Average Error Controlled Bayesian Sample Size Determination Methods

Sujit K. Ghosh

NC State University
DEPARTMENT OF STATISTICS

https://statistics.sciences.ncsu.edu/people/sghosh2

Presented at:

2022 ISBA World Meeting
Place Bonaventure, Montreal, Canada
https://isbawebmaster.github.io/ISBA2022/

Outline

- Limitations of Classical Methods
- Bayesian Average Errors
- Bayes Factor as Test Statistic
- Numerical Illustrations
- R package: `BAEssd`

Sample Size Determination

- Sample size determination is critical in designing medical studies
- Failure to consider sample size calculations prior to a study can have severe consequences:
  - Studies may lack power to detect clinically important effects
  - An unnecessary number of subjects may be enrolled
- E.g., the study GUSTO III with over 15,000 patients has been found under-powered to assess non-inferiority
- There are a variety of approaches to sample size determination:
  - Adcock (1997): provides an comprehensive review of various approaches
  - Inoue, Berry and Parmigiani (2005): a general framework that connects the classical and Bayesian perspectives

A safety study: Rosuvastatin therapy

- Avis et al.(2010) reported a clinical trial to determine the efficacy of *rosuvastatin therapy* for lowering cholesterol in children with familial hypercholesterolemia
- The treatment with a 20mg dose of rosuvastatin was found effective in lowering cholesterol (against placebo)
- However, the study was not powered on the secondary safety endpoints (e.g., adverse effects of 20mg of rosuvastatin)
- Suppose we want to conduct follow-up studies to assess the safety of rosuvastatin in children
- Avis et al. (2010) reported that 54% and 55% of children experienced adverse events in the placebo and rosuvastatin group
- *Can we use the results of this previous study (as prior knowledge) to determine sample sizes?*

- Consider comparing event rates of two groups based on dichotomous data
- θ_0: true (unknown) event rate of control group

  θ_1: true (unknown) event rate of experimental group
- The goal is to compare the hypotheses:

$$H_0 : \theta_0 = \theta_1 \text{ vs. } H_1 : \theta_0 \neq \theta_1$$

- Qn.: *How many subjects should we sample from each group to make a decision?*
- For classical methods, the target is to control two error rates:
  - *Type I error rate* below α (e.g., 0.05)
  - *Type II error rate* below β (e.g., 0.20)

    or equivalently *the power* above $1-\beta$ (e.g., 0.80)
- For simplicity, assume $n_1 = n_2 = n$ subjects would be sampled

- Classical (frequentist) solution:

$$n \geq \frac{\left(Z_\alpha\sqrt{2\bar{\theta}(1-\bar{\theta})} + Z_\beta\sqrt{\theta_0\left(1-\theta_0\right)+\theta_1\left(1-\theta_1\right)}\right)^2}{\left(\theta_1-\theta_0\right)^2} \tag{1}$$

  where $\bar{\theta} = \left(\theta_0+\theta_1\right)/2$ and Z_α denotes the $1-\alpha$ percentile of a standard normal distribution (e.g., $Z_{0.05} = 1.645$)

- Some obvious but critical issues:
  - n depends on *posited values for the parameters of interest* !!
  - What happens to above solution in (1) if indeed H_0 were true?
  - *No uncertainty about the posited values are accommodated*
  - Pivot quantities not guaranteed to exist (Adcock, 1997)
  - Normal approximations may be questionable (M’Lan, 2008)
  - *Wouldn’t large sample based approximations lead to larger sample?*

Limitations of Classical Methods

- Calculation of a Type-II error rate often requires the user to posit a value for theparameter under the alternative
- Positing suitable values becomes more difficult when the hypotheses are composite
- Sample size calculations under the classical framework are often based on a pivot quantity
- However, the existence of a pivot quantity is not guaranteed, even in common settings
- How to deal with nuisance parameters involved in a composite hypothesis ?
- Elimination via conditioning statistic or estimate of nuisance parameters can rarely be done in practice

Bayesian Testing Frameowrk

- Consider the general set-up of a Bayesian model:

  $$X|\theta \sim f(x|\theta) \text{ and } \theta \sim \pi(\theta) \text{ where } \theta \in \Theta \text{ and } x \in \mathcal{X}$$

- $f(x|\theta)$: joint density of the vector of observations X given θ
- $\pi(\theta)$: prior density of the vector of parameters θ
- Our goal is to compare: $H_0 : \theta \in \Theta_0$ vs. $H_1 : \theta \in \Theta_1$ where $\Theta_0 \cap \Theta_1 = \emptyset$ and $\Theta_0 \cup \Theta_1 \subseteq \Theta$
- Example: if $X_j|\theta_j \sim Bin(n_j, \theta_j)$ for $j = 0, 1$, we have $X = (X_1, X_2)$ and $\theta = (\theta_0, \theta_1) \in \Theta = [0, 1]^2 \equiv [0, 1] \times [0, 1]$
- $H_0 : \theta_0 = \theta_1 \quad \Rightarrow \quad \Theta_0 = \{\theta_0 = \theta_1 : \theta \in [0, 1]^2\}$ and
  $H_1 : \theta_0 \neq \theta_1 \quad \Rightarrow \quad \Theta_1 = \{\theta_0 \neq \theta_1 : \theta \in [0, 1]^2\}$

- We assume: $\Pr_\pi[\theta \in \Theta_j] = \int_{\Theta_j} \pi(\theta)\, d\theta > 0$ for $j = 0, 1$
- In other words, *apriori we shouldn't rule out the possibility of any of the hypotheses*
- Otherwise, no amount of data can test the validity of a hypothesis if a positive probability is not assigned to that hypothesis
- Notice that if we use the usual conjugate prior $\theta_j \sim Beta(a_j, b_j)$ for $j = 0, 1$, the condition $\Pr[\theta \in \Theta_0] = \Pr[\theta_1 = \theta_0] > 0$ is violated!
- Instead we could use the following (conjugate) prior:

$$\pi(\theta) = u\mathbb{I}\left(\theta_0 = \theta_1 = \eta\right) p_{(a_0,b_0)}(\eta) + (1-u)\mathbb{I}\left(\theta_0 \neq \theta_1\right) p_{(a_1,b_1)}(\theta_0) p_{(a_2,b_2)}(\theta_1)$$

where $u = Pr(\theta_1 = \theta_2)$ and $p_{(a,b)}(\theta)$ denotes a Beta(a, b) density

- In above, we can use any other continuous distribution replacing Beta(a, b)
- However, if we are comparing $H_0 : \theta_0 \leq \theta_1$ vs. $H_1 : \theta_0 > \theta_1$, then we can use the usual conjugate prior distributions

- Thus prior distributions should be chosen carefully based on the hypotheses being tested (making sure hypotheses are not ruled out *apriori*)

- In general, one may choose prior distributions satisfying the following condition: $\Pr[\theta \in \Theta_0] \approx \Pr[\theta \in \Theta_1] \approx 0.5$

- In the previous example choosing $u = 0.5$ guarantees the above requirement $\Pr[\theta \in \Theta_0] = \Pr[\theta \in \Theta_1] = 0.5$

- In other words, *apriori* we are not be overly biased in favor of one of the hypotheses (being tested)

- Notice that relatively non-informative priors can be used that also simultaneously satisfy above prior unbiasedness requirement

- E.g., in the previous example of testing $H_0 : \theta_0 = \theta_1$, we can choose to use Beta$(0.5, 0.5)$ (Jeffrey's prior) or the flat Beta$(1, 1)$ prior by choosing $a_0 = b_0 = a_1 = b_1 = a_2 = b_2 = 0.5$ or $= 1$

Bayesian Average Errors for Hypotheses Tests

- Within a frequentist framework, hypotheses are tested by carefully controlling the familiar *Type I & II* errors
- Regulatory purposes and various scientific considerations often necessitates the control of such error probabilities
- Bayesian sample size determination methods are often criticized as not being able to control the error probabilities for testing hypotheses
- This aspect has remained a stumbling block against the automatic adoption of Bayesian methods in clinical trials (by regulatory agencies)
- So, *can we built Bayesian methods that allow controlling such error probabilities?*
- More fundamentally, *how do we define similar error probabilities when parameters are random (with assigned prior distributions)?*

- $T(X)$: a "test statistic" measuring the evidence favoring the alternative hypothesis
- Decision rule: Reject the null hypothesis (in favor of the alternative) if $T(X) > t$ for some cut-off value t
- *How would we choose the cut-off value t?*
- Consider **Bayesian Average Error (AE) rates**:
  $AE_1(t) = \Pr[T(X) > t|\theta \in \Theta_0]$ and $AE_2(t) = \Pr[T(X) \leq t|\theta \in \Theta_1]$
- Above error rates are to be distinguished from the classical errors
- The conditional probability $\Pr[T(X) > t|\theta \in \Theta_j]$ is well defined only when $\Pr[\theta \in \Theta_j] > 0$ for $j = 0, 1$
- The quantity $(1 - AE_2(t))$ may be considered as the average power of the test
- Notice that $AE_j(t)$ does not require the user to posit a value of parameters under (null and alternative) hypotheses

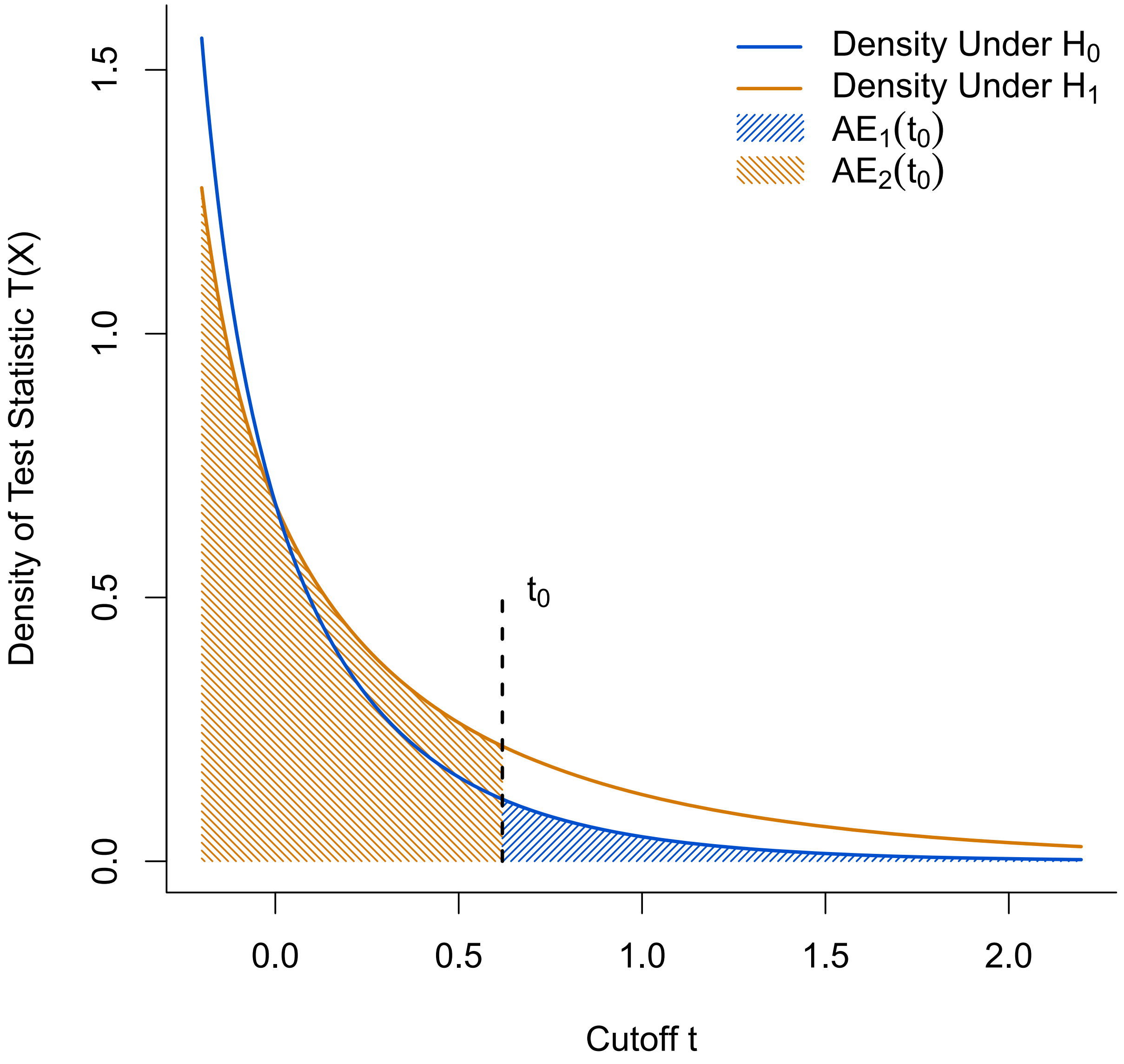
- The calculation of $AE_j(t)$ is straightforward even when there are nuisance parameters in the composite hypotheses
- Given a prior $\theta \sim \pi(\theta)$ and sampling model $X|\theta \sim f(x|\theta)$, we can compute Bayesian average Type I error probability:

$$\begin{aligned} AE_1(t) &= \Pr[T(X) > t|\theta \in \Theta_0] = \frac{\Pr[T(X) > t, \theta \in \Theta_0]}{\Pr[\theta \in \Theta_0]} \\ &= \frac{\int_{T(x)>t} \int_{\Theta_0} f(x|\theta)\pi(\theta)\, d\theta\, dx}{\int_{\Theta_0} \pi(\theta)\, d\theta} = \int_{T(x)>t} m_0(x)\, dx \end{aligned}$$

  where $m_0(x) = \frac{\int_{\Theta_0} f(x|\theta)\pi(\theta)\, d\theta}{\int_{\Theta_0} \pi(\theta)\, d\theta}$ denotes the marginal distribution of the data under the null hypothesis

- Thus, *we no longer need to obtain a pivot quantity or conditioning statistic to eliminate nuisance parameters*
- However, we do need to compute above (possibly high dimensional) integrals

- Thus, in practice we will often need to employ numerical integration methods (e.g., MCMC methods) to compute both types of Bayesian Average Errors
- Moreover, such computations need to be done in an efficient manner so that we can compute $AE_j(t)$ for any given $t \in \mathbb{R}$
- Notice that $AE_1(t) \leq \sup_{\theta \in \Theta_0} \mathrm{Pr}_\theta[T(X) > t]$ for any $t \in \mathbb{R}$
- In above, the bound is precisely the frequentist level of significance that is controlled to be below a prescribed value (e.g. ≤ 0.05)
- Note that $AE_1(t) = \mathrm{Pr}_{m_0}[T(X) > t]$ is a non-increasing function in t while $AE_2(t) = \mathrm{Pr}_{m_1}[T(X) \leq t]$ is a non-decreasing function
- Thus, as the cut-off t is altered, there is a trade-off between these two Bayesian average error rates
- Hence, we can find a cutoff t that bounds either AE_1 or AE_2 or a weighted average of these Bayesian average errors


Density Under H_0
Density Under H_1
$AE_1(t_0)$
$AE_2(t_0)$
t_0
Density of Test Statistic T(X)
0.0
0.5
1.0
1.5
0.0
0.5
1.0
1.5
2.0
Cutoff t

- A reasonable approach is to choose a cutoff t that allows for both error rates to be controlled simultaneously
- Hence, consider a *Total Weighted Error (TWE)* criterion:

$$TWE(t, w) = wAE_1(t) + (1 - w)AE_2(t)$$

  where $w \in [0, 1]$ is specified *a priori*

- The weight w can be used to place more emphasis on controlling one type of error over the other
- Given a value of $w \in [0, 1]$, the optimal cutoff $t_0(w)$ is defined as:

$$t_0(w) = \arg\min_t TWE(t, w)$$

- Thus the decision rule becomes: Reject H_0 if $T(X) > t_0(w)$
- *How do we compute $t_0(w)$? How do we find the “optimal” $T(X)$?*

Bayes Factor as Test Statistic

- Consider the *Bayes Factor* in favor of the alternative H_1:

$$BF(X) = \left(\frac{\Pr(\theta \in \Theta_1|X)}{\Pr(\theta \in \Theta_0|X)}\right) \Big/ \left(\frac{\Pr(\theta \in \Theta_1)}{\Pr(\theta \in \Theta_0)}\right)$$

- Test statistic: $T(X) = \log BF(X)$
- It is well-known that $T(x) = \log m_1(x) - \log m_0(x)$ where $m_j(x)$ denotes the marginal density under hypothesis H_j for $j = 0, 1$
- Recall that

$$m_j(x) = \frac{\int_{\Theta_j} f(x|\theta)\pi(\theta)\ d\theta}{\int_{\Theta_j} \pi(\theta)\ d\theta} \quad \text{for} \quad j = 0, 1$$

- Thus $T(X) > 0$ would favor H_1. BUT...Is 0 a good cutoff value?

Why should we use Bayes Factor (BF) as a test statistic?

It turns out that BF is indeed optimal among all test functions in the following sense:

Theorem 1. *(Reyes and Ghosh, 2013) Consider testing the hypothesis as described previously. Let BF(X) denote the Bayes factor and let*

$$\varphi(X) : \mathcal{X} \to [0, 1]$$

represent a randomized test for the hypothesis. Then, for a given value of $w \in (0, 1)$, $\hat{\varphi}(X)$ minimizes $TWE(t, w)$ where

$$\hat{\varphi}(X) = \mathbb{I}\left(BF(X) > \frac{w}{1 - w}\right).$$

Implications:

- $T(X) = \log(BF(X))$ **is optimal among all test functions**
- $t_0(w) = \log \frac{w}{1-w}$ **(universally!)**

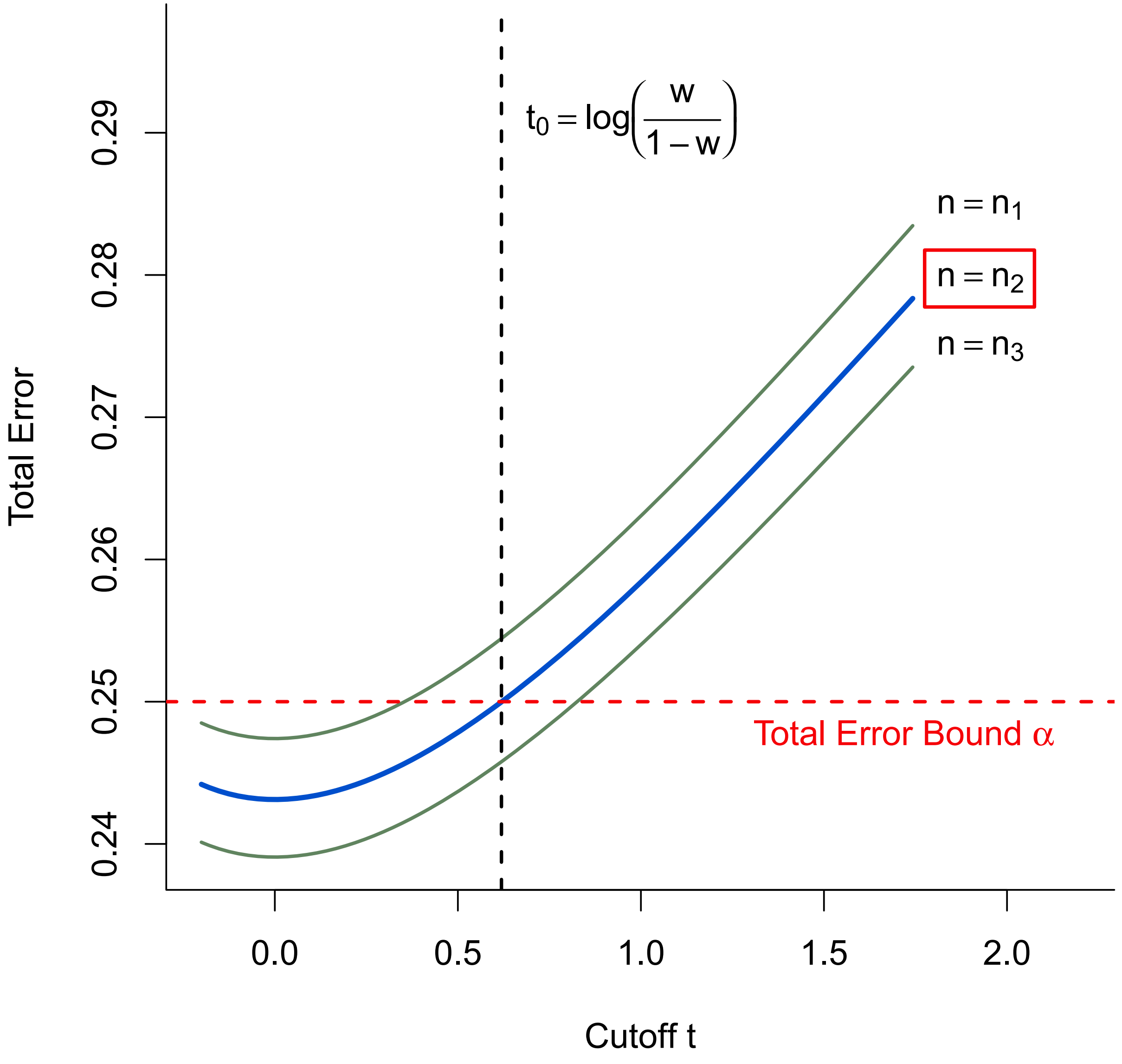
Bayesian Sample Size Determination

- The goal of any test is to control the two errors AE_1 and AE_2
- Given $\alpha, \beta \in (0,1)$, we usually take a two-step approach:
  - Bound $AE_1 \leq \alpha$ by finding a cutoff value t
  - Obtain n such that $AE_2 \leq \beta$
- Alternatively, we can also use a single step approach:

  Given a $w \in (0,1)$, obtain the minimum n such that

$$TE(t_0(w)) \leq \alpha + \beta$$

  where $TE(t) = AE_1(t) + AE_2(t)$ denotes the Total Error (TE)
- Notice that $TE(t) = 2\,TWE(t, 0.5)$
- Hence, $w = 0.5$ provides the smallest sample size


$t_0 = \log\left(\frac{w}{1-w}\right)$
$n = n_1$
$\boxed{n = n_2}$
$n = n_3$
Total Error Bound α
Total Error
0.24
0.25
0.26
0.27
0.28
0.29
0.0
0.5
1.0
1.5
2.0
Cutoff t

- For a fixed total error bound (e.g., $TE \leq \alpha + \beta$), the weight that will produce the smallest sample size is $w = 0.5$
- If $\Pr(\theta \in \Theta_0) \approx \Pr(\theta \in \Theta_1)$ then $w = 0.5$ is equivalent to rejecting the null H_0 when $\Pr(\theta \in \Theta_0|X) < \Pr(\theta \in \Theta_1|X)$
- Choosing $w = 0.5$ seems a good rule of thumb if there is no strongly preferred bound on AE_1 or AE_2
- What if the goal is to control AE_1 below α?

Theorem 2. *(Osman and Ghosh, 2011) Consider testing the hypothesis as described previously. Let $T(X) = \log BF(X)$ denote the test statistic with cutoff $t_0(w) = \log(w/(1-w))$ for a given $w \in (0,1)$. There exists $w_0 \in (0,1)$ such that for any $w > w_0$, we have,*

$$AE_1(t_0(w)) \leq TWE(t_0(w), w) \leq 1 - w$$

Implication: If we want $AE_1 \leq \alpha$ then choose $w = 1 - \alpha$

Numerical Illustrations

Consider again comparing two binomial proportions:

$X_j|\theta_j \sim Bin(n_j, \theta_j)$ for $j = 0, 1$

Want to compare: $H_0 : \theta_0 = \theta_1$ vs. $H_1 : \theta_0 \neq \theta_1$

Prior distributions:

- Under H_0: Assume $\theta_0 = \theta_1 = \eta \sim Beta(a_0, b_0)$ w.p. u
- Under H_1: Assume $\theta_j \sim Beta(a_{j+1}, b_{j+1})$ for $j = 0, 1$ w.p. $1 - u$

In other words, if $\theta = (\theta_0, \theta_1)$, we have

$$\pi(\theta) = u\mathbb{I}\left(\theta_0 = \theta_1 = \eta\right) p_{(a_0,b_0)}(\eta) + (1-u)\mathbb{I}\left(\theta_0 \neq \theta_1\right) p_{(a_1,b_1)}(\theta_0)p_{(a_2,b_2)}(\theta_1)$$

We set $u = 0.5$ and $TE \leq 0.25$ for all calculations

Prior Parameters						Results			
a_0	b_0	a_1	b_1	a_2	b_2	w	n	AE$_1$	AE$_2$
1	1	1	1	1	1	0.99	285	0.0001	0.2498
1	1	1	1	1	1	0.95	202	0.0011	0.2482
1	1	1	1	1	1	0.90	172	0.0028	0.2467
1	1	1	1	1	1	0.50	111	0.0429	0.2065
1	1	1	1	1	1	0.10	827	0.2018	0.0479

Recall that $a_0 = b_0 = 1$ correspond to $U(0, 1)$ prior on η under H_0 and $a_1 = b_1 = a_2 = b_2 = 1$ correspond $U(0, 1)$ priors on θ_0 and θ_1 under H_1

Notice that for this example $w = 0.5$ not only provides smallest sample size of 111 but it also ensures $AE_1 \approx 0.05$ and $AE_2 \approx 0.2$ as desired by regulatory agencies

Prior Parameters						Results			
a_0	b_0	a_1	b_1	a_2	b_2	w	n	AE$_1$	AE$_2$
1	1	$15/16$	$5/16$	$5/16$	$15/16$	0.99	52	0.0001	0.2485
1	1	$15/16$	$5/16$	$5/16$	$15/16$	0.95	37	0.0012	0.2487
1	1	$15/16$	$5/16$	$5/16$	$15/16$	0.90	32	0.0028	0.2452
1	1	$15/16$	$5/16$	$5/16$	$15/16$	0.50	20	0.0554	0.1916
1	1	$15/16$	$5/16$	$5/16$	$15/16$	0.10	136	0.2019	0.0472

Recall that $a_0 = b_0 = 1$ correspond to $U(0,1)$ prior on η under H_0 and $a_1 = b_2 = 15/16$ and $b_1 = a_2 = 5/16$ correspond to highly skewed priors on θ_0 and θ_1 under H_1

Here again for this case $w = 0.5$ not only provides smallest sample size of 20 but it also ensures $AE_1 \approx 0.05$ and $AE_2 \approx 0.2$

In fact, we can choose w to ensure $AE_1 \leq 0.05$ as closely as possible and $AE_2 \leq 0.2$ as closely as possible

A Comparison with classical methods:

	$d = \theta_1 - \theta_0$					
	0	0.1	0.2	0.3	0.4	0.5
n_c	∞	392	97	43	24	15
$n_{w=0.9}$	172	159	127	87	54	32
$n_{w=0.5}$	111	103	82	56	35	20
$n_{w=0.1}$	827	762	603	404	240	136

Recall that the classical sample size formula:

$$n_c = \frac{\left(Z_\alpha\sqrt{2\bar{\theta}(1-\bar{\theta})} + Z_\beta\sqrt{\theta_0\left(1-\theta_0\right)+\theta_1\left(1-\theta_1\right)}\right)^2}{\left(\theta_1-\theta_0\right)^2}$$

We have used $\alpha = 0.05$ and $\beta = 0.20$

Back to Rosuvastatin Therapy

- Using the Avis et al. (2010) study, we chooses the following prior parameters

(1) Under H_0: $\eta \sim$ Beta with mean 0.545 & variance 0.125

(2) Under H_1: $\theta_0(\theta_1) \sim$ Beta with mean 0.54 (0.55) with a variance of 0.125 for the placebo (rosuvastatin) group

- We set $u = 0.5$ and $TWE \leq \alpha + \beta = 0.15$
- Using $w = 0.5$, required sample size is $\mathbf{n} = \mathbf{243}$ subjects for each treatment arm, yielding an $AE_1 = 0.021$ and $AE_2 = 0.129$
- Reyes and Ghosh (2011) presents results based on a second study to determine if the treatment impairs renal function
- The change in Glomerular Filtration Rate (GFR) from baseline through 12 weeks of treatment is considered as the response

R package: BAEssd

Download the R package from CRAN site:

https://cran.r-project.org/web/packages/BAEssd/

```
#install the package
> install.packages('BAEssd')
#load the package after installation
> library(BAEssd)
#generate suite of function by specifying prior
> fn=binom2.2sided(prob=0.5,a0=1,b0=1,a1=1,b1=1,a2=1,b2=1)
#attach the suite
> attach(fn)
#compure log(BF) for a given data
> logbf(n=30,x=c(12,22))
[1] 2.170515
```

```
#compute the log marginal densities
> logm(n=30,x=c(12,22))
$logm0
[1] -9.03849
$logm1
[1] -6.867974
$logm
[1] -7.453058

> ssd.binom(alpha=0.25,w=0.5,logm=logm,two.sample=TRUE)

Bayesian Average Error Sample Size Determination
Call: ssd.binom(alpha = 0.25, w = 0.5, logm = logm, two.sample = TRUE)
Sample Size:  111
Total Average Error:  0.2494102
Acceptable sample size determined!

> ssd.binom(alpha=0.25,w=0.95,logm=logm,two.sample=TRUE)
```

```
Bayesian Average Error Sample Size Determination
Call: ssd.binom(alpha = 0.25, w = 0.95, logm = logm, two.sample = TRUE)
Sample Size:  202
Total Average Error:  0.2493688
Acceptable sample size determined!

> ssd.binom(alpha=0.2,w=0.5,logm=logm,two.sample=TRUE)

Bayesian Average Error Sample Size Determination
Call: ssd.binom(alpha = 0.2, w = 0.5, logm = logm, two.sample = TRUE)
Sample Size:  192
Total Average Error:  0.1998955
Acceptable sample size determined!
```

Questions?

Osman, M. and Ghosh, S. K. (2011). Semiparametric Bayesian Testing Procedure for Noninferiority Trials with Binary Endpoints, *Journal of Biopharmaceutical Statistics*, **21**, 920-937: http://dx.doi.org/10.1080/10543406.2010.544526

Reyes, E. M. and Ghosh, S. K. (2013). Bayesian Average Error Based Approach to Sample Size Calculations for Hypothesis Testing, *Journal of Biopharmaceutical Statistics*, **23**, 569-588: https://doi.org/10.1080/10543406.2012.755994

R package: https://cran.r-project.org/web/packages/BAEssd/