in Measurements over Time*: We presume that the correlation between SBP measurements made on the same subject 6 months apart does not depend upon treatment. We presume $\rho_1 = \rho_0 = 0.4$
- *Treatment effect*: We presume that prior pilot studies have suggested that the treatment might provide a benefit that corresponds to an average decreased SBP of 10 mmHg compared to what it might otherwise have been in the absence of treatment, and that the treatment does not affect the variation in SBP measurements. Hence, if this hypothesis is true, among subjects receiving the experimental treatment during the first 6 months post study accrual, the average SBP at 6 months would be 153 mmHg, with a standard deviation of 27.386 mmHg.

2.5 Other Study Design Parameters

In all scenarios considered:

- *Test type*: We presume a test of the null hypothesis versus a one-sided lesser alternative.
- *Type I error*: We presume that standards of evidence demand a one-sided type I error of $\alpha = 0.025$.
- *Design power*: We presume that we desire 90% statistical power to reject the null hypothesis when the alternative hypothesis is in fact true.
- *Sequential sampling plan*: We presume that up to four equally spaced analyses will be performed. A one-sided symmetric design (Emerson & Fleming, 1989) having O'Brien-Fleming boundary relationships will be used.
- *Probability model*: Statistical inference will summarize treatment effect using means. Statistical tests will be as described in section 1.
- *Design task*: We want to find the sample size required to satisfy the above criteria.

3 Specification of Selected Possible Clinical Trial Designs

In the general setting described in section 2, there are a number of alternative study designs that can be considered:

- One Arm Study: Mean SBP 6 Months after Accrual
- One Arm Study: Mean Change in SBP During 6 Months Following Accrual
- Two Arm Study: Difference Between Study Arms in Mean SBP 6 Months after Randomization
- Two Arm Study: Difference Between Study Arms in Mean Change in SBP 6 Months after Randomization
- Two Arm Study: Difference Between Study Arms in Mean SBP at 6 Months after Randomization with Adjustment for Baseline in a Regression Model
- Crossover Study: Difference Between Study Periods in Mean SBP 6 Months after Start of Treatment

We address each of these in the following subsections. For each scenario, we relate the notation used in section 1 (Y_{ki}, σ_k^2, and θ) to the notation used to describe the potentially available data in section 2 (X_{kit}, μ_{kt}, τ_{kt}^2, and ρ_k).

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)
```

3.1 One Arm Study: Mean SBP 6 Months after Accrual

We consider a one arm, uncontrolled study that will compare the mean SBP after 6 months of treatment to our belief (based on historical data?) about the mean SBP after 6 months on study in the absence of a treatment effect. *(We do not imagine anyone would actually plan such a study, because we derived our hypotheses presuming that we knew in advance what the mean SBP was at time of study accrual and how the mean would change over time in the absence of treatment. Neither of these sorts of presumptions are likely to be reliable, and hence it is highly unlikely that an investigator would plan such a study.)*

In designing this RCT, we would regard that

- The sample size on placebo is $n_0 = 0$.
- The measurement made on the ith subject is the SBP measurement made at 6 months post study accrual: $Y_{1i} = X_{1i6}$.
- The treatment effect is defined as the mean SBP at 6 months: $\theta = \mu_{16}$
- The variability of the clinical measurement is $\sigma_1^2 = \tau_{16}^2$.
- The null hypothesis of no treatment effect corresponds to the presumed mean SBP at baseline plus the presumed change in SBP over 6 months in the absence of a treatment effect: $\theta_0 = 160 + 3 = 163$ mmHg.
- The alternative hypothesis corresponds to a mean SBP that is 10 mm Hg lower than what it would be under the null hypothesis: $\theta_1 = 163 - 10 = 153$ mmHg.

The following R code would produce a clinical trial design meeting these criteria:

```
> dsn1 <- seqDesign(
+         prob.model= "mean",
+         arms= 1,
+         null.hypothesis= 163,
+         alt.hypothesis= 153,
+         sd= 27.386,
+         test.type= "less",
+         size= 0.025,
+         power= 0.90,
+         nbr.analyses= 4,
+         design.family= "X",
+         P= 1
+         )
> dsn1

Call:
seqDesign(prob.model = "mean", arms = 1, null.hypothesis = 163,
    alt.hypothesis = 153, sd = 27.386, nbr.analyses = 4, test.type = "less",
    size = 0.025, power = 0.9, design.family = "X", P = 1)

PROBABILITY MODEL and HYPOTHESES:
   Theta is mean response
   One-sided hypothesis test of a lesser alternative:
           Null hypothesis : Theta >= 163    (size  = 0.025)
    Alternative hypothesis : Theta <= 153    (power = 0.900)
   (Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                     Efficacy Futility
    Time 1 (N= 20.55) 138.7970 175.1015
    Time 2 (N= 41.10) 150.8985 163.0000
    Time 3 (N= 61.65) 154.9323 158.9662
    Time 4 (N= 82.21) 156.9492 156.9492
```

Because

- the mean probability model is the default probability model,
- a one-sided test type is the default and the lesser hypothesis can be inferred from the values of the null and alternative hypotheses,
- the default type I error in a one-sided hypothesis test is 0.025,
- the family of designs on the estimate (sample mean) scale (the unified family) is the default,
- the O'Brien-Fleming boundary relationship (corresponding to `P=1`) is the default boundary relationship, and
- argument names can be abbreviated to the length adequate to distinguish them from all other arguments,

the same design could have been obtained by the following shorter code:

```
> dsn1 <- seqDesign( arms= 1, null= 163, alt= 153, sd= 27.386, power= 0.90, nbr= 4)
> dsn1

Call:
seqDesign(arms = 1, null.hypothesis = 163, alt.hypothesis = 153,
    sd = 27.386, nbr.analyses = 4, power = 0.9)

PROBABILITY MODEL and HYPOTHESES:
   Theta is mean response
   One-sided hypothesis test of a lesser alternative:
           Null hypothesis : Theta >= 163    (size  = 0.025)
    Alternative hypothesis : Theta <= 153    (power = 0.900)
   (Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                      Efficacy Futility
    Time 1 (N= 20.55) 138.7970 175.1015
    Time 2 (N= 41.10) 150.8985 163.0000
    Time 3 (N= 61.65) 154.9323 158.9662
    Time 4 (N= 82.21) 156.9492 156.9492
```

From the above printout of the study design, we see

- The study would require a maximum sample size of 82.21. (The fractional sample size is sometimes relevant in "information" scale monitoring, but it is clearly not of interest in this case. In real life, we would likely round the sample size up to 83.)
- At a first analysis that occurred when 25% of the maximal sample size had accrued, an observed average SBP of 138.797 mmHg or less would correspond to a recommendation that the study be stopped and the null hypothesis be rejected. An observed average SBP of 175.1015 mmHg or greater would correspond to a recommendation that the study be stopped with a decision not to reject the null hypothesis. If the observed average SBP is between those two values, the recommendation would be to continue the trial to accrue another 25% of the planned maximal sample size.
- Similar interpretations apply to the stopping boundaries at the second and third planned analyses.
- According to these boundaries, the trial would only continue to the maximal sample size only if the observed mean SBP were between 154.9323 mmHg and 158.9662 mmHg at the third analysis.
- If the trial continued to the last analysis, an observed average SBP below 156.9492 mmHg would correspond to a decision to reject the null hypothesis. Otherwise, we would not reject the null hypothesis.

It should be noted that the above boundaries are specific to the number and schedule of analyses, as well as being specific to the presumed variability of the measurements. During the actual conduct of the study, the boundaries would typically be modified to account for the actual observed variability and the actual schedule of analyses. (See tutorials on implementation of stopping rules.)

3.2 One Arm Study: Mean Change in SBP During 6 Months Following Accrual

We consider a one arm, uncontrolled study that will compare the mean change in SBP during 6 months of treatment to our belief (based on historical data?) about the mean change in SBP after 6 months on study

in the absence of a treatment effect. *(This sort of a study design is regrettably often proposed. Sometimes the investigator will erroneously claim that "each subject serves as his/her own control". But generally the assumption will be that there would be no difference in mean SBP over time in the absence of a treatment effect, rather than the mean change of 3 mmHg we imagined. However, this is still an uncontrolled study that does not allow for the possibility of aging, disease progression, calendar time trends, regression to the mean, or the "Hawthorne" effect, among other known and unknown factors that might lead to changes in SBP. Again, we present this example more to show how RCTdesign is used, than to serve as an example of good clinical trial design.)*

In designing this RCT, we would regard that

- The sample size on placebo is $n_0 = 0$.
- The measurement made on the ith subject is the change in SBP measurements during the six months on treatment: $Y_{1i} = X_{1i6} - X_{1i0}$.
- The treatment effect is defined as the mean change in SBP during 6 months post study accrual: $\theta = \mu_{16} - \mu_{10}$
- The variability of the clinical measurement can be calculated from the hypothesized variability of measurements made at a single point in time and the hypothesized correlation between measurements made on the same individual 6 months apart: $\sigma_1^2 = \tau_{16}^2 + \tau_{10}^2 - 2\rho_1\tau_{16}\tau_{10}$. Because we hypothesize that $\tau_{k6}^2 = \tau_{k0}^2$, we have $\sigma_1^2 = 2\tau_{10}^2(1-\rho_1) = 2 \times 27.386^2 \times (1-0.4) = 900$, so $\sigma_1 = 30$. (Of course, if we had data available on the change in SBP over 6 months, we could have just found the standard deviation of those measurements.)
- The null hypothesis of no treatment effect corresponds to the the presumed change in SBP over 6 months in the absence of a treatment effect: $\theta_0 = 3$ mmHg.
- The alternative hypothesis corresponds to a mean SBP that is 10 mm Hg lower than what it would be under the null hypothesis: $\theta_1 = 3 - 10 = -7$ mmHg.

The following R code would produce a clinical trial design meeting these criteria:

```
> dsn2 <- seqDesign(
+         prob.model= "mean",
+         arms= 1,
+         null.hypothesis= 3,
+         alt.hypothesis= -7,
+         sd= 30,
+         test.type= "less",
+         size= 0.025,
+         power= 0.90,
+         nbr.analyses= 4,
+         design.family= "X",
+         P= 1
+         )
> dsn2

Call:
seqDesign(prob.model = "mean", arms = 1, null.hypothesis = 3,
    alt.hypothesis = -7, sd = 30, nbr.analyses = 4, test.type = "less",
    size = 0.025, power = 0.9, design.family = "X", P = 1)
```

```
PROBABILITY MODEL and HYPOTHESES:
   Theta is mean response
   One-sided hypothesis test of a lesser alternative:
           Null hypothesis : Theta >=  3    (size  = 0.025)
    Alternative hypothesis : Theta <= -7    (power = 0.900)
   (Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                     Efficacy Futility
    Time 1 (N= 24.66) -21.2030  15.1015
    Time 2 (N= 49.32)  -9.1015   3.0000
    Time 3 (N= 73.99)  -5.0677  -1.0338
    Time 4 (N= 98.65)  -3.0508  -3.0508
```

Because

- the mean probability model is the default probability model,
- a one-sided test type is the default and the lesser hypothesis can be inferred from the values of the null and alternative hypotheses,
- the default type I error in a one-sided hypothesis test is 0.025,
- the family of designs on the estimate (sample mean) scale (the unified family) is the default,
- the O'Brien-Fleming boundary relationship (corresponding to `P=1`) is the default boundary relationship, and
- argument names can be abbreviated to the length adequate to distinguish them from all other arguments,

the same design could have been obtained by the following shorter code:

```
> dsn2 <- seqDesign( arms= 1, null= 3, alt= -7, sd= 30, power= 0.90, nbr= 4)
> dsn2

Call:
seqDesign(arms = 1, null.hypothesis = 3, alt.hypothesis = -7,
    sd = 30, nbr.analyses = 4, power = 0.9)

PROBABILITY MODEL and HYPOTHESES:
   Theta is mean response
   One-sided hypothesis test of a lesser alternative:
           Null hypothesis : Theta >=  3    (size  = 0.025)
    Alternative hypothesis : Theta <= -7    (power = 0.900)
   (Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                     Efficacy Futility
    Time 1 (N= 24.66) -21.2030  15.1015
    Time 2 (N= 49.32)  -9.1015   3.0000
    Time 3 (N= 73.99)  -5.0677  -1.0338
    Time 4 (N= 98.65)  -3.0508  -3.0508
```

From the above printout of the study design, we see

- The study would require a maximum sample size of 98.65. (The fractional sample size is sometimes relevant in "information" scale monitoring, but it is clearly not of interest in this case. In real life, we would likely round the sample size up to 99.)
- At a first analysis that occurred when 25% of the maximal sample size had accrued, an observed average change in SBP of -21.203 mmHg or less would correspond to a recommendation that the study be stopped and the null hypothesis be rejected. An observed average change in SBP of 15.1015 mmHg or greater would correspond to a recommendation that the study be stopped with a decision not to reject the null hypothesis. If the observed average SBP is between those two values, the recommendation would be to continue the trial to accrue another 25% of the planned maximal sample size.
- Similar interpretations apply to the stopping boundaries at the second and third planned analyses.
- According to these boundaries, the trial would only continue to the maximal sample size only if the observed mean change in SBP were between -5.0677 mmHg and -1.0338 mmHg at the third analysis.
- If the trial continued to the last analysis, an observed average change in SBP below -3.0508 mmHg would correspond to a decision to reject the null hypothesis. Otherwise, we would not reject the null hypothesis.

Note that the change in measurements is more variable than the final measurement, and thus this design requires a larger sample size than a design based solely on the final measurement. This is because the correlation was less than 0.5. However, despite this loss of precision, this design would likely seem more acceptable because it does not rely on an absolute criterion for the final blood pressure alone. Basing decisions on the change would undoubtedly seem more intuitive to an interested observer, even though there are still many problems with the uncontrolled study.

3.3 Two Arm Study: Difference Between Study Arms in Mean SBP 6 Months after Randomization

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. In the uncontrolled study of section 3.1, we had to rely on prior estimates of what the average blood pressure would be in an untreated population 6 months after study accrual. In the uncontrolled study of section 3.2, we looked at the change in SBP, but we still had to rely on prior estimates of how aging, time trends in other behaviors, regression to the mean, etc. might affect what the mean change in SBP would be in the absence of treatment. Clearly neither of these approaches are very rigorous scientifically. ("There are two kinds of researchers: those with a lot of enthusiasm and no controls, or those with a lot of controls and no enthusiasm.")

A randomized study comparing the experience of a group of patients receiving the experimental treatment to the experience of a group of otherwise similar patient receiving placebo solves this problem. The concurrent independent control group allows an unbiased estimate of what the clinical outcome would be in an untreated population.

In this section we consider a clinical trial design in which the outcome is based only on the SBP measurement made 6 months after randomization.

In designing this RCT, we would regard that

- We decide to use 1:1 randomization, so the sample sizes are the same on the experimental treatment and on placebo: $n_1 = n_0$.

- The measurement made on the ith subject is the SBP measurement made six months after randomization: $Y_{ki} = X_{ki6}$.
- The treatment effect is defined as the difference between the study arms (treatment minus control) in mean SBP 6 months post randomization: $\theta = \mu_{16} - \mu_{06}$.
- The variability of the clinical measurement is hypothesized to be the same on each study arm: $\tau^2_{k6} = 27.386^2$.
- The null hypothesis of no treatment effect corresponds to a difference of 0 between the means: $\theta_0 = 0$ mmHg.
- The alternative hypothesis corresponds to a mean SBP on the experimental treatment arm that is 10 mm Hg lower than what it would be on the placebo arm: $\theta_1 = -10$ mmHg.

The following R code would produce a clinical trial design meeting these criteria:

```
> dsn3 <- seqDesign(
+         prob.model= "mean",
+         arms= 2,
+         ratio= c(1,1),
+         null.hypothesis= c(163,163),
+         alt.hypothesis= c(153,163),
+         sd= c(27.386, 27.386),
+         test.type= "less",
+         size= 0.025,
+         power= 0.90,
+         nbr.analyses= 4,
+         design.family= "X",
+         P= 1
+         )
> dsn3

Call:
seqDesign(prob.model = "mean", arms = 2, null.hypothesis = c(163,
    163), alt.hypothesis = c(153, 163), sd = c(27.386, 27.386),
    ratio = c(1, 1), nbr.analyses = 4, test.type = "less", size = 0.025,
    power = 0.9, design.family = "X", P = 1)

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=  82.21) -24.2030  12.1015
    Time 2 (N= 164.41) -12.1015   0.0000
    Time 3 (N= 246.62)  -8.0677  -4.0338
    Time 4 (N= 328.82)  -6.0508  -6.0508
```

Because

- the mean probability model is the default probability model,
- a two arm study is the default,
- 1:1 randomization is the default,
- we do not really need to know the means for each study arm, only the difference (note that the hypotheses printed in the above output just provide the difference between the means),
- in a two arm study, the default null hypothesis is a difference of 0 across study arms (or a mean of 0 on each arm, which corresponds to a difference of 0),
- the default assumes the same variability (standard deviation) on both arms,
- a one-sided test type is the default and the lesser hypothesis can be inferred from the values of the null and alternative hypotheses,
- the default type I error in a one-sided hypothesis test is 0.025,
- the family of designs on the estimate (sample mean) scale (the unified family) is the default,
- the O'Brien-Fleming boundary relationship (corresponding to `P=1`) is the default boundary relationship, and
- argument names can be abbreviated to the length adequate to distinguish them from all other arguments,

the same design could have been obtained by the following shorter code:

```
> dsn3 <- seqDesign( alt= -10, sd= 27.386, power= 0.90, nbr= 4)
> dsn3

Call:
seqDesign(alt.hypothesis = -10, sd = 27.386, nbr.analyses = 4,
    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=  82.21) -24.2030  12.1015
    Time 2 (N= 164.41) -12.1015   0.0000
    Time 3 (N= 246.62)  -8.0677  -4.0338
    Time 4 (N= 328.82)  -6.0508  -6.0508
```

From the above printout of the study design, we see

- The study would require a maximum sample size of 328.82 on the two treatment arms combined. (The fractional sample size is sometimes relevant in "information" scale monitoring, but it is clearly not of interest in this case. In real life, we would likely round the sample size up to 330– 115 on each arm.)
- At a first analysis that occurred when 25% of the maximal sample size had accrued, an observed difference between study arms in average SBP (experimental treatment average minus control average) of -24.203 mmHg or less would correspond to a recommendation that the study be stopped and the null hypothesis be rejected. An observed difference between study arms in average SBP of 12.1015 mmHg or greater would correspond to a recommendation that the study be stopped with a decision not to reject the null hypothesis. If the observed average SBP is between those two values, the recommendation would be to continue the trial to accrue another 25% of the planned maximal sample size.
- Similar interpretations apply to the stopping boundaries at the second and third planned analyses.
- According to these boundaries, the trial would only continue to the maximal sample size only if the observed difference in mean SBP were between -8.0677 mmHg and -4.0338 mmHg at the third analysis.
- If the trial continued to the last analysis, an observed difference in average SBP below -3.0508 mmHg would correspond to a decision to reject the null hypothesis. Otherwise, we would not reject the null hypothesis.

Note that this design does not presume that you know the mean in an untreated group, so a larger sample size is required to accomodate the variability in the resulting estimate. The end result is a total sample size that is 4 times greater than what it would be if you knew what the control group's true mean SBP was.

3.4 Two Arm Study: Difference Between Study Arms in Mean Change in SBP 6 Months after Randomization

In this section we consider a clinical trial design in which the outcome is based on the change in SBP measurements during the 6 months after randomization. *(This is probably the analysis that would be chosen by most people, unless they were aware of results about the ANCOVA model given below. While it seems natural to believe that taking differences should be the more statistically precise analysis method, in this example we illustrate a setting in which this approach is less precise than the previous analysis that ignores the baseline measurement.)*

In designing this RCT, we would regard that

- We decide to use 1:1 randomization, so the sample sizes are the same on the experimental treatment and on placebo: $n_1 = n_0$.
- The measurement made on the ith subject is the change in SBP measurements during the six months after randomization: $Y_{ki} = X_{ki6} - X_{ki0}$.
- The treatment effect is defined as the difference between the study arms (treatment minus control) in mean change in SBP over the 6 months post randomization: $\theta = \mu_{16} - \mu_{06}$.
- The variability of the clinical measurement can be calculated from the hypothesized variability of measurements made at a single point in time and the hypothesized correlation between measurements made on the same individual 6 months apart: $\sigma_1^2 = \tau_{16}^2 + \tau_{10}^2 - 2\rho_1\tau_{16}\tau_{10}$. Because we hypothesize that $\tau_{k6}^2 = \tau_{k0}^2$, we have $\sigma_1^2 = 2\tau_{10}^2(1-\rho_1) = 2 \times 27.386^2 \times (1-0.4) = 900$, so $\sigma_1 = 30$. (Of course, if we had data available on the change in SBP over 6 months, we could have just found the standard deviation of those measurements.)

- The null hypothesis of no treatment effect corresponds to a difference of 0 between the means: $\theta_0 = 0$ mmHg.
- The alternative hypothesis corresponds to a mean SBP on the experimental treatment arm that is 10 mm Hg lower than what it would be on the placebo arm: $\theta_1 = -10$ mmHg.

The following R code would produce a clinical trial design meeting these criteria:

```
> dsn4 <- seqDesign(
+         prob.model= "mean",
+         arms= 2,
+         ratio= c(1,1),
+         null.hypothesis= c(3,3),
+         alt.hypothesis= c(-7,3),
+         sd= c(30,30),
+         test.type= "less",
+         size= 0.025,
+         power= 0.90,
+         nbr.analyses= 4,
+         design.family= "X",
+         P= 1
+         )
> dsn4

Call:
seqDesign(prob.model = "mean", arms = 2, null.hypothesis = c(3,
    3), alt.hypothesis = c(-7, 3), sd = c(30, 30), ratio = c(1,
    1), nbr.analyses = 4, test.type = "less", size = 0.025, power = 0.9,
    design.family = "X", P = 1)

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
```

Because

- the mean probability model is the default probability model,
- a two arm study is the default,
- 1:1 randomization is the default,

- we do not really need to know the means for each study arm, only the difference (note that the hypotheses printed in the above output just provide the difference between the means),
- in a two arm study, the default null hypothesis is a difference of 0 across study arms (or a mean of 0 on each arm, which corresponds to a difference of 0),
- the default assumes the same variability (standard deviation) on both arms,
- a one-sided test type is the default and the lesser hypothesis can be inferred from the values of the null and alternative hypotheses,
- the default type I error in a one-sided hypothesis test is 0.025,
- the family of designs on the estimate (sample mean) scale (the unified family) is the default,
- the O'Brien-Fleming boundary relationship (corresponding to `P=1`) is the default boundary relationship, and
- argument names can be abbreviated to the length adequate to distinguish them from all other arguments,

the same design could have been obtained by the following shorter code:

```
> dsn4 <- seqDesign( alt= -10, sd= 30, power= 0.90, nbr= 4)
> dsn4

Call:
seqDesign(alt.hypothesis = -10, sd = 30, nbr.analyses = 4, 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
```

From the above printout of the study design, we see

- The study would require a maximum sample size of 394.59 on the two treatment arms combined. (The fractional sample size is sometimes relevant in "information" scale monitoring, but it is clearly not of interest in this case. In real life, we would likely round the sample size up to 396– 198 on each arm.)
- At a first analysis that occurred when 25% of the maximal sample size had accrued, an observed difference between study arms in average SBP (experimental treatment average minus control average) of -24.203 mmHg or less would correspond to a recommendation that the study be stopped and the null hypothesis be rejected. An observed difference between study arms in average SBP of 12.1015 mmHg or

greater would correspond to a recommendation that the study be stopped with a decision not to reject the null hypothesis. If the observed average SBP is between those two values, the recommendation would be to continue the trial to accrue another 25% of the planned maximal sample size.

- Similar interpretations apply to the stopping boundaries at the second and third planned analyses.
- According to these boundaries, the trial would only continue to the maximal sample size only if the observed difference in mean SBP were between -8.0677 mmHg and -4.0338 mmHg at the third analysis.
- If the trial continued to the last analysis, an observed difference in average SBP below -3.0508 mmHg would correspond to a decision to reject the null hypothesis. Otherwise, we would not reject the null hypothesis.

Note that this design is less precise than the design based on using only the final SBP measurements. Examination of the formula for variance of a difference shows that unless the correlation between measurements is at least 0.5, the variability of the treatment effect based on mean change is greater than the variability of the treatment effect based only on mean SBP at the end of the study. Of course, the fact that we are free to choose between the two measures arises out of randomization: $\mu_{10} = \mu_{00}$. Hence for an arbitrary scalar c in the two arm study setting, an estimator $\hat{\theta}_c$ defined as

$$\hat{\theta}_c = (\overline{X}_{16} - c\overline{X}_{10}) - (\overline{X}_{06} - c\overline{X}_{00})$$

would have $E(\hat{\theta}_c) = \mu_{16} - \mu_{06} = \theta$. The choice $c = 0$ corresponds to using only the final SBP measurements for each individual as in section 3.3, and the choice $c = 1$ corresponds to using the change in SBP measurements as in this section.

Though all such estimators are thus unbiased for θ, they differ in their variability. We find

$$Var(\hat{\theta}_c) = \tau_{16}^2 + c^2\tau_{10}^2 - 2c\rho_1\tau_{16}\tau_{10} + \tau_{06}^2 + c^2\tau_{00}^2 - 2c\rho_0\tau_{06}\tau_{00}$$
$$\tau_{16}^2 + \tau_{06}^2 + 2c^2\tau_{10}^2 - 2c\tau_{10}(\rho_1\tau_{16} + \rho_0\tau_{06})$$

where we have used the equality of variances at baseline that arises through randomization. In the setting where treatment does not affect within group variability or within group correlation, we would have

$$Var(\hat{\theta}_c) = \tau_{16}^2(2 + 2c^2 - 4c\rho_1).$$

Differentiating with respect to c shows that the minimal variance is achieved in this case when $c = \rho_1$. It can further be shown that if the SBP measurements were normally distributed in this homoscedastic RCT setting, the estimator $\hat{\theta}_\rho$ is the uniform minimum variance unbiased estimator of θ.

(We note again the reliance on randomization to ensure that $\hat{\theta}_c$ was unbiased for θ for all choices of c. In observational data, the measure of treatment effect based only on final measurements ($\mu_{16} - \mu_{06}$) may be different from the measure of treatment effect based on the change in measurements ($(\mu_{16} - \mu_{10}) - (\mu_{06} - \mu_{00})$) . Therefore, in observational data, a decision would have to be made according to which measure was scientifically most relevant.)

3.5 Two Arm Study: Difference Between Study Arms in Mean SBP at 6 Months after Randomization with Adjustment for Baseline in a Regression Model

In this section we consider a clinical trial design in which the outcome is based on the SBP measurement 6 months after randomization, but the analysis is adjusted for the baseline in a linear regression analysis.

Because the linear model includes a binary indicator of treatment assignment along with the continuous baseline measurement, this is commonly referred to as the analysis of covariance, or ANCOVA, model.

The linear model is thus

$$E[X_{ki6} \mid k, X_{ki0}] = \beta_0 + \beta_1 X_{ki0} + \theta k,$$

where regression parameter β_1 will be involve the τ_{kt}'s and the ρ_k's. In the setting in which the treatment has no effect on within group variability or within group correlation,

$$\beta_1 = \rho_1.$$

Hence, the ANCOVA model will correspond closely to the estimator $\hat{\theta}_\rho$ as described at the end of section 3.4. We will thus base our design on the assumption that the least squares estimate

$$\hat{\theta} \doteq \hat{\theta}_c = (\overline{X}_{16} - \rho_1 \overline{X}_{10}) - (\overline{X}_{06} - \rho \overline{X}_{00}).$$

In designing this RCT, we would regard that

- We decide to use 1:1 randomization, so the sample sizes are the same on the experimental treatment and on placebo: $n_1 = n_0$.
- The measurement made on the ith subject is the change in SBP measurements during the six months after randomization: $Y_{ki} = X_{ki6} - \rho_k X_{ki0}$.
- The treatment effect is defined as the difference between the study arms (treatment minus control) in mean change in SBP over the 6 months post randomization: $\theta = \mu_{16} - \mu_{06}$.
- The variability of the clinical measurement can be calculated from the hypothesized variability of measurements made at a single point in time and the hypothesized correlation between measurements made on the same individual 6 months apart: $\sigma_k^2 = \tau_{k6}^2 + \rho_k^2 \tau_{k0}^2 - 2\rho_k^2 \tau_{k6} \tau_{k0}$. Because we hypothesize that $\tau_{k6}^2 = \tau_{k0}^2$, we have $\sigma_k^2 = \tau_{10}^2(1 - \rho_1^2) = 27.386^2 \times (1 - 0.4^2) = 630$, so $\sigma_1 = 25.10$.
- The null hypothesis of no treatment effect corresponds to a difference of 0 between the means: $\theta_0 = 0$ mmHg.
- The alternative hypothesis corresponds to a mean SBP on the experimental treatment arm that is 10 mm Hg lower than what it would be on the placebo arm: $\theta_1 = -10$ mmHg.

The following R code would produce a clinical trial design meeting these criteria:

```
> dsn5 <- seqDesign(
+         prob.model= "mean",
+         arms= 2,
+         ratio= c(1,1),
+         null.hypothesis= c(163, 163),
+         alt.hypothesis= c(153, 163),
+         sd= c(25.1, 25.1),
+         test.type= "less",
+         size= 0.025,
+         power= 0.90,
+         nbr.analyses= 4,
+         design.family= "X",
+         P= 1
+         )
> dsn5
```

```
Call:
seqDesign(prob.model = "mean", arms = 2, null.hypothesis = c(163,
    163), alt.hypothesis = c(153, 163), sd = c(25.1, 25.1), ratio = c(1,
    1), nbr.analyses = 4, test.type = "less", size = 0.025, power = 0.9,
    design.family = "X", P = 1)

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=  69.05) -24.2030  12.1015
    Time 2 (N= 138.11) -12.1015   0.0000
    Time 3 (N= 207.16)  -8.0677  -4.0338
    Time 4 (N= 276.22)  -6.0508  -6.0508
```

Because

- we obtain the same estimator of treatment effect if we analyze the mean change in SBP over six months adjusted for baseline as we would obtain if we analyze the mean SBP at 6 months adjusted for baseline (only the estimate of β_1 would change, and that difference would be 1.0),
- the mean probability model is the default probability model,
- a two arm study is the default,
- 1:1 randomization is the default,
- we do not really need to know the means for each study arm, only the difference (note that the hypotheses printed in the above output just provide the difference between the means),
- in a two arm study, the default null hypothesis is a difference of 0 across study arms (or a mean of 0 on each arm, which corresponds to a difference of 0),
- the default assumes the same variability (standard deviation) on both arms,
- a one-sided test type is the default and the lesser hypothesis can be inferred from the values of the null and alternative hypotheses,
- the default type I error in a one-sided hypothesis test is 0.025,
- the family of designs on the estimate (sample mean) scale (the unified family) is the default,
- the O'Brien-Fleming boundary relationship (corresponding to `P=1`) is the default boundary relationship, and
- argument names can be abbreviated to the length adequate to distinguish them from all other arguments,

the same design could have been obtained by the following shorter code:

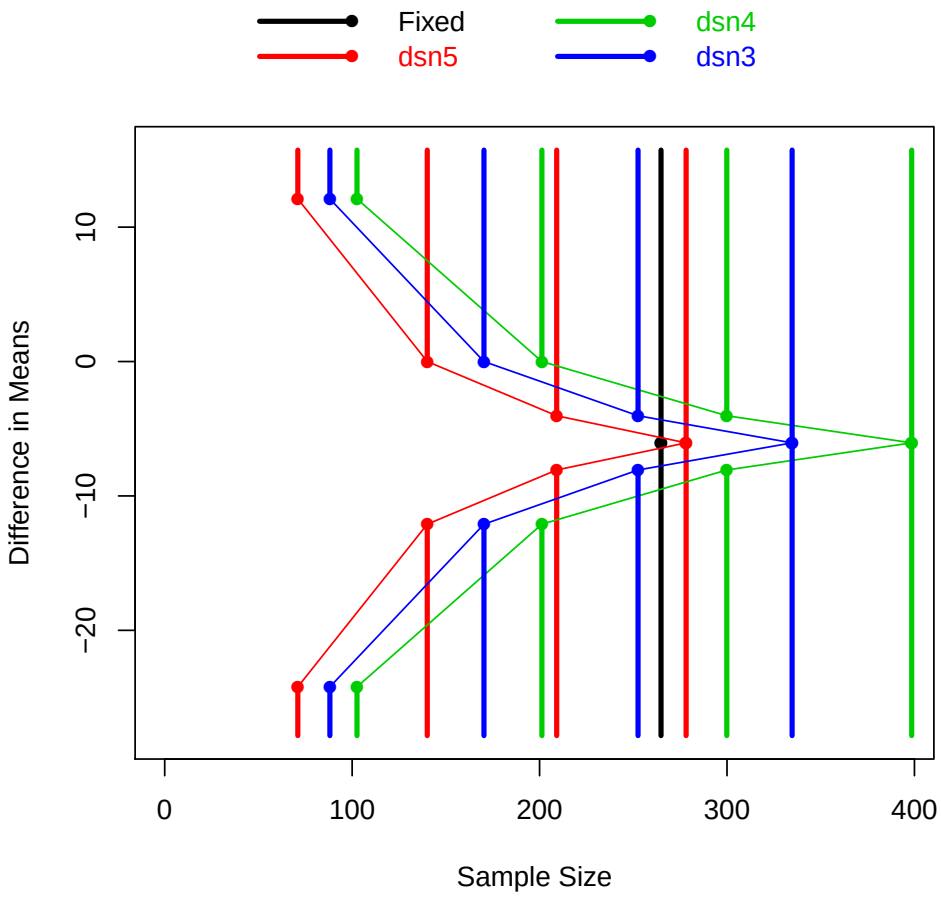
```
> dsn5 <- seqDesign( alt= -10, sd= 25.1, power= 0.90, nbr= 4)
> dsn5

Call:
seqDesign(alt.hypothesis = -10, sd = 25.1, nbr.analyses = 4,
    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=  69.05) -24.2030  12.1015
    Time 2 (N= 138.11) -12.1015   0.0000
    Time 3 (N= 207.16)  -8.0677  -4.0338
    Time 4 (N= 276.22)  -6.0508  -6.0508
```

From the above printout of the study design, we see

- The study would require a maximum sample size of 276.22 on the two treatment arms combined. (The fractional sample size is sometimes relevant in "information" scale monitoring, but it is clearly not of interest in this case. In real life, we would likely round the sample size up to 278– 139 on each arm.)
- At a first analysis that occurred when 25% of the maximal sample size had accrued, an observed difference between study arms in average SBP (experimental treatment average minus control average) of -24.203 mmHg or less would correspond to a recommendation that the study be stopped and the null hypothesis be rejected. An observed difference between study arms in average SBP of 12.1015 mmHg or greater would correspond to a recommendation that the study be stopped with a decision not to reject the null hypothesis. If the observed average SBP is between those two values, the recommendation would be to continue the trial to accrue another 25% of the planned maximal sample size.
- Similar interpretations apply to the stopping boundaries at the second and third planned analyses.
- According to these boundaries, the trial would only continue to the maximal sample size only if the observed difference in mean SBP were between -8.0677 mmHg and -4.0338 mmHg at the third analysis.
- If the trial continued to the last analysis, an observed difference in average SBP below -3.0508 mmHg would correspond to a decision to reject the null hypothesis. Otherwise, we would not reject the null hypothesis.

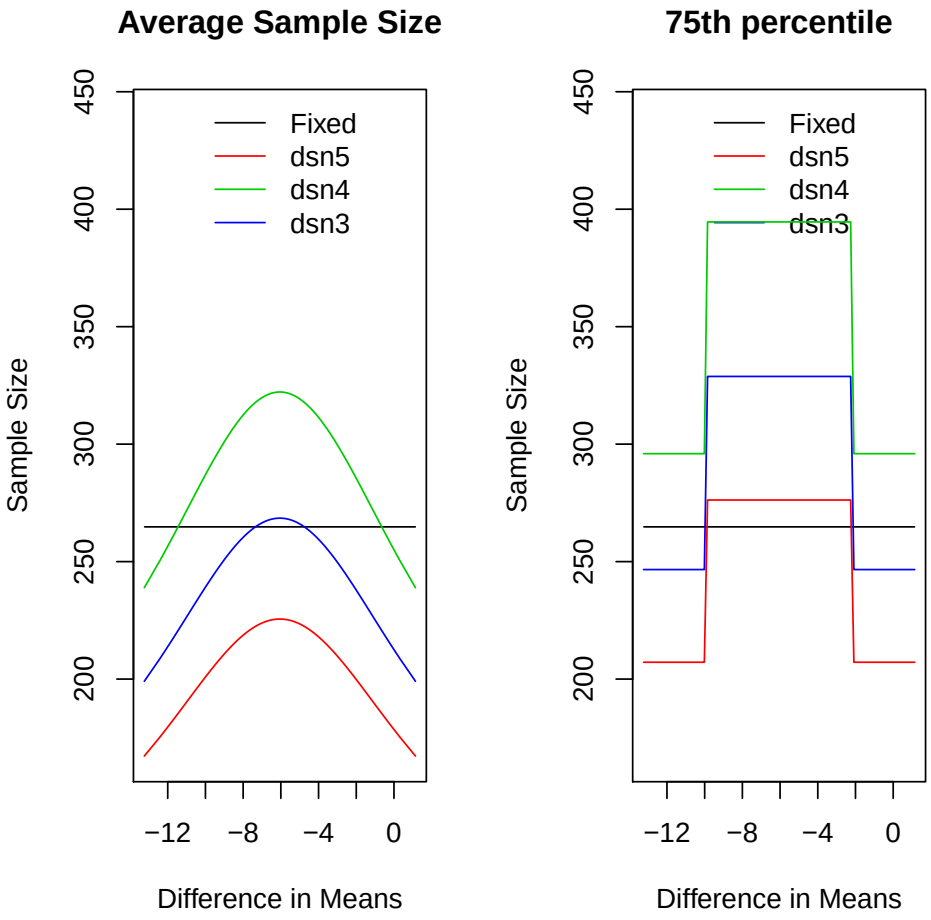
We can plot the difference in the boundaries obtained for all three of the two study arm strategies. In this plot, the x axis is the sample size, hence we can see the difference in maximum sample size requirements under the alternative analysis strategies.

```
> plot(dsn5,dsn4,dsn3)
```



Similarly we can plot the ASN (average sample N) curves for the three designs, as well as a fixed sample design that has the same power to detect the design alternative. Easily seen is the marked efficiency provided by the ANCOVA model.

```
> seqPlotASN(dsn5,dsn4,dsn3)
```



3.6 Crossover Study: Difference Between Study Periods in Mean SBP 6 Months after Start of Treatment

We can also consider a RCT design in which each subject truly serves as his/her own control: a randomized crossover study. The actual analysis of such a study must consider the possiibility of period effects and carryover effects. But for the purposes of planning the study, we can think about it in the following simplified (simplistic?) fashion that assumes no carryover effects exist.

In such a study, subjects are randomized in a 1:1 ratio to one of two strategies:

- A group that will receive the experimental treatment for 6 months, then undergo a "washout period" of, say, 6 months, and then receive placebo for 6 months. In keeping with the hypotheses that we have used in other settings, for this group we might presume
  - The SBP measurements at the time of accrual average 160 mmHg.
  - The SBP measurements at the end of the first 6 month accrual period (during which this group of patients was receiving the experimental treatment) average 163 mmHg in the absence of a treatment effect, and average 153 mm Hg if the treatment works as well as hoped.

  - During the washout period, any treatment effect is erased, but any time trend in measurements due to aging, progression of disease, etc. persist. Hence, we might imagine that the SBP measurements at the end of the washout period average 166 mmHg.
  - The SBP measurements at the end of the second 6 month treatment period (during which time this group of patients was receiving placebo) average 169 mmHg (reflecting trends over time due to aging, progression of disease, etc.).

- A group that will receive the placebo for 6 months, then undergo a "washout period" of, say, 6 months, and then receive the experimental treatment for 6 months. For this group we might presume
  - The SBP measurements at the time of accrual average 160 mmHg.
  - The SBP measurements at the end of the first 6 month accrual period (during which this group of patients was receiving the placebo) average 163 mmHg.
  - During the washout period any time trend in measurements due to aging, progression of disease, etc. persists. Hence, we might imagine that the SBP measurements at the end of the washout period average 166 mmHg.
  - The SBP measurements at the end of the second 6 month treatment period (during which time this group of patients was receiving the experimental treatment) 169 mmHg in the absence of a treatment effect, and average 159 mm Hg if the treatment works as well as hoped.

We also have to consider the correlation between the four measurements made on each individual. Two patterns that might be considered include

- "Exchangeable", in which case each pair of observations made on the same subject have correlation ρ, or
- "Auto-regressive 1", in which case the correlation between two adjacent measurements 6 months apart is ρ, the correlation between two measurements made 12 months apart is ρ^2, and the correlation between two measurements made 18 months apart is ρ^3.

In analyzing these data, we have to decide how to use the baseline measurements at the start of each period. The issues are much the same as considered above for the ANCOVA model, though the existence of four measurements that might be used does make the model a bit more complex:

- Under the exchangeable correlation structure, there is little reason to consider the baseline measurements. We can again use estimators of treatment response for each individual that are of the form $Y_i = (X_{11i} - aX_{10i}) - (X_{01i} - aX_{00i})$ and search for the value of a that provides the smallest variance under the exchangeable correlation structure. That optimal value turns out to be $a = 0$, so the direct comparison of the measurements made at the end of each period is the best approach. The variability of the clinical measurement Y_i (which would equal the difference of the measurement made at the end of 6 months receiving the experimental therapy minus the measurement made at the end of 6 months receiving placebo) would be found to be $\sigma^2 = 2\tau^2(1-\rho)$. Using the presumed values, we would thus find $\sigma = 30$.
- Under the auto-regressive 1 correlation structure, there is some advantage in considering the baseline measurements at the start of each period. We can again use estimators of treatment response for each individual that are of the form $Y_i = (X_{11i} - aX_{10i}) - (X_{01i} - aX_{00i})$. In this case, the optimal choice is $a = \rho/2$. With that choice, we find $\sigma^2 = 2\tau^2(1-\rho^2)(1-a\rho+a^2) = 2\times 27.386^2 times 0.84\times 0.96 = 1209.59$, so $\sigma = 34.779$ mmHg.

We thus create clinical trial designs according to

- We essentially have a one arm study with clinical outcomes being the within subject contrasts.
- The measurement made on the ith subject is the difference between adjusted 6 month change in measurements made on the experimental treatment minus the adjusted 6 month change in measurements made on the placebo arm: $Y_i = (X_{1i6} - aX_{1i0}) - (X_{0i6} - aX_{0i0})$, where $a = 0$ if we presume exchangeability and $a = \rho/2 = 0.2$ if we presume auto-regressive 1 correlations structure.
- The treatment effect is defined as the difference between the study periods (treatment minus control) in mean change in SBP over the 6 months post randomization: $\theta = \mu_{16} - \mu_{06}$.
- The variability of the clinical measurement can be calculated from the hypothesized structure for the variance-covariance matrix as discussed above.
- The null hypothesis of no treatment effect corresponds to a difference of 0 between the means: $\theta_0 = 0$ mmHg.
- The alternative hypothesis corresponds to a mean SBP on the experimental treatment arm that is 10 mm Hg lower than what it would be on the placebo arm: $\theta_1 = -10$ mmHg.

The following R code would produce a clinical trial design meeting these criteria:

```
> dsn6exch <- seqDesign( arms= 1, alt.= -10, sd= 30, power= 0.90, nbr= 4)
> dsn6exch

Call:
seqDesign(arms = 1, alt.hypothesis = -10, sd = 30, nbr.analyses = 4,
    power = 0.9)

PROBABILITY MODEL and HYPOTHESES:
   Theta is mean response
   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= 24.66) -24.2030  12.1015
    Time 2 (N= 49.32) -12.1015   0.0000
    Time 3 (N= 73.99)  -8.0677  -4.0338
    Time 4 (N= 98.65)  -6.0508  -6.0508

> dsn6ar1 <- seqDesign( arms= 1, alt.= -10, sd= 34.779, power= 0.90, nbr= 4)
> dsn6ar1

Call:
seqDesign(arms = 1, alt.hypothesis = -10, sd = 34.779, nbr.analyses = 4,
    power = 0.9)

PROBABILITY MODEL and HYPOTHESES:
   Theta is mean response
   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=  33.14) -24.2030  12.1015
    Time 2 (N=  66.29) -12.1015   0.0000
    Time 3 (N=  99.43)  -8.0677  -4.0338
    Time 4 (N= 132.58)  -6.0508  -6.0508
```

We note that crossover designs are not that widely used in phase III clinical trials, despite the clear cost savings in terms of the number of subjects. In the hypothetical setting we consider here, the calendar time could have been prolonged (depending on the accrual time). But more importantly, it is difficult to be sure that any washout period is sufficiently long to guarantee no carryover effects of a previous treatment on efficacy and safety endpoints.

4 Specification of the Test in Simulations: `var.equal` and `var.true`

As noted above in section 1, RCTdesign by default presumes the use of the t test in one arm settings and the use of the t test that allows for the possibility of unequal variances in two arm settings. However, a user can specify the use of the t test that presumes equal variances or the Z test using known variances when performing simulations. (The Z test is primarily of interest for validation of the numerical routines, as usual standards for analysis would dictate a t test.)

We illustrate the impact of these choices using RCTdesign's facility for simulating operating characteristics. Our interest here is just demonstrating the differences between the versions of the t test and the Z test, hence we do not demonstrate all functionality of the simulation routines here.

In our examples we will use a two arm study in which we have purposely set the sample size to be quite low: we want to illustrate the differences between the choice of tests, and if our sample sizes are moderate to large, asymptotic theory will often dictate that all of our examples below would have given similar results. Hence we create a design that will result in a small sample size. We do this by updating the design from section 3.3 to detect a larger alternative.

```
> dsn7 <- update(dsn3, alt= -25)
> dsn7

Call:
seqDesign(alt.hypothesis = -25, sd = 27.386, nbr.analyses = 4,
    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 <= -25    (power = 0.900)
   (Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                      Efficacy Futility
```

```
    Time 1 (N= 13.15) -60.5076  30.2538
    Time 2 (N= 26.31) -30.2538   0.0000
    Time 3 (N= 39.46) -20.1692 -10.0846
    Time 4 (N= 52.61) -15.1269 -15.1269
```

If we want to examine operating characteristics such as power and ASN when computed using numerical integration, we could just execute `seqPlotPlower(dsn7)` and `seqPlotASN(dsn7)`. However, as our goal is to also plot simulated operating characteristics, we first must create a `"seqOC"` object containing the simulation results. We do this by executing the following R code, which will by default consider 50 different values of `theta` and will perform the requested 1,000 simulations for each one. By default, the data are generated from a normal distribution. (We "set the seed" for the random number generator in order that we will be able to reproduce the results later.)

```
> dsn7OC <- seqOC( dsn7, Nsimul=1000, seed=0 )
```

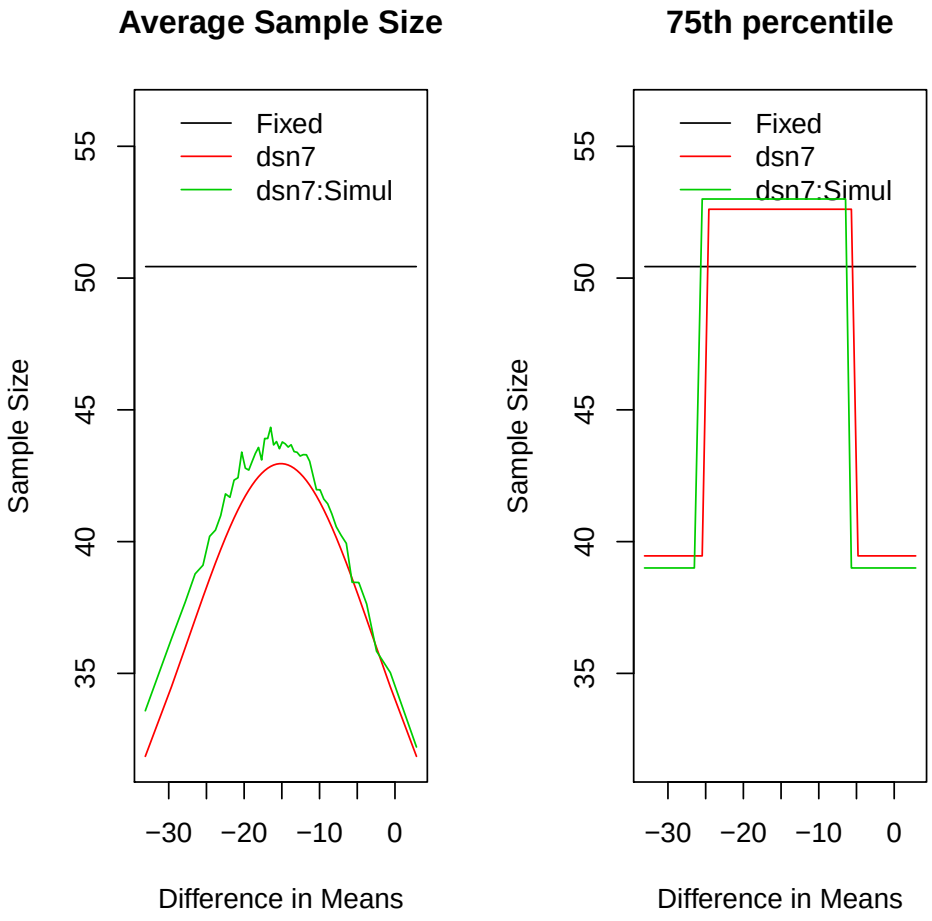
Though suppressed from this printout, warnings would be displayed about rounded sample sizes. The `"seqDesign"` object specified analyses performed after 13.15, 26.31, 39.46, and 52.61 subjects' data were available. Clearly these fractional subjects are impossible in this case, so `rSeqMean2()` (which is called by `seqOC()` through `rSeq()`) will round the sample sizes:

```
> seqExtract(dsn7OC$AsympOC,"design")

Call:
seqDesign(alt.hypothesis = -25, sd = 27.386, nbr.analyses = 4,
    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 <= -25    (power = 0.900)
   (Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                     Efficacy Futility
    Time 1 (N= 13.15) -60.5076  30.2538
    Time 2 (N= 26.31) -30.2538   0.0000
    Time 3 (N= 39.46) -20.1692 -10.0846
    Time 4 (N= 52.61) -15.1269 -15.1269

> seqExtract(dsn7OC$SimulOC,"design")

Call:
seqDesign(alt.hypothesis = -25, sd = 27.386, nbr.analyses = 4,
    sample.size = Nj, 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  = 0.0250)
```

```
    Alternative hypothesis : Theta <= -25    (power = 0.9025)
   (Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                    Efficacy Futility
    Time 1 (N= 13) -61.3987  31.2786
    Time 2 (N= 26) -30.6993   0.5792
    Time 3 (N= 39) -20.4662  -9.6539
    Time 4 (N= 53) -15.0601 -15.0601
```

We can now plot the ASN curves for both the numerically integrated and simulated results.

```
> seqPlotASN(dsn7OC)
```



We note that the agreement between the ASN curves is not particularly good, especially under the more extreme alternative hypotheses. There are multiple interacting factors that could potentially contribute to poor agreement between simulated results and the numerically integrated results in RCTdesign:

1. *Shape of Data Distribution*: In small samples, the statistics assumed by the probability model and

implemented by `rSeq()` are exact only for normally distributed data. Use with other distributions requires moderate to large sample sizes before the asymptotic results provide good approximations.

2. *Mean-Variance Relationships*: The mean probability model presumes that the variance of the data is unaffected by the treatment effect. In many distributions, however, differences in means will generally lead to differences in variance. This will typically not matter under the null hypothesis, because we usually base our test statistics on variability under the null hypothesis. Furthermore, in two arm studies, the null hypothesis is most often one of equality of means across study arms. However, a mean-variance relationship can greatly impact the true power of the test relative to that estimated by the RCTdesign numerical integration routines.

3. *Choice of test statistic*: The Z tests assumed by the numerical routines' probability model (and implemented by `rSeq()` when `var.true=TRUE`) are based on using the known variance. The t tests allow for the use of the sample variance as an estimate of the true variance. When using the t test that presumes equal variances (implemented by `rSeq()` when `var.equal=TRUE`), true differences between variances can lead to incorrect inference when sample sizes are not equal. The t test that allows for the possibility of unequal variances is valid for testing differences in means in large samples.

4. *Sample size*: Owing to the central limit theorem, in large samples this probability model will provide valid statistical hypothesis testing (i.e., correct type I error) for any distribution with finite variance, so long as the variances across groups is modeled appropriately. In small samples, the tests can be conservative or anti-conservative.

5. *Number of simulations*: Even when distributional theory holds exactly, a small number of simulations can mean that the simulated operating characteristics are not sufficiently precise.

6. *Lack of concordance between designs*: As noted above, simulations are conducted assuming that the number of observations is always an integer, while the numerical integrations allow for the "statistical information" at an analysis might be any real number.

Addressing the last point first, we modify our clinical trial design to be based on analyses at 14, 28, 42, and 56 subjects' data accrued (choosing these values to have balance between treatment groups at each interim analysis). In modifying the design, we can choose to continue with a design alternative of -25 mmHg (in which case the power will need to be calculated)

```
> dsn7a <- update(dsn7, sample.size=c(14, 28, 42, 56), power="calculate")
> dsn7a

Call:
seqDesign(alt.hypothesis = -25, sd = 27.386, nbr.analyses = 4,
    sample.size = c(14, 28, 42, 56), 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  = 0.0250)
    Alternative hypothesis : Theta <= -25    (power = 0.9168)
   (Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                   Efficacy Futility
    Time 1 (N= 14) -58.6483  29.3241
    Time 2 (N= 28) -29.3241   0.0000
```

```
    Time 3 (N= 42) -19.5494  -9.7747
    Time 4 (N= 56) -14.6621 -14.6621
```

or we can choose to continue to focus on 90% power (in which case the design alternative will need to be calculated)

```
> dsn7p <- update(dsn7, sample.size=c(14, 28, 42, 56), alt=, test.type="less")
> dsn7p

Call:
seqDesign(sd = 27.386, nbr.analyses = 4, sample.size = c(14,
    28, 42, 56), 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 <= -24.23    (power = 0.900)
   (Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                   Efficacy Futility
    Time 1 (N= 14) -58.6483  29.3241
    Time 2 (N= 28) -29.3241   0.0000
    Time 3 (N= 42) -19.5494  -9.7747
    Time 4 (N= 56) -14.6621 -14.6621
```

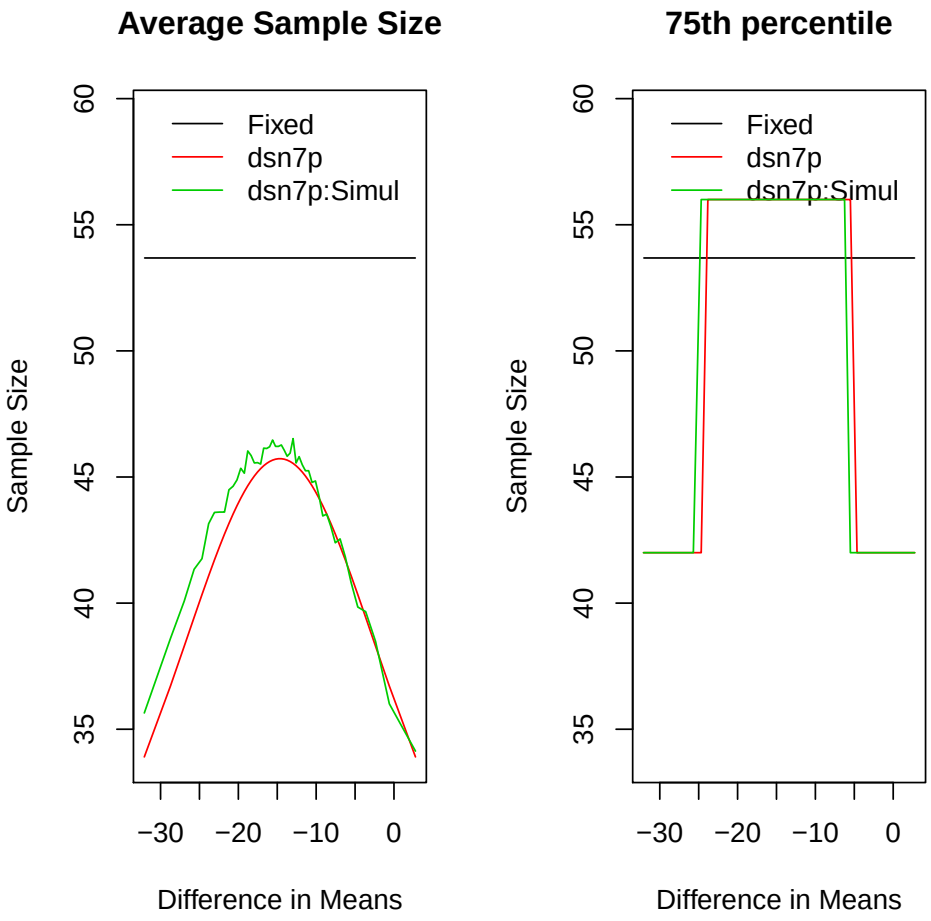
(Note that in each case we had to "undefine" either the alternative hypothesis or the power in order to get `seqDesign()` to use the sample size specification exactly. Otherwise, the `sample.size` argument would have only been used to define the relative spacing of analyses. When we removed the specification of the alternative hypothesis, we also had to specify `test.type="less"`, else the default value of a greater alternative hypothesis would have been used.)

These two designs are truly the same stopping rule. All that has differed is how the alternative hypothesis is defined. When we ask for the power of `dsn7p` under the alternative hypothesis of $\theta = -25$ mmHg, we get the same power as shown for `dsn7a`:

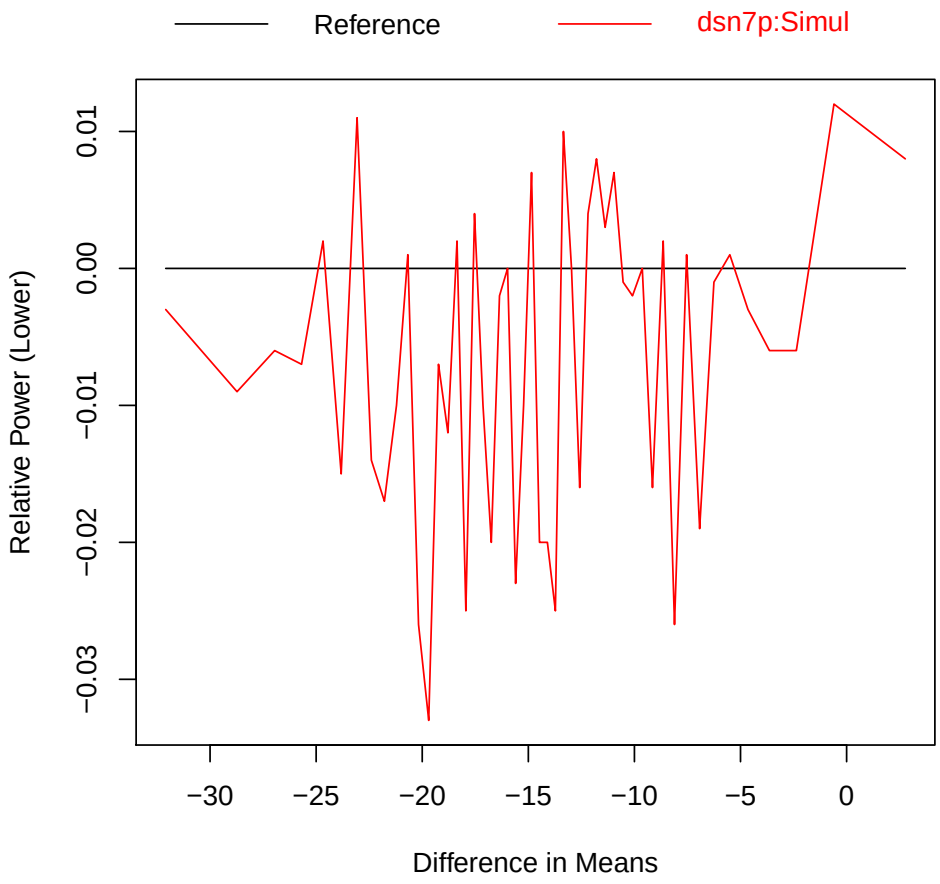
```
> seqOC(dsn7p , theta= -25)

### Asymptotic Operating Characteristics
Operating characteristics at theta= -25
ASN= 40.0497
Expected theta= -26.8971
Lower Power= 0.9168

Stopping Probabilities:
                 Lower Null  Upper  Total
Analysis time 1 0.0108    0 0.0001 0.0109
Analysis time 2 0.3276    0 0.0078 0.3354
Analysis time 3 0.4062    0 0.0298 0.4360
Analysis time 4 0.1724    0 0.0454 0.2178
```

We can now examine a more fair comparison of the numerically integrated and simulated operating characteristics.

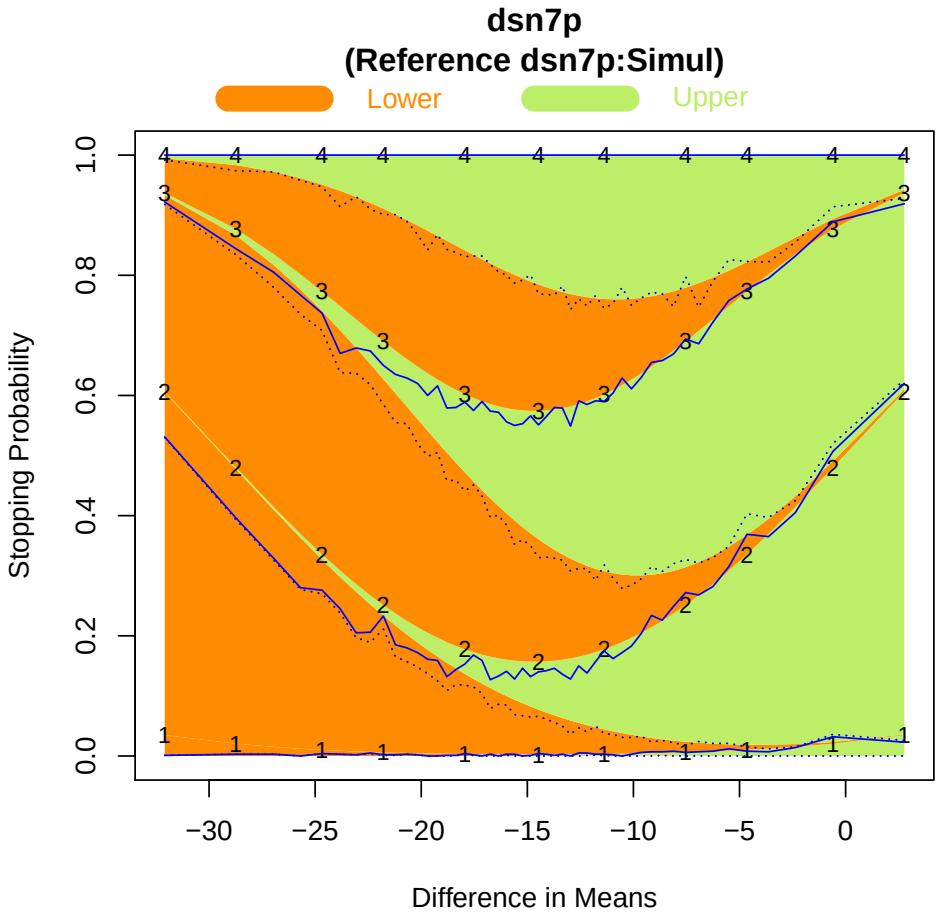
```
> dsn7pOC <- seqOC(dsn7p, Nsimul=1000, seed=0)
> seqPlotASN(dsn7pOC)
```




```
> seqPlotPower( dsn7pOC$SimulOC, reference=dsn7pOC$AsympOC, fixed=FALSE)
```




```
> seqPlotStopProb(dsn7pOC$AsympOC, reference=dsn7pOC$SimulOC)
```



In the following subsections, we can explore how better agreement between the numerically integrated and simulated results can be obtained.

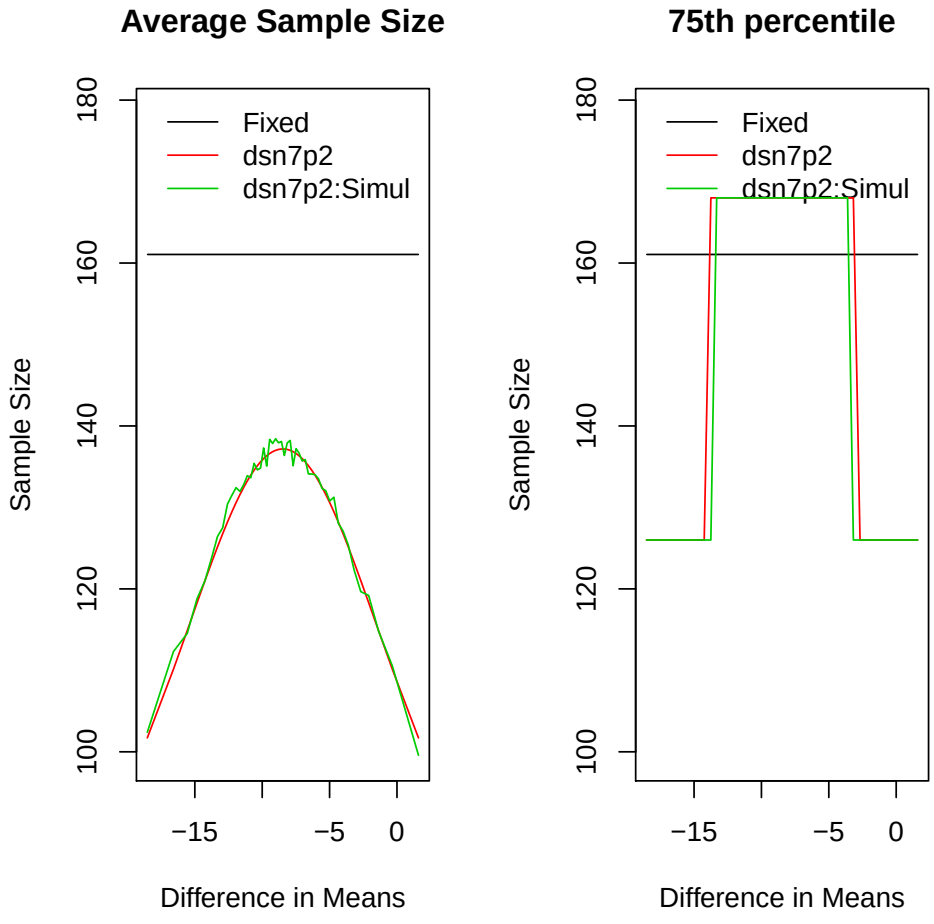
4.1 Effect of Sample Size on Agreement with Numerical Integration

As noted above, a part of the lack of agreement between the simulated results and the numerically integrated results can be due to the sample size not being large enough to show good agreement between the Z statistic distribution (used in the numerical integrations) and the t distribution (used in the simulations). (Note that we are using normally distributed simulated data by default, so the discordance between the t and Z distribution is the major problem.) We thus consider increasing the sample size threefold:

```
> dsn7p2 <- update(dsn7p, sample.size=3 * sampleSize(dsn7p))
> dsn7p2

Call:
seqDesign(sd = 27.386, nbr.analyses = 4, sample.size = 3 * sampleSize(dsn7p),
    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 <= -13.99    (power = 0.900)
   (Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
                    Efficacy Futility
    Time 1 (N=  42) -33.8606  16.9303
    Time 2 (N=  84) -16.9303   0.0000
    Time 3 (N= 126) -11.2869  -5.6434
    Time 4 (N= 168)  -8.4652  -8.4652
```

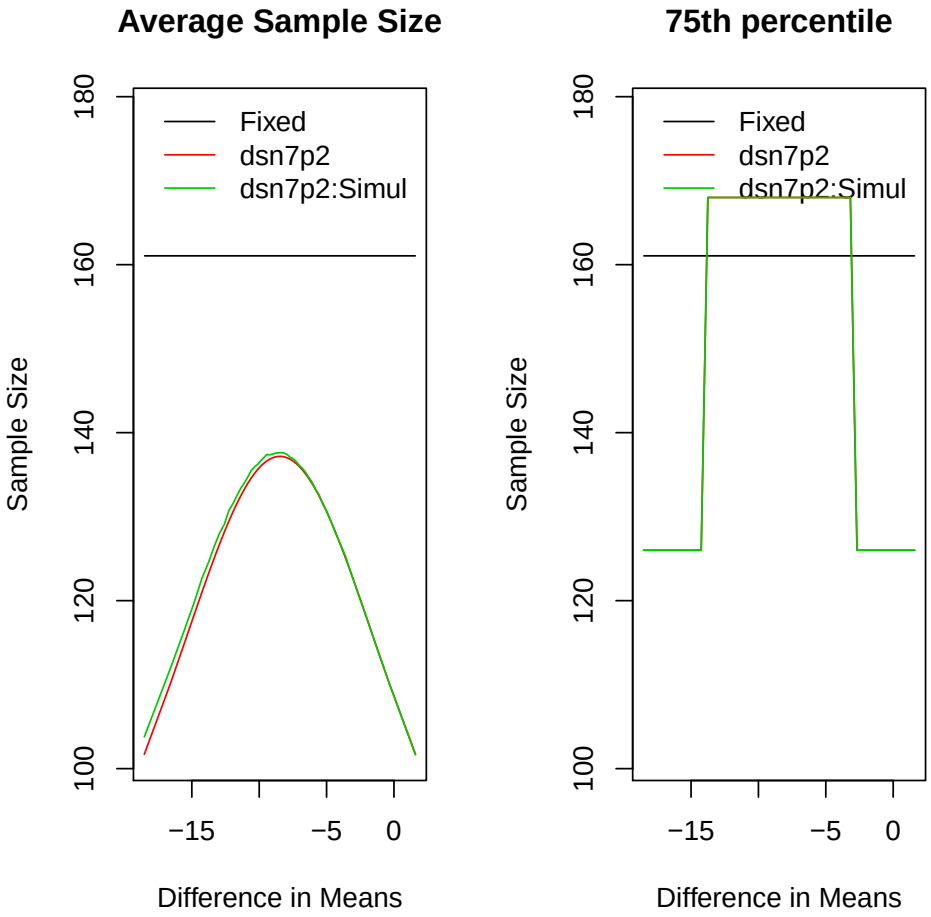
```
> dsn7p2OC <- seqOC(dsn7p2, Nsimul=1000, seed=0)
> seqPlotASN(dsn7p2OC)
```



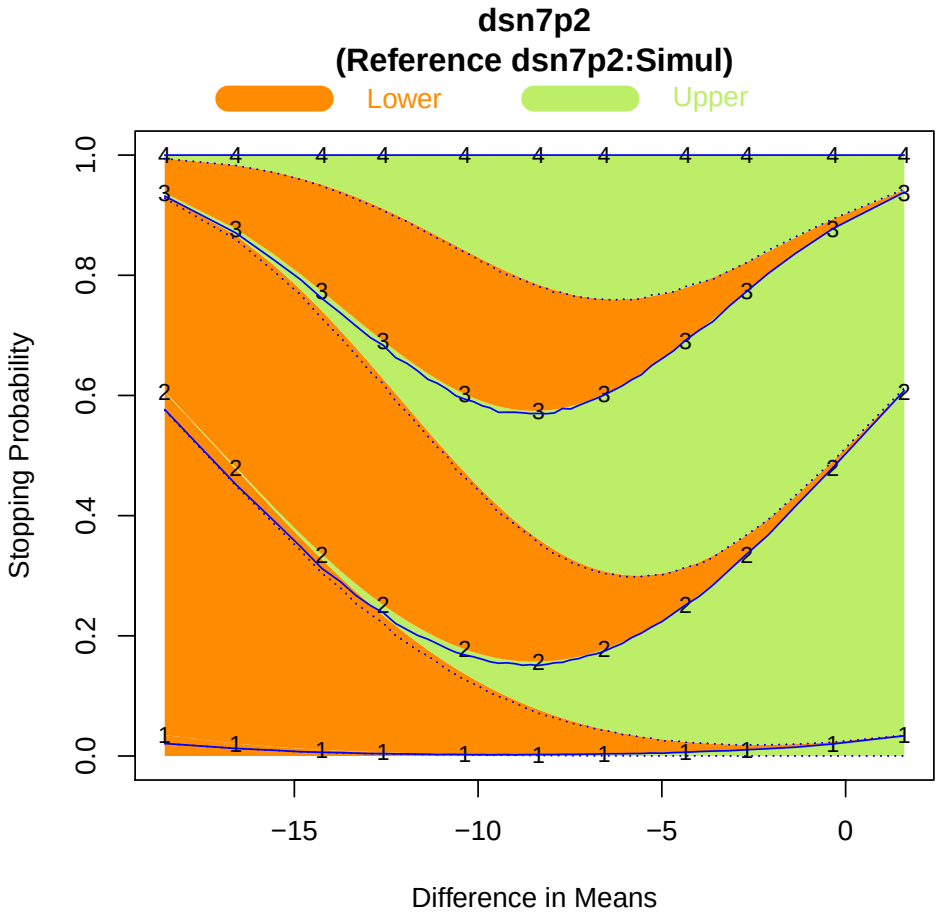
4.2 Effect of Number of Simulations on Agreement with Numerical Integration

The design with the increased sample size showed much better agreement with the numerically integrated results, however, the simulated ASN curve shows a lack of smoothness that is due to the low number of simulations. We can thus increase the number of simulations. This can be done by creating an entirely new `"seqOC"` object, or we can append new results to those simulations we have already performed (thereby saving some computational effort). We thus consider increasing the number of simulations to 100,000:

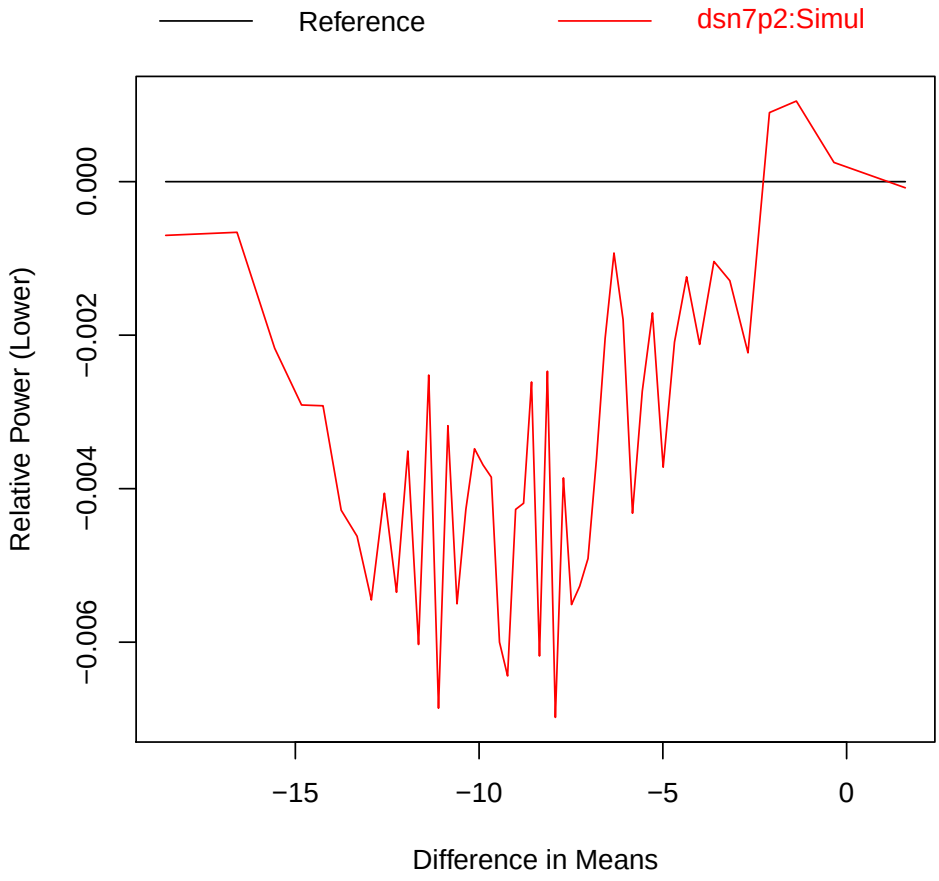
```
> dsn7p2OC <- seqOC(dsn7p2OC, Nsimul=99000, seed=1)
> seqPlotASN(dsn7p2OC)
```




```
> seqPlotStopProb( dsn7p2OC$AsympOC, reference=dsn7p2OC$SimulOC)
```




```
> seqPlotPower( dsn7p2OC$SimulOC, reference=dsn7p2OC$AsympOC, fixed=FALSE)
```



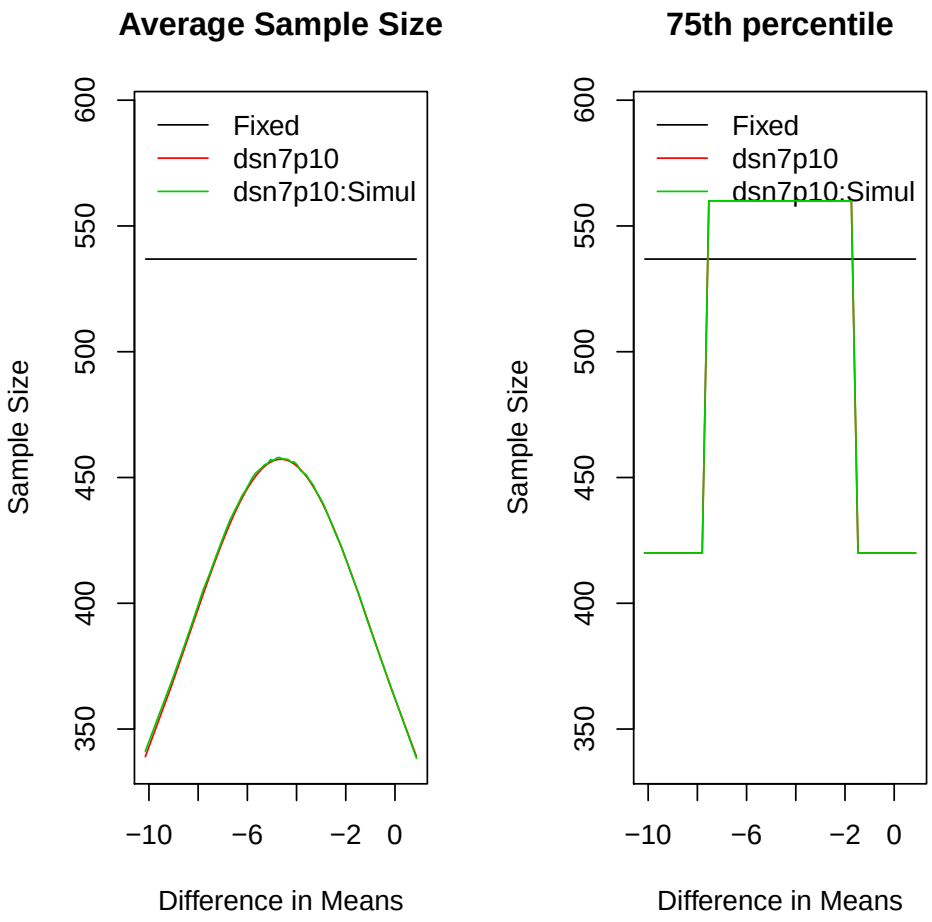
With this increased precision, we see that there is still some discrepancy between the simulated and numerically integrated operating characteristics with a sample size of 42 at the first of four analyses.

We can increase the sample size further and obtain good agreement.

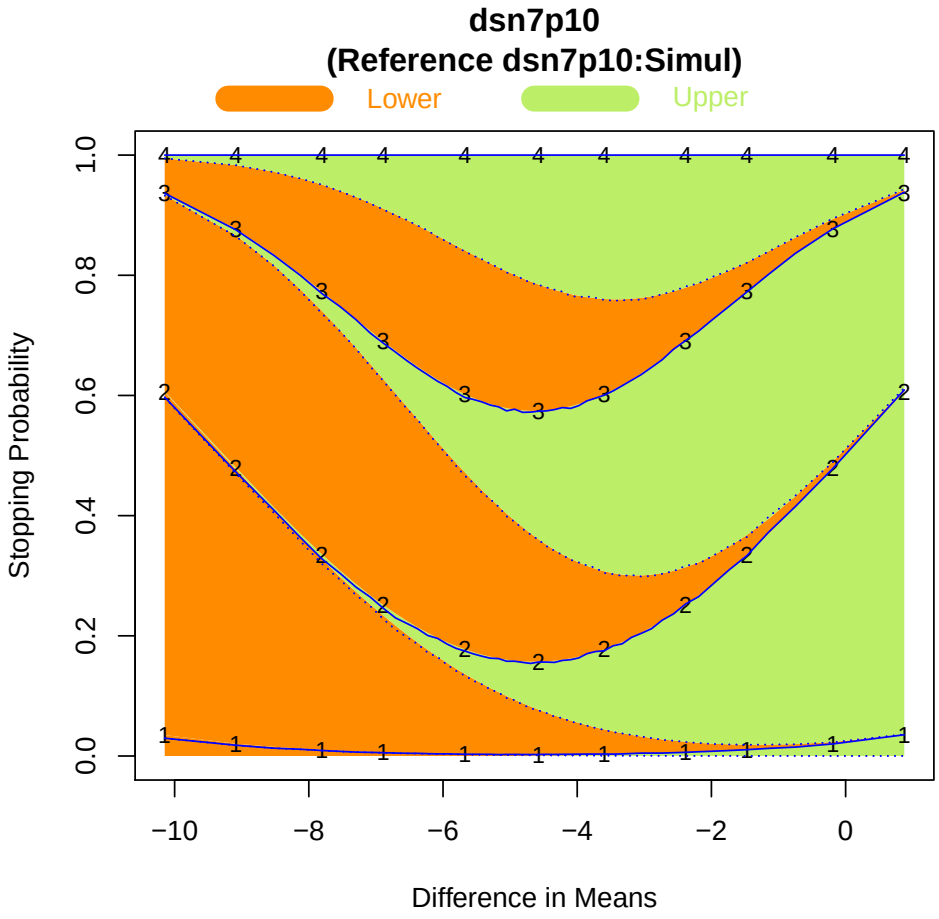
```
> dsn7p10 <- update(dsn7p, sample.size=10 * sampleSize(dsn7p))
> dsn7p10

Call:
seqDesign(sd = 27.386, nbr.analyses = 4, sample.size = 10 * sampleSize(dsn7p),
    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 >=  0.000    (size  = 0.025)
    Alternative hypothesis : Theta <= -7.663    (power = 0.900)
   (Emerson & Fleming (1989) symmetric test)

STOPPING BOUNDARIES: Sample Mean scale
```

```
                  Efficacy Futility
Time 1 (N= 140) -18.5462   9.2731
Time 2 (N= 280)  -9.2731   0.0000
Time 3 (N= 420)  -6.1821  -3.0910
Time 4 (N= 560)  -4.6366  -4.6366
```

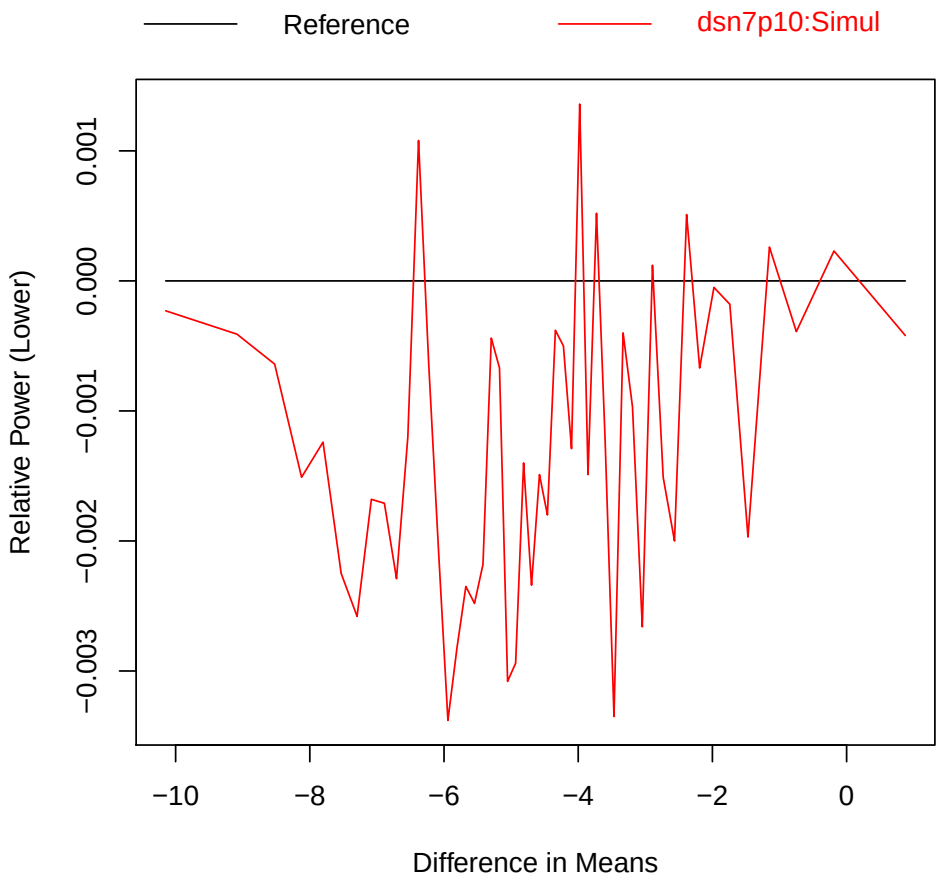
```
> dsn7p10OC <- seqOC(dsn7p10, Nsimul=100000, seed=0)
> seqPlotASN(dsn7p10OC)
```




```
> seqPlotStopProb( dsn7p10OC$AsympOC, reference=dsn7p10OC$SimulOC)
```



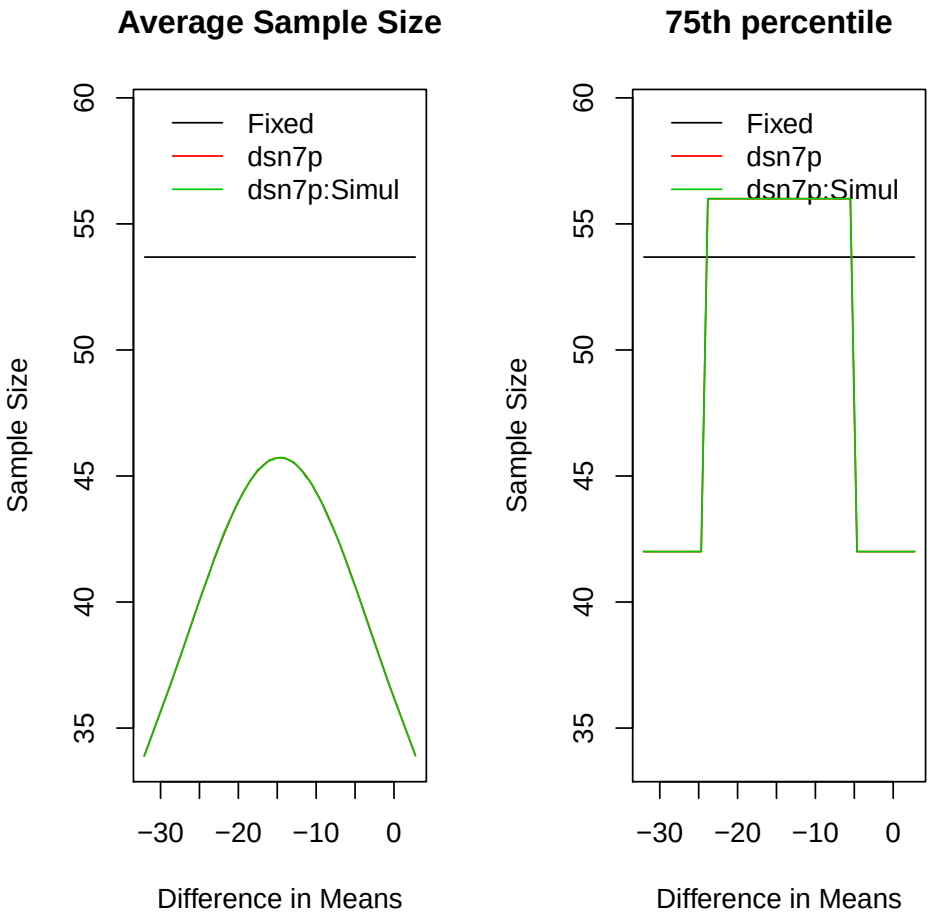

```
> seqPlotPower( dsn7p10OC$SimulOC, reference=dsn7p10OC$AsympOC, fixed=FALSE)
```



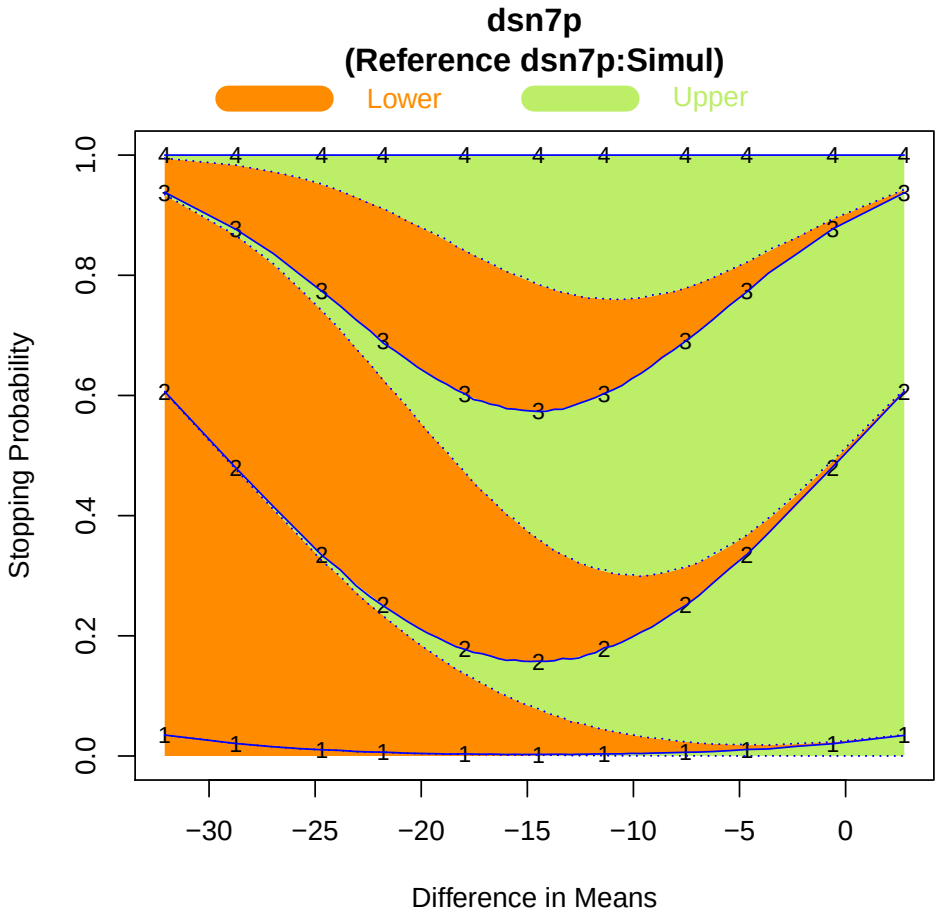
4.3 Effect of Choice of Statistic on Agreement with Numerical Integration

In the previous sections we attributed a part of the disagreement between simulations and the numerically integrated operating characteristics to the fact that at low sample sizes, there is still disagreement between the Z and t distributions. We can demonstrate this by using the original small sample size and specifying that the simulated RCTs should be analyzed with Z statistics using the known variances.

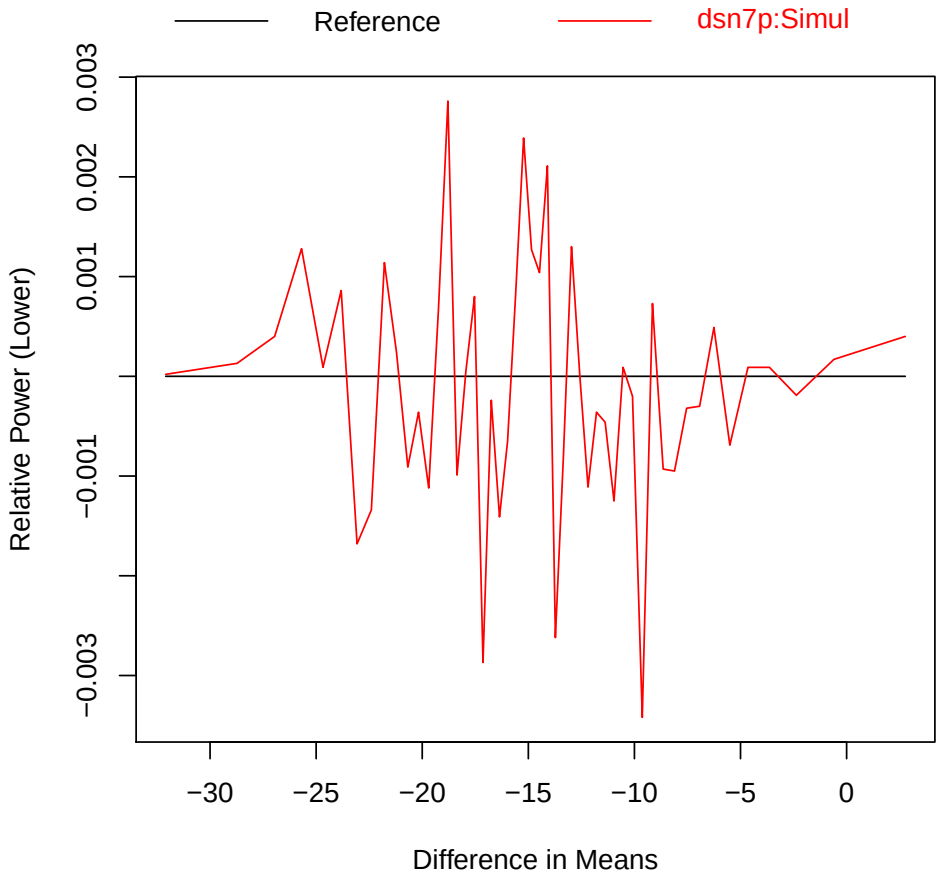
```
> dsn7pZOC <- seqOC(dsn7p, Nsimul=100000, seed=0, var.true=TRUE)
> seqPlotASN(dsn7pZOC)
```




```
> seqPlotStopProb( dsn7pZOC$AsympOC, reference=dsn7pZOC$SimulOC)
```




```
> seqPlotPower( dsn7pZOC$SimulOC, reference=dsn7pZOC$AsympOC, fixed=FALSE)
```



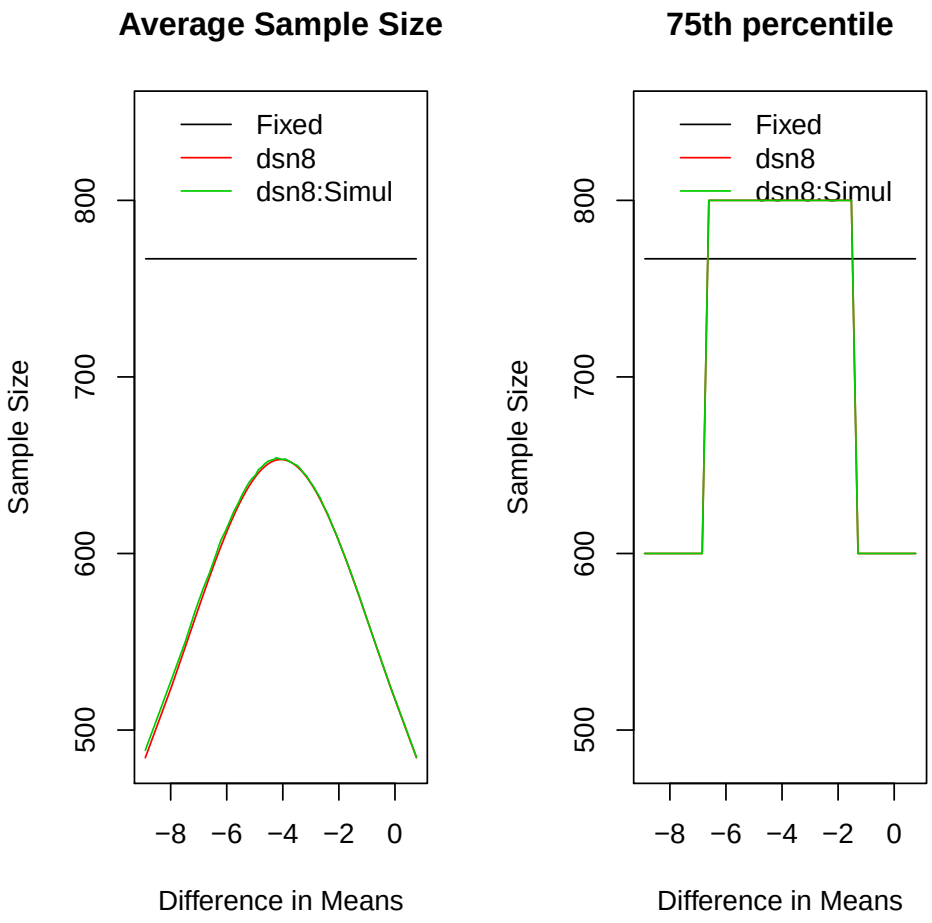
It is not standard to use the Z statistic when comparing means across study arms, because we rarely, if ever, know the true variance. On the other hand, when estimating variances there are two versions of the t statistic that are used: the t test that allows for the possibility of unequal variances (which is the default statistic in RCTdesign) and the t test that presumes the equality of variances (this can be specified using the argument `var.equal=TRUE` in `rSeq()` or `seqOC()`).

As described above in section 1, erroneously assuming equal variances poses no large problem when sample sizes are equal. However, if sample sizes are unequal and variances are unequal, the use of the t test that presumes equal variances can lead to either anti-conservative or conservative inference. To illustrate this point, we will use a RCT design with a large sample size, in order to eliminate the possibility that discrepancies between the simulated and numerically integrated operating characteristics were due to issues with small sample disagreements between the standard normal and t distributions.

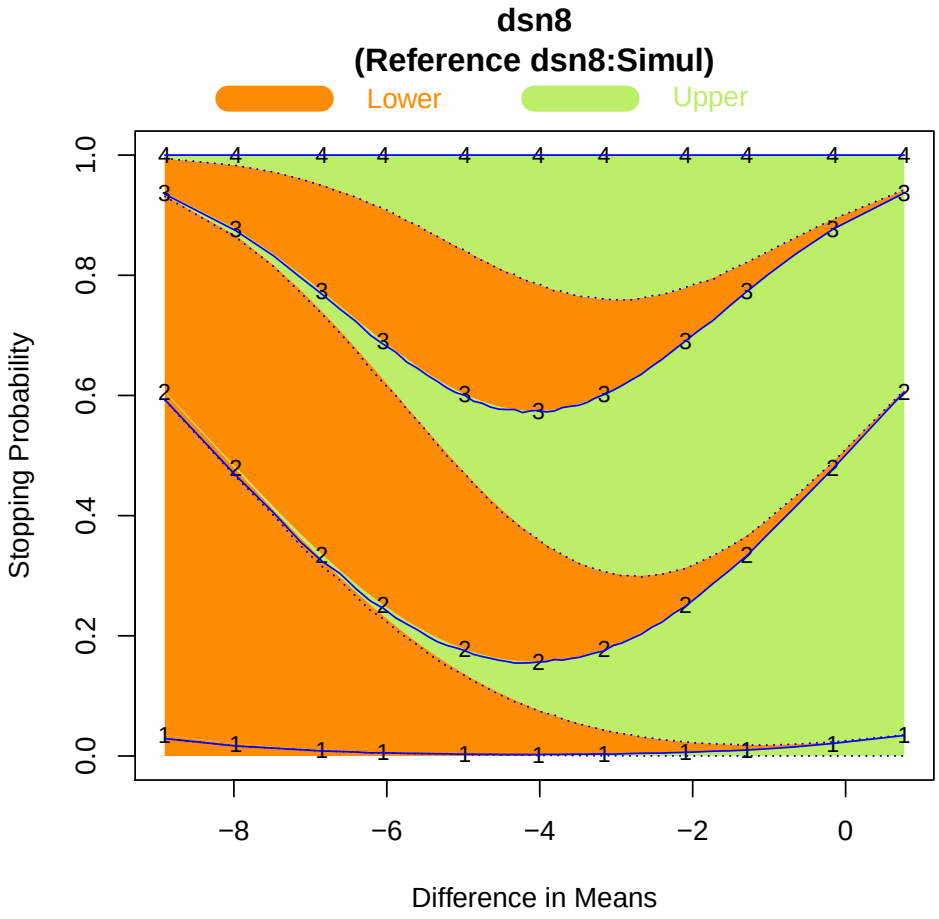
We use 2:1 randomization ratio and assume standard deviations are 20 and 30, respectively, on the experimental treatment and control arms. We first demonstrate appropriate type I error and when using the default t test that allows for the possibility of unequal variance.

```
> dsn8 <- seqDesign( ratio=2, sd= c(20,30), nbr=4, sample.size=800,
+         power= 0.90, test.type="less")
> dsn8OC <- seqOC(dsn8, Nsimul=100000, seed=0)
```

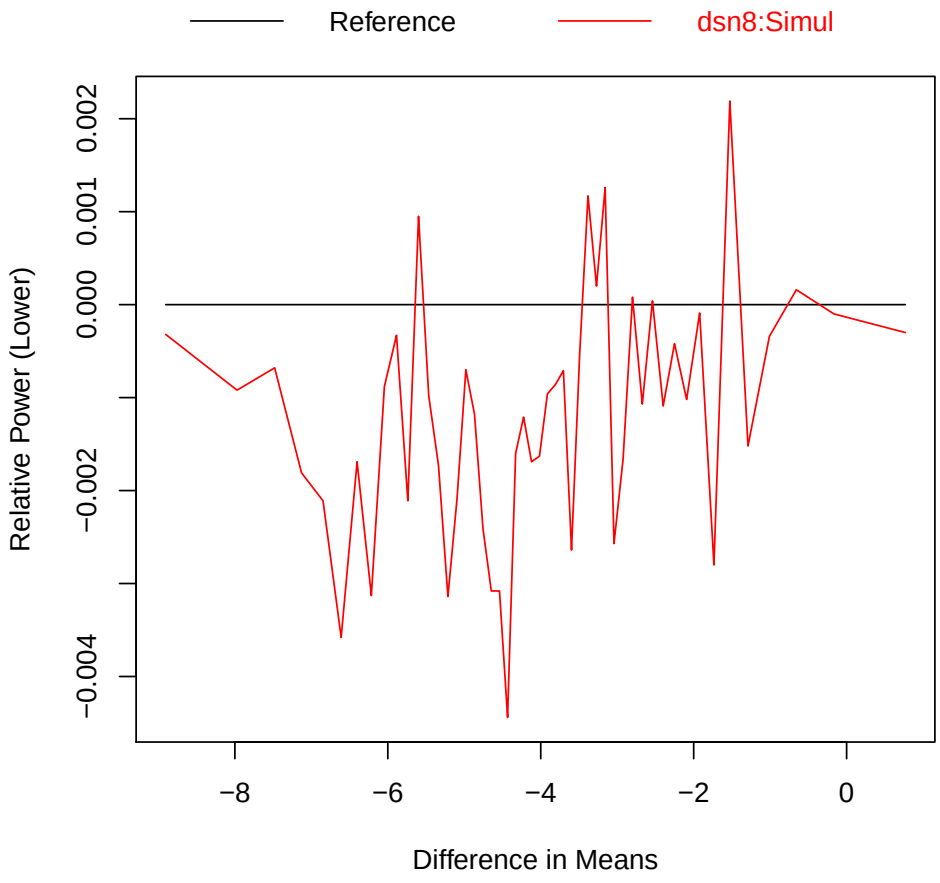
```
> seqPlotASN(dsn8OC)
```




```
> seqPlotStopProb( dsn8OC$AsympOC, reference=dsn8OC$SimulOC)
```

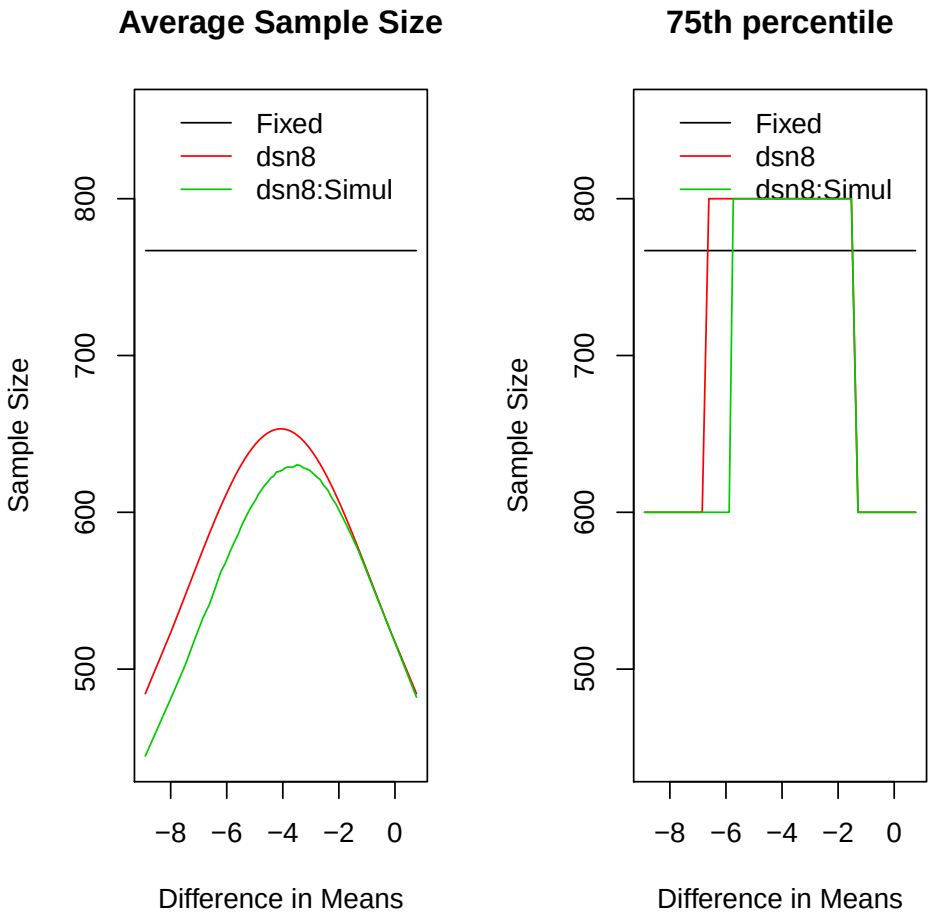



```
> seqPlotPower( dsn8OC$SimulOC, reference=dsn8OC$AsympOC, fixed=FALSE)
```

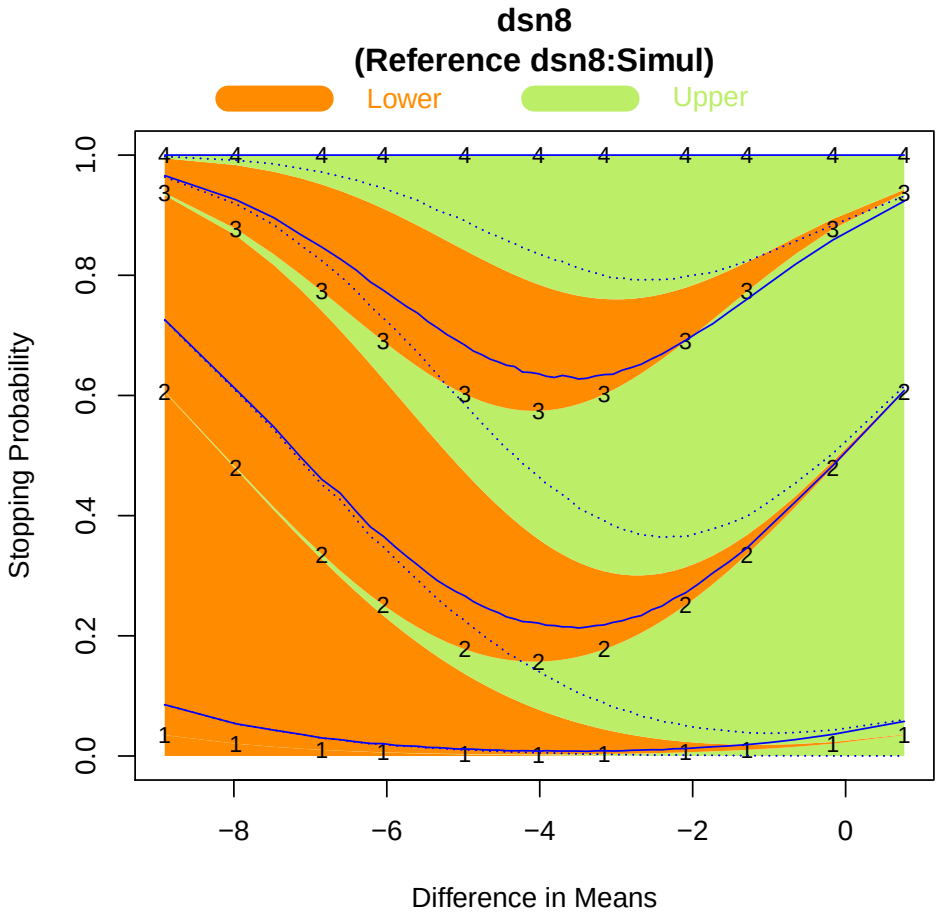


Now we use the t test that presumes equal variances. Given that the larger sample size will be on the arm having the smaller inference, we would expect anti-conservative inference: the type I error from the simulations is expected to be greater than the desired 0.025 level.

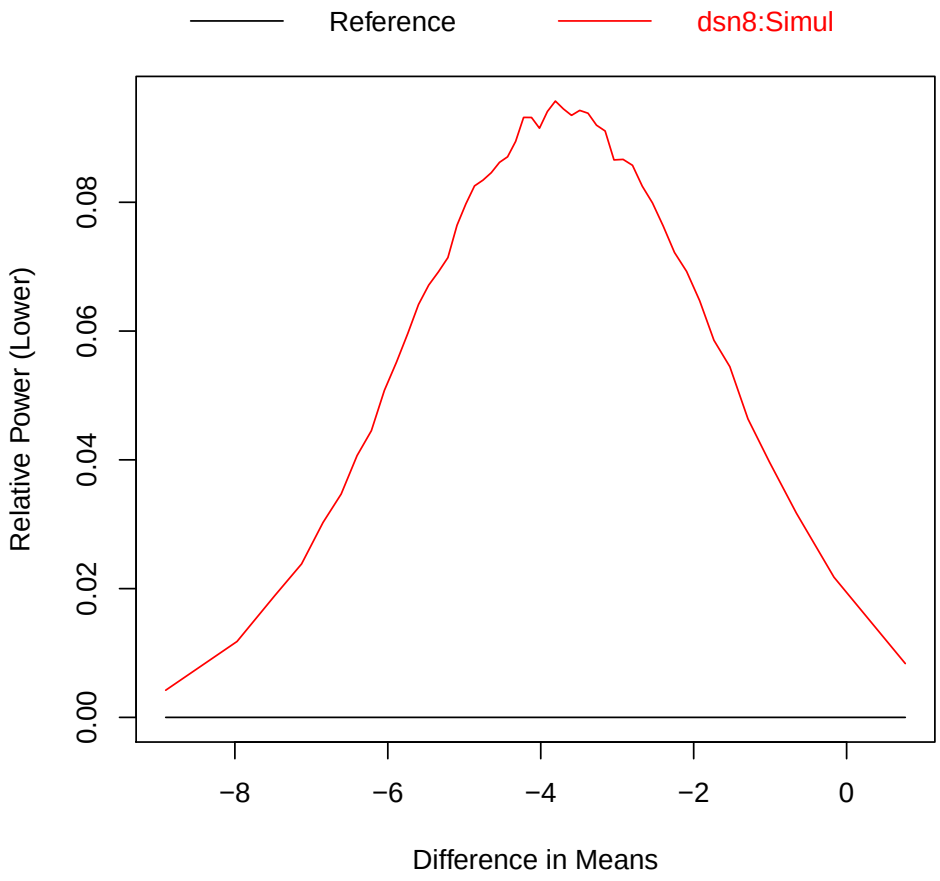
```
> dsn8OCeq <- seqOC(dsn8, Nsimul=100000, seed=0, var.equal=TRUE)
> seqPlotASN(dsn8OCeq)
```




```
> seqPlotStopProb( dsn80Ceq$AsympOC, reference=dsn80Ceq$SimulOC)
```

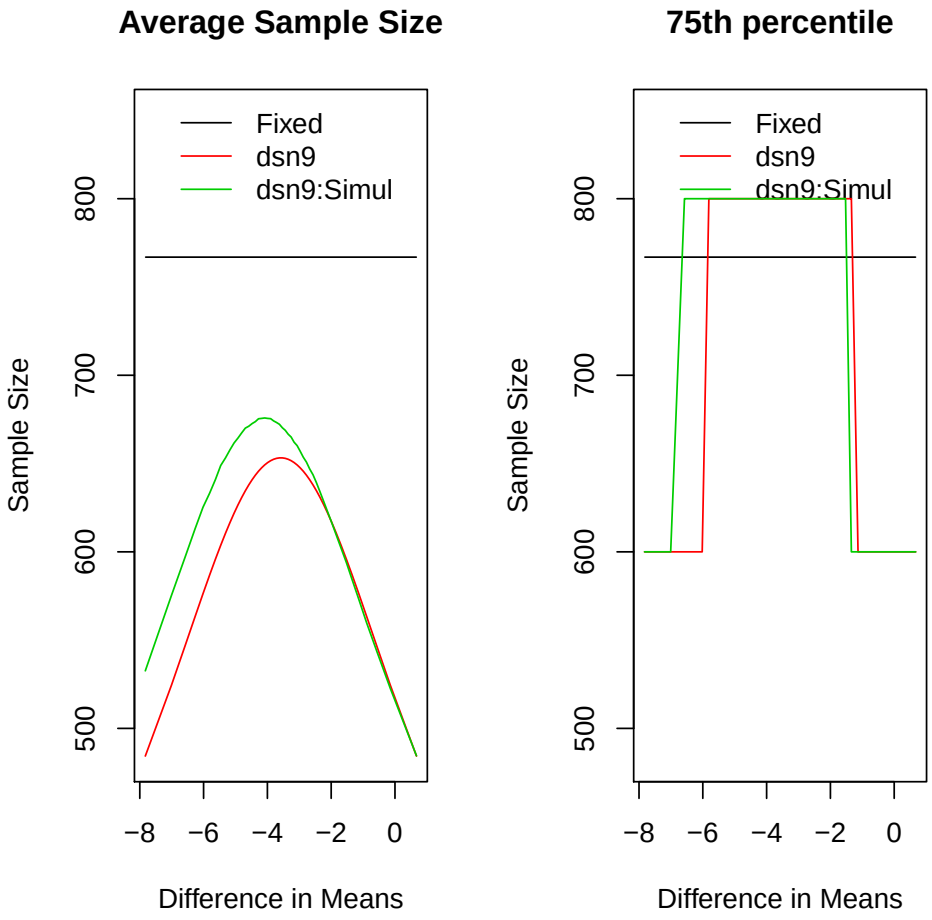



```
> seqPlotPower( dsn8OCeq$SimulOC, reference=dsn8OCeq$AsympOC, fixed=FALSE)
```

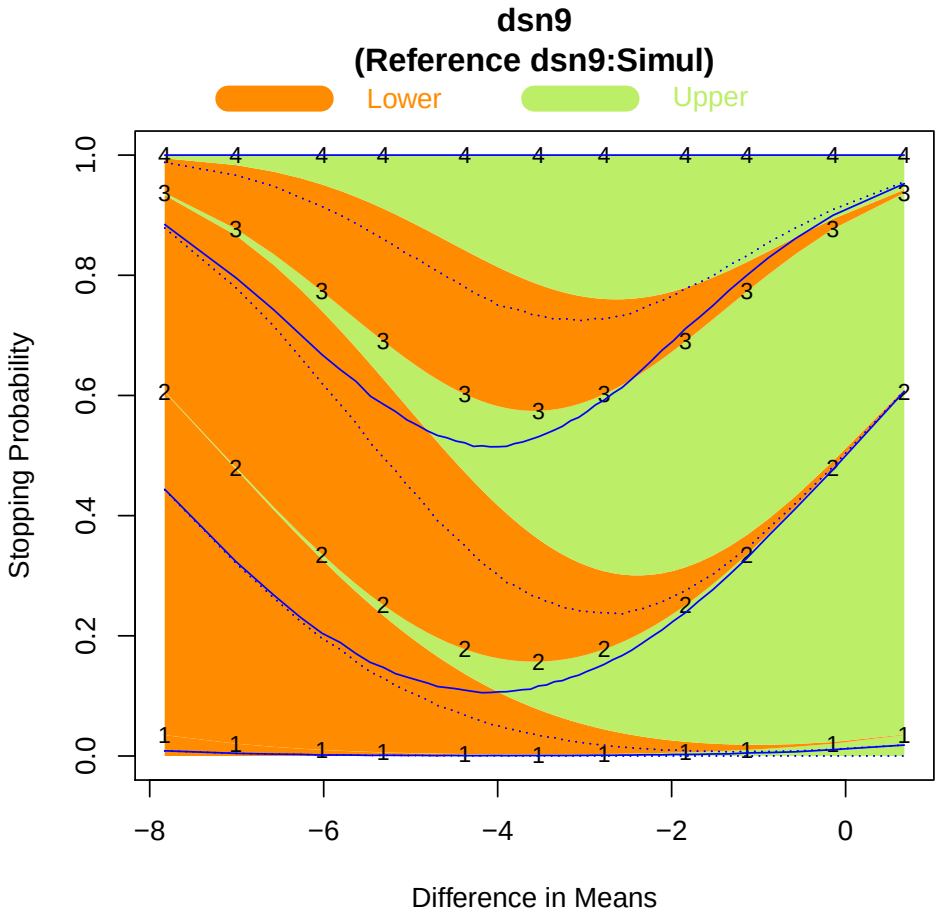


Using the t test that presumes equal variances when the larger sample size will be on the arm having the larger inference, will lead to conservative inference: the type I error from the simulations is expected to be less than the desired 0.025 level and the power will similarly be less.

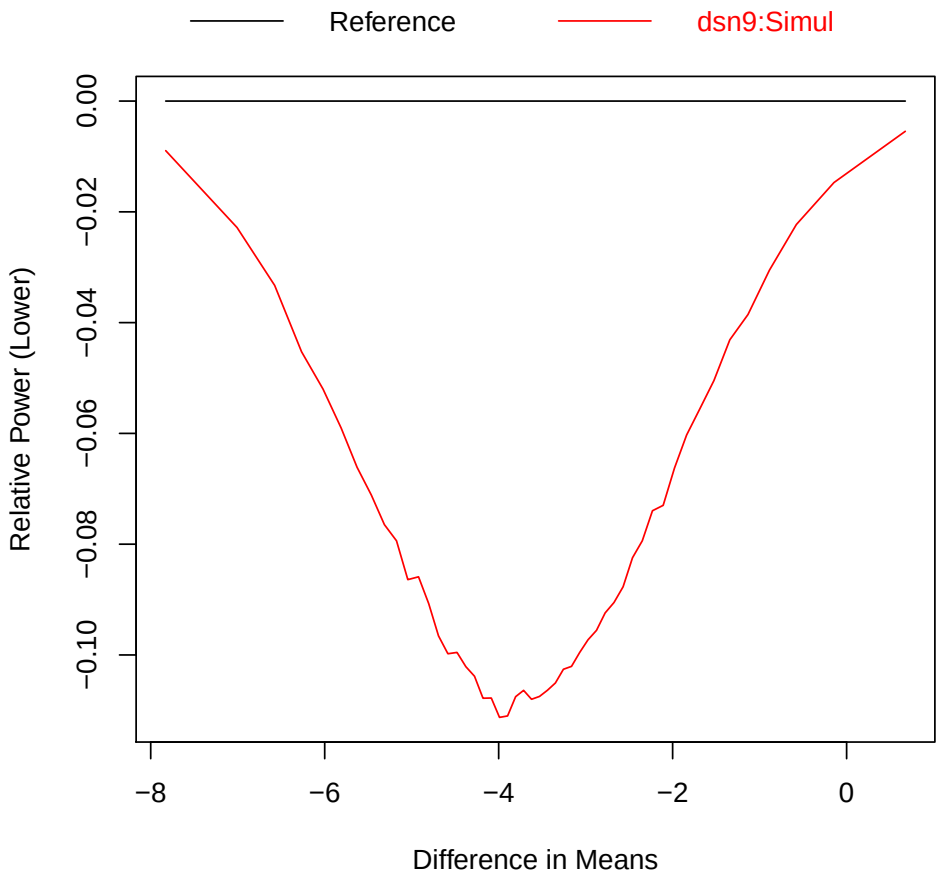
```
> dsn9 <- seqDesign( ratio=2, sd= c(30,20), nbr=4, sample.size=800,
+         power= 0.90, test.type="less")
> dsn9OCeq <- seqOC(dsn9, Nsimul=100000, seed=0, var.equal=TRUE)
> seqPlotASN(dsn9OCeq)
```




```
> seqPlotStopProb( dsn90Ceq$AsympOC, reference=dsn90Ceq$SimulOC)
```




```
> seqPlotPower( dsn90Ceq$SimulOC, reference=dsn90Ceq$AsympOC, fixed=FALSE)
```



Note that the illustrated poor behavior of the t test that presumes unequal variances in these settings does not go away as the sample sizes increase.

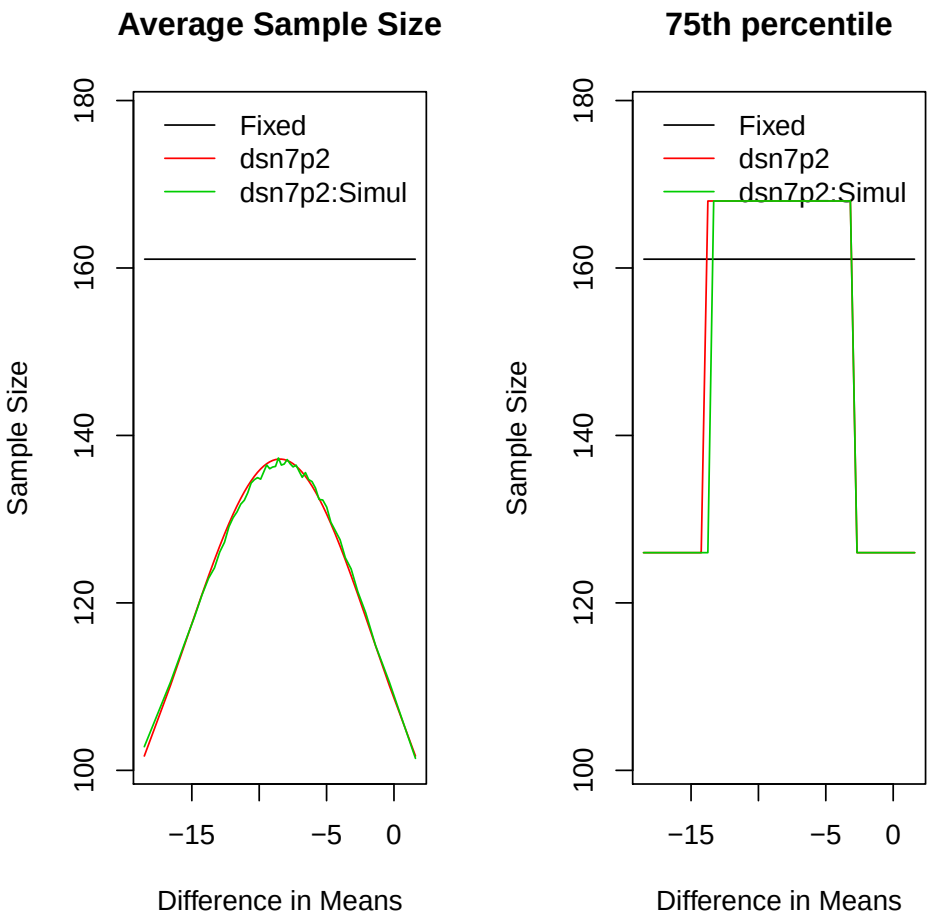
4.4 Effect of Data Distributional Shape

In the previous examples, we simulated normally distributed data. The Z test and t tests were developed to be exact or very good approximations in small samples. We can explore the behavior of the t test in settings in which the data distribution are markedly nonnormal using the `rSeq()` facility to simulate data from a user specified function. We will use a shifted exponential distribution with the `dsn7p2` and `dsn7p10` designs that have maximal sample sizes of 168 and 560, respectively. Because all data will be generated in these simulations (instead of just the sufficient statistics) we reduce the number of simulations for each case to 5-10,000.

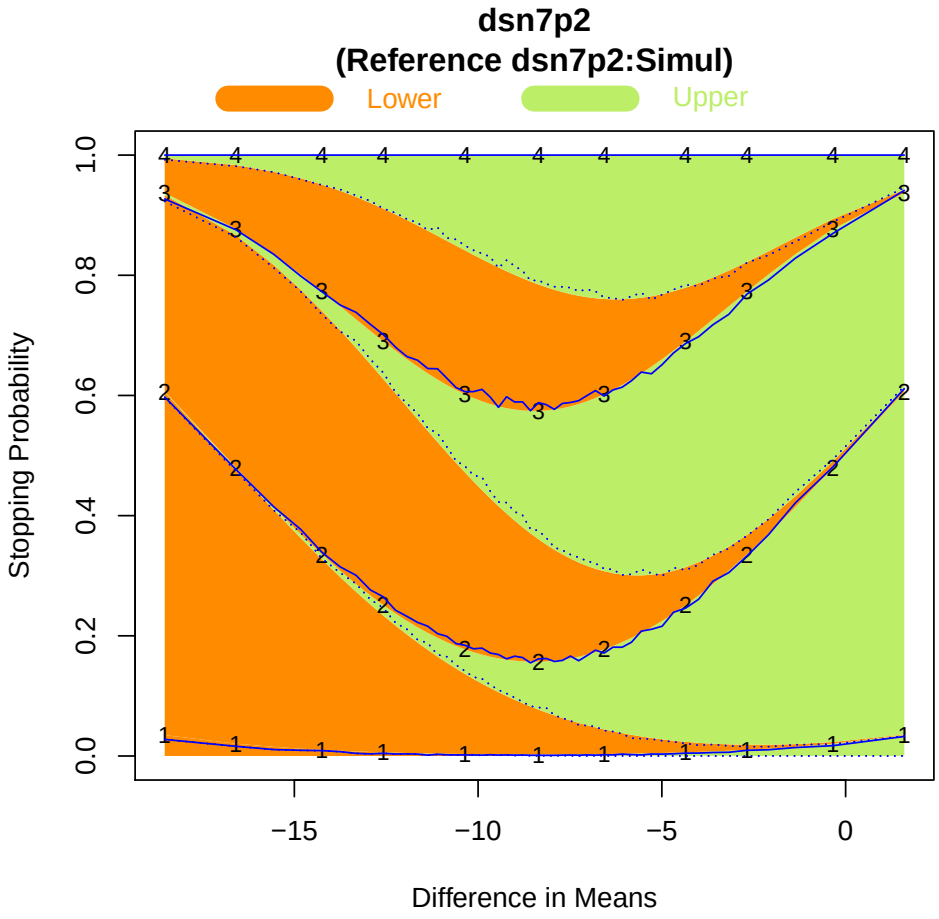
We write a function `shiftExp (n, armMean, args)` that will return data having mean `armMean`. The standard deviation will be passed to the function as `args`, and because we are going to use this in a homoscedastic (equal variance) setting, we only have to pass that single standard deviation.

```
> shiftExp <- function (n, armMean, args) {
+         args * rexp(n) + armMean
```

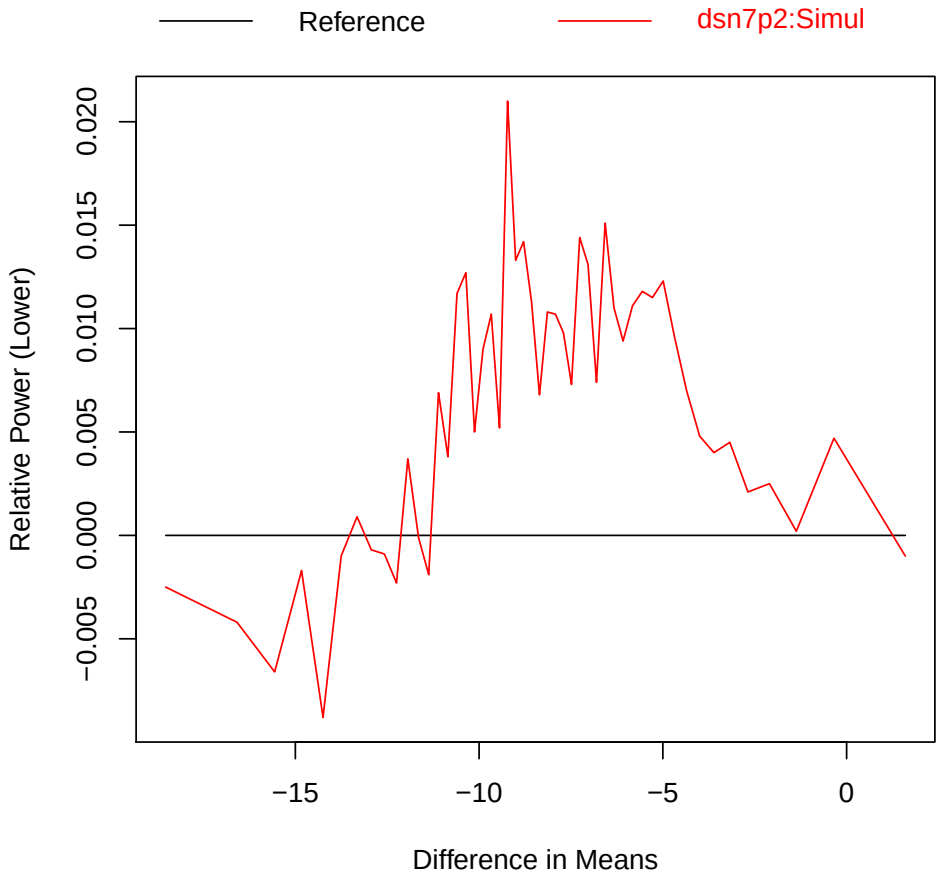
```
+         }
> dsn7p2OCexp <- seqOC(dsn7p2, Nsimul= 10000, seed=0, distn= shiftExp,
+         distnArgs= sqrt(seqExtract(dsn7p,"variance"))[1])
> seqPlotASN(dsn7p2OCexp)
```




```
> seqPlotStopProb( dsn7p2OCexp$AsympOC, reference=dsn7p2OCexp$SimulOC)
```

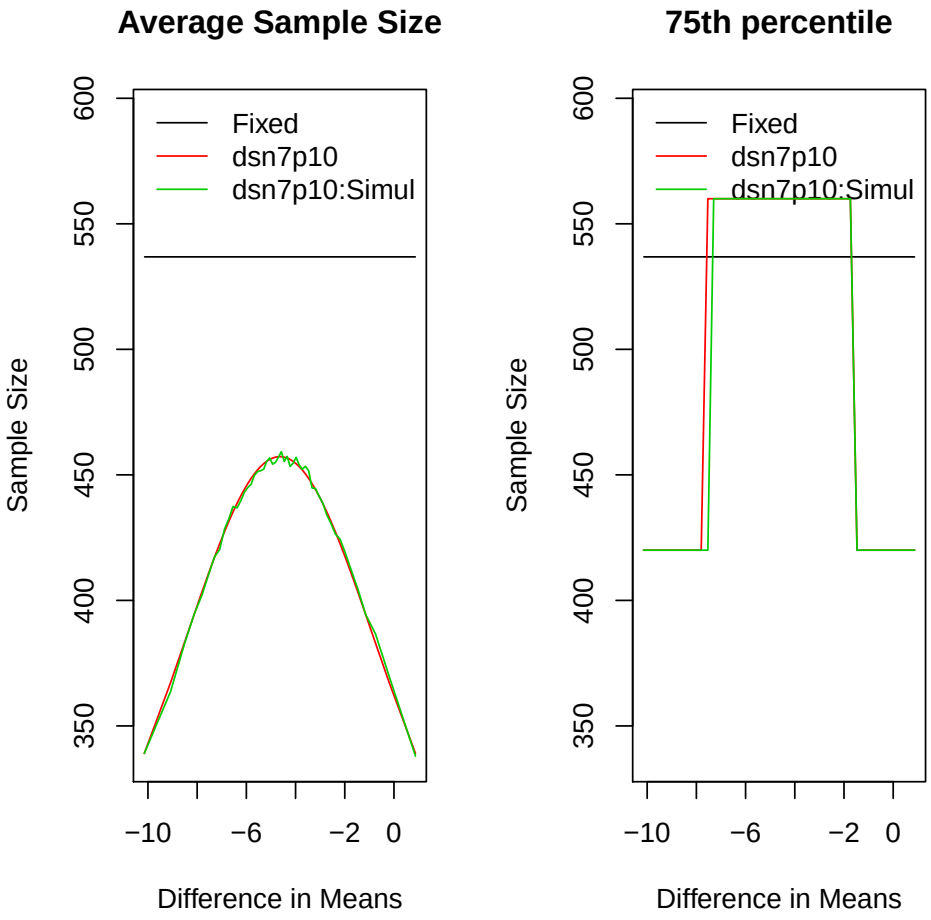



```
> seqPlotPower( dsn7p2OCexp$SimulOC, reference=dsn7p2OCexp$AsympOC, fixed=FALSE)
```

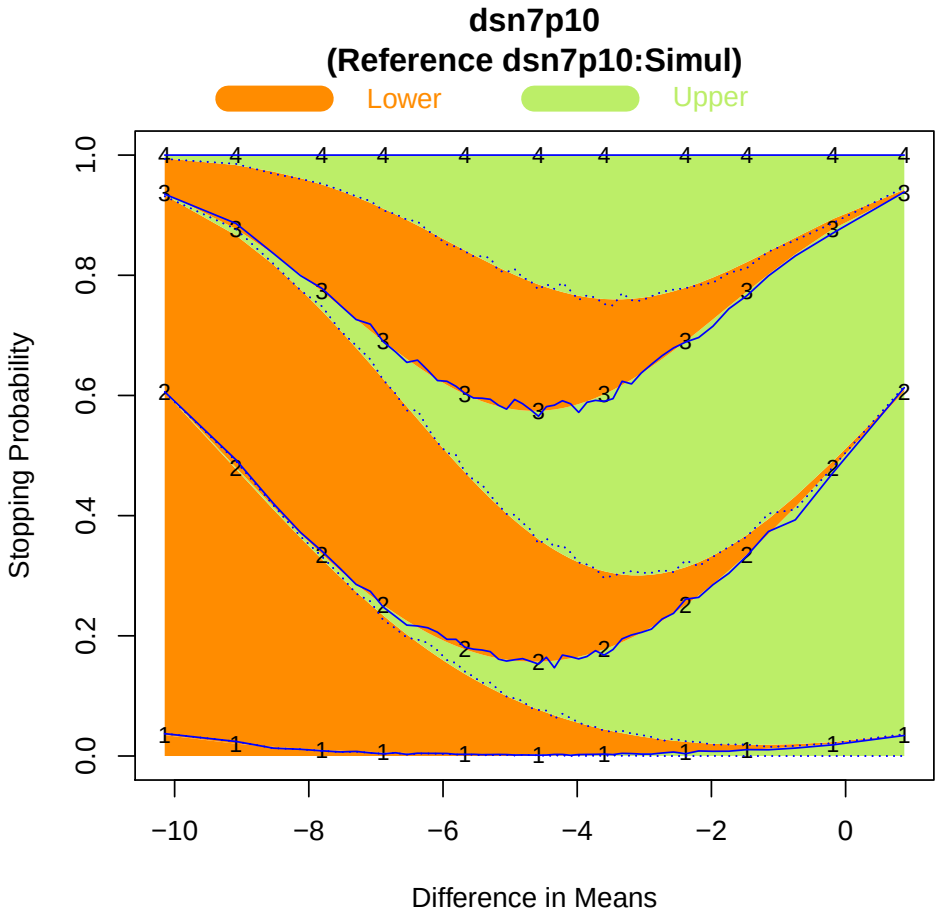


Note the worse agreement between the simulated and numerically integrated operating characteristics than was observed in section 4.1 for normally distributed data. We now consider the behavior with the larger sample size (which results can be compared to those for the same design with normally distributed data in section 4.2.

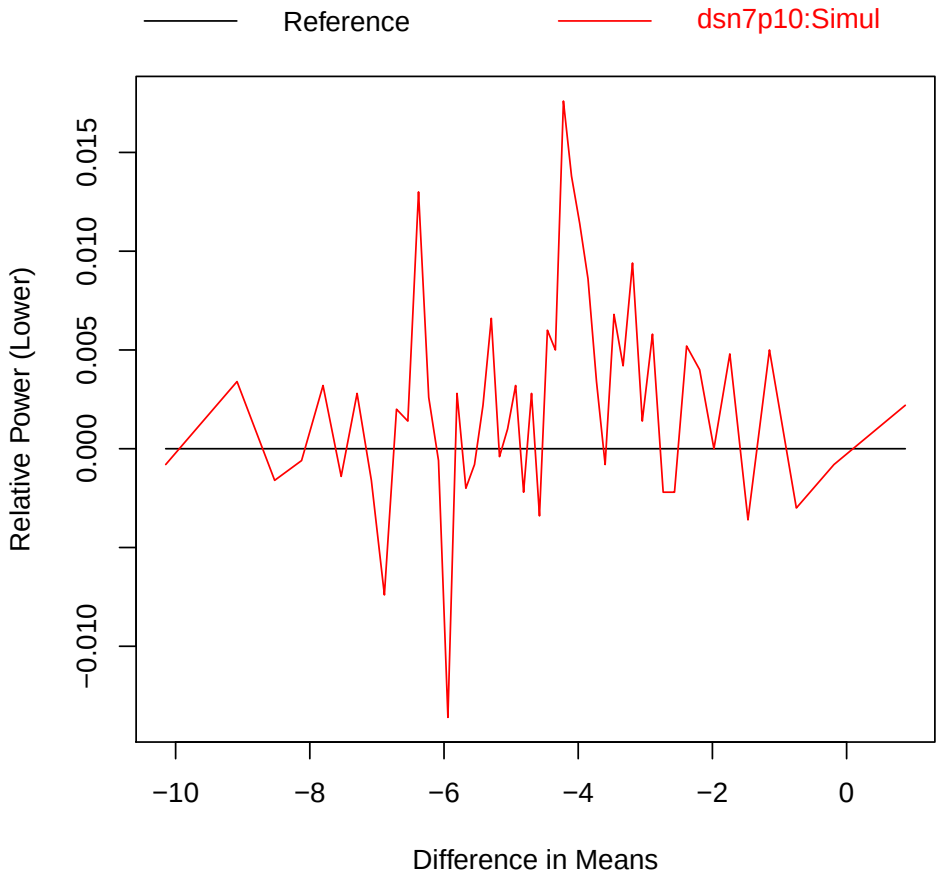
```
> shiftExp <- function (n, armMean, args) {
+         args * (1 - rexp(n)) + armMean
+         }
> dsn7p10OCexp <- seqOC(dsn7p10, Nsimul= 5000, seed=0, distn= shiftExp,
+         distnArgs= sqrt(seqExtract(dsn7p,"variance"))[1])
> seqPlotASN(dsn7p10OCexp)
```




```
> seqPlotStopProb( dsn7p10OCexp$AsympOC, reference=dsn7p10OCexp$SimulOC)
```




```
> seqPlotPower( dsn7p10OCexp$SimulOC, reference=dsn7p10OCexp$AsympOC, fixed=FALSE)
```



From the above it can be seen that nonnormality of data is not a very big issue with even moderate sample sizes: The exponential distribution has an excess kurtosis of 6.

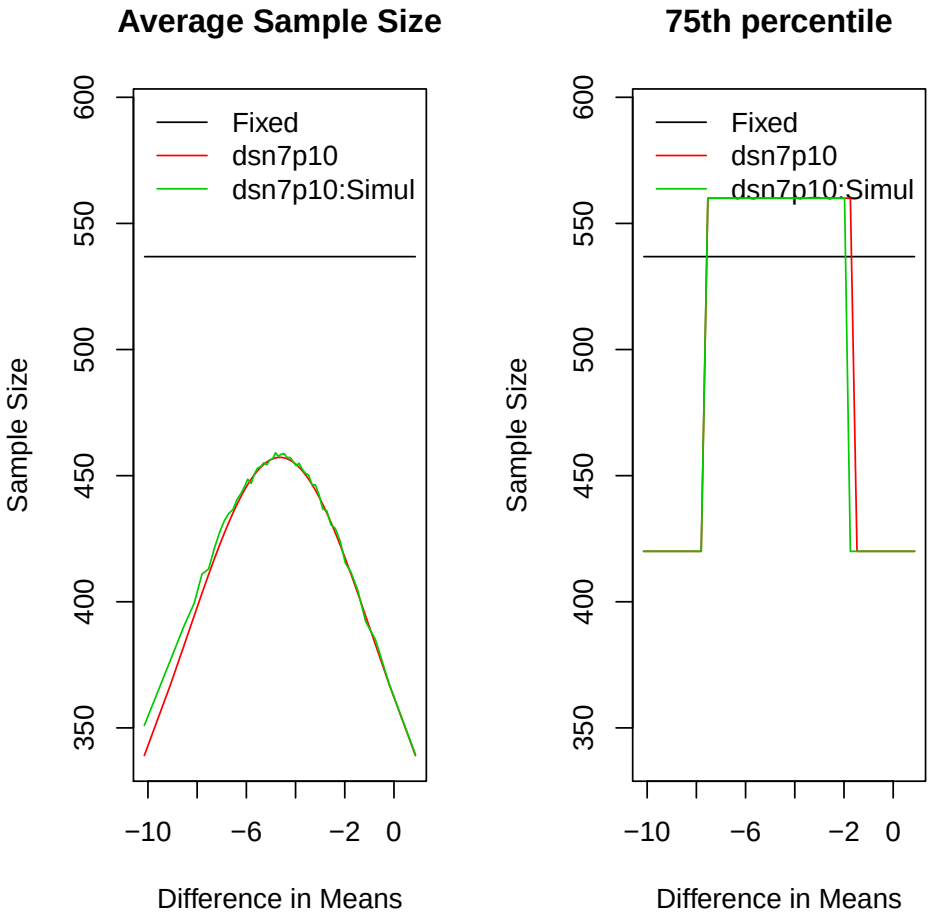
4.5 Effect of Mean-Variance Relationship

A bigger issue arises from a mean-variance relationship. It is not typical to incorporate a mean-variance relationship into the t test. However, there are many physiologic processes that would tend to mean that a treatment that affected the mean would also affect the variance. This does not necessarily mean that the type I error is affected: f the only cause of heteroscedasticity in the data is a mean-variance relationship, the variances would be equal under the null hypothesis. It can affect the statistical power in a meaningful way.

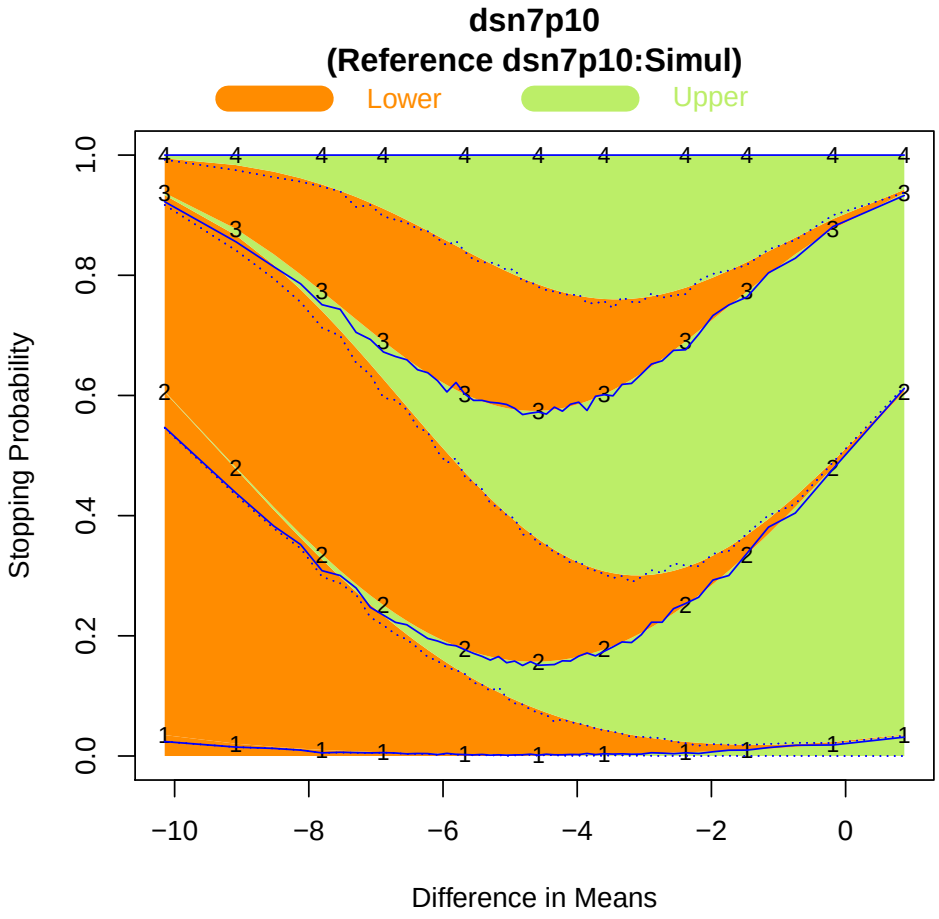
We explore this by again using the facility for user-defined data generation functions. In this case we create a normal mixture model in which the treatment is imagined to affect only 50% of the population. Hence, if the population effect is -10 mmHg, that is achieved by half the population having no benefit, and the other half averaging a decrease of 20 mmHg. We imagine that each half of the population has the same variance conditional on their status as responders or nonresponders. Hence, when the treatment has an effect, there is greater variability of response on the treatment arm than the control arm.

```
> normMix <- function (n, armMean, args) {
```

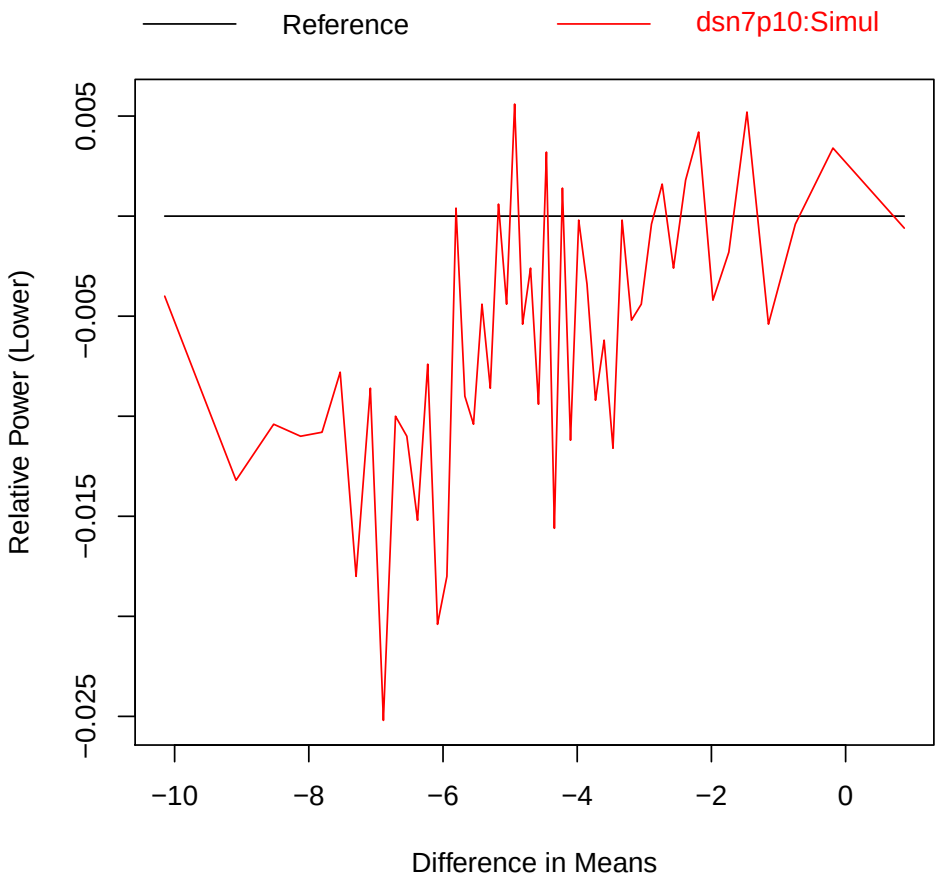
```
+         rnorm(n,0,args) + rbinom(n,1,0.5) * armMean * 2
+         }
> dsn7p10OCmix <- seqOC(dsn7p10, Nsimul= 5000, seed=0, distn= normMix,
+         distnArgs= sqrt(seqExtract(dsn7p,"variance"))[1])
> seqPlotASN(dsn7p10OCmix)
```




```
> seqPlotStopProb( dsn7p10OCmix$AsympOC, reference=dsn7p10OCmix$SimulOC)
```




```
> seqPlotPower( dsn7p10OCmix$SimulOC, reference=dsn7p10OCmix$AsympOC, fixed=FALSE)
```



Evident is the loss of power relative to the estimation from the numerical integration. Remedies to this problem require explicit modeling of the mean-variance relationship when designing the study. It should be noted, however, that as the sample size increases, the range of treatment effects between the null and design alternative decreases. Hence, any mean-variance relationship matters less. Nonetheless, there are times that it is prudent to consider how mean-variance relationships might affect study power relative to that predicted by the usual mean probability models.

Mean-variance relationships are of great interest in the proportion, odds, and rate probability models, and more discussion of such issues can be found in those tutorials.