Sequential Design Estimators and Confidence Intervals

May 16, 2012

Setting and Notation

- Throughout this document, the notation $P(A;\mu^*)$ and $\mathrm{E}(X;\mu^*)$ denote, respectively, the probability of the event A and the expectation of the random variable X when the true value of the mean μ is taken to be μ^*.
- One-sample setting: $X_1,\ldots,X_n \overset{iid}{\sim} \text{Normal}(\mu,1)$.
- Goal: Test hypotheses and perform inference regarding the value of the population mean μ.
- Null hypothesis: $H_0 : \mu = 0$
- Alternative hypothesis: $H_A : \mu > 0$ (one sided test of greater alternative).
- Sequential test with three analysis times: $(N_1 = 100, N_2 = 200, N_3 = 300)$
- O'Brien-Fleming stopping boundary for a level $\alpha = 0.025$ test of the null hypothesis:

Analysis Time	Sample Size	Lower a Boundary	Upper d Boundary
Time 1	100	-0.1149	0.3447
Time 2	200	0.0574	0.1723
Time 3	300	0.1149	0.1149

- This boundary has power $1-\beta = 0.975$ to detect a difference of $\mu = 0.230$. It has power $1-\beta = 0.80$ to detect a difference of $\mu = 0.164$.
- For a given experiment, let (M,S) be the bivariate sufficient statistic that results at the end of the experiment, where M is the sample size at which the trial is stopped, and $S = S_M = \bar{X}_M = \frac{1}{M}\sum_{i=1}^{M} X_i$ is the sample mean when the trial is stopped. Both M and S are random quantities, as the trial may stop at any of the planned analysis times. We will denote the observed value by lower case: (m,s) are the observed values of (M,S) for a particular run of the experiment.

Ordering of Outcome Space

Many statistical procedures depend upon an ordering of the sample space (outcome space). We need a way to identify which of two possible outcomes are 'larger' under a particular hypothesized value of the true parameter μ. That is, which of the two outcomes is more consistent with the null hypothesis, and which is more consistent with the alternative?

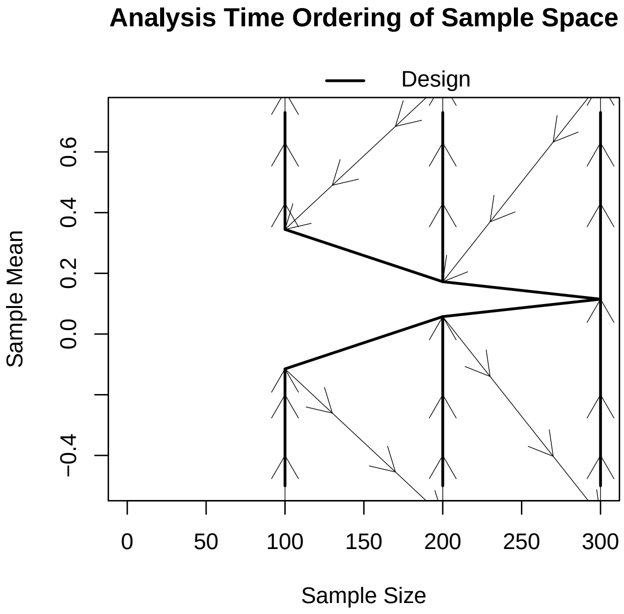
For fixed-sample testing, the sufficient statistic for this problem is the sample mean, and an obvious ordering is to say sample 1 is larger than sample 2 if $\bar{X}_{(1)} > \bar{X}_{(2)}$ when we are testing against a greater alternative. Any sample with a larger sample mean is more consistent with the alternative than a sample with a smaller mean.

For sequential testing, the sufficient statistic is bivariate: (M, S), and therefore ordering the sample space is more challenging. Several possible orderings have been proposed to try to capture which outcomes are stronger evidence for the alternative $H_A : \mu > 0$.

- Analysis Time Ordering: An outcome is considered more evidence for the alternative if it stops earlier in favor of the alternative, and less evidence if it stops earlier in favor of the null.

$$(M_{(1)}, S_{(1)}) \succ (M_{(2)}, S_{(2)}) \qquad \text{if} \qquad \begin{cases} M_{(1)} = M_{(2)}, & S_{(1)} > S_{(2)} \\ M_{(1)} > M_{(2)}, & S_{(2)} < a_{M_{(2)}} \\ M_{(1)} < M_{(2)}, & S_{(1)} > d_{M_{(1)}} \end{cases}$$

  where $a_{M_{(2)}}$ is the lower boundary at the stopping time $M_{(2)}$ for the sample 2, and $d_{M_{(1)}}$ is the upper boundary at the stopping time $M_{(1)}$ for the sample 1. So if sample 2 stops earlier and fails to reject the null, it is less consistent with the alternative than sample 1 is. If sample 1 stops earlier and rejects the null, it is more consistent with the alternative hypothesis than sample 2 is. The following figure illustrates the analysis time ordering.

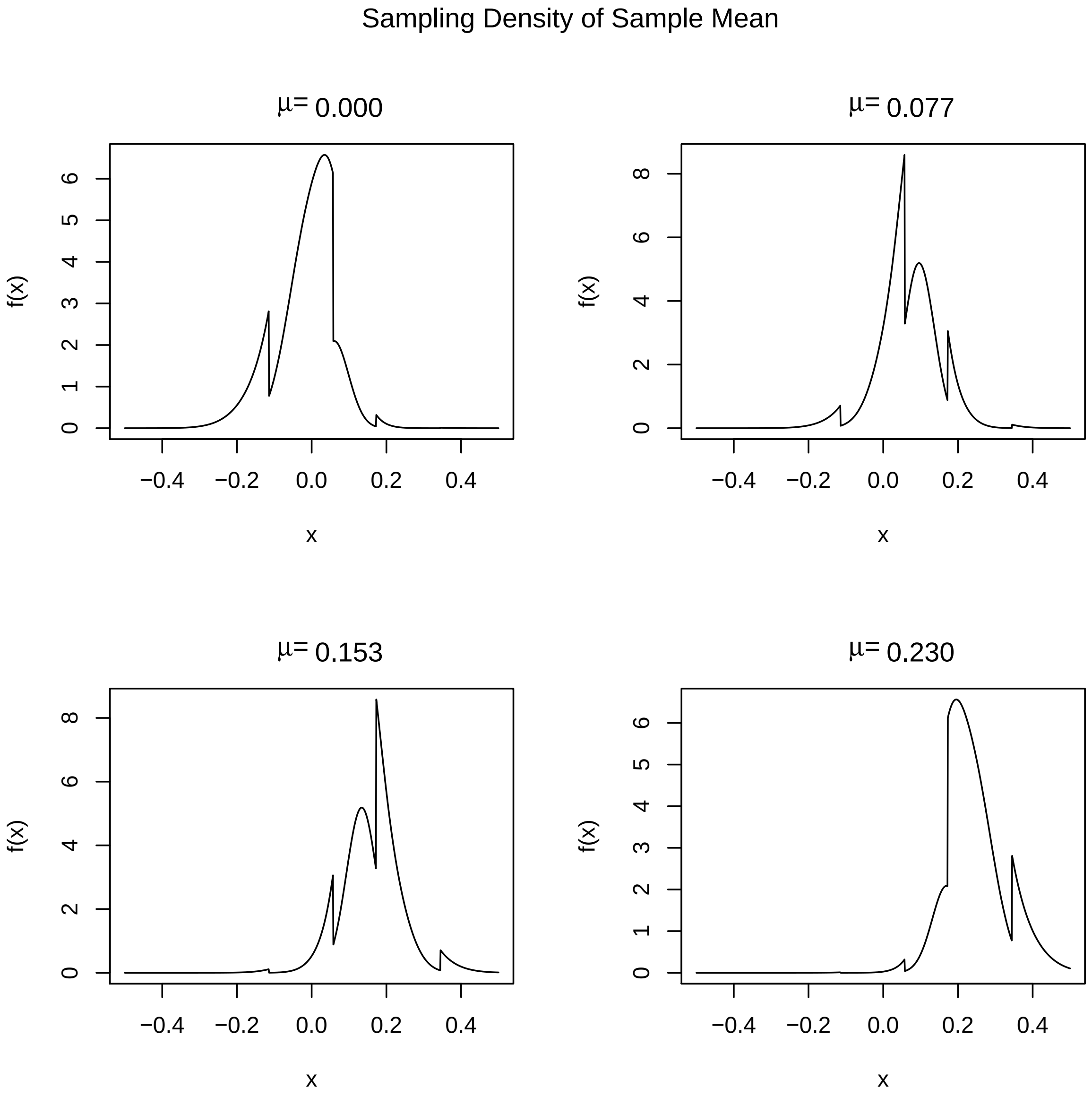


- Sample Mean Ordering: Outcomes are ordered solely on the basis of their sample mean, without regard to when the study is stopped.

$$(M_{(1)}, S_{(1)}) \succ (M_{(2)}, S_{(2)}) \qquad \text{if} \qquad S_{(1)} > S_{(2)}$$

Estimation

Estimation following a group sequential stopping rule is trickier because of the bias produced by the stopping rule. The maximum likelihood estimate of μ is still $\hat{\mu} = \bar{X}$, but now $\mathrm{E}(\hat{\mu}) \neq \mu$ in general. The following figure illustrates the sampling distribution of the sample mean $\bar{X}$ when the stopping boundaries given above are used, for different values of the true population mean μ.



Several improved approaches to estimation have been considered. We will consider three estimators:

- Maximum Likelihood Estimate: $\hat{\mu}_{\mathrm{MLE}} = \bar{X}_M = S_M$
- Bias-adjusted Mean: $\hat{\mu}_{\mathrm{BAM}}$ is the value of μ^* satisfying $\mathrm{E}\left[(M,S);\mu^*\right] = (m,s)$; that is, the value of the mean for which the observed statistic is the expected value under that mean.
- Median-unbiased Mean: $\hat{\mu}_{\mathrm{MUE}}$ is the value of μ^* satisfying $P\left((M,S) \succ (m,s);\mu^*\right) = 0.5$; that is, the value of the mean for which the observed statistic would be the median of the sampling distribution under that mean. ***Note that this estimator depends on the ordering of the outcome space.***

The values of these estimators for a particular observed statistic value (m, s) may be found by numerical integration or simulation; that is, we can find the expectation (or median) of (M, S) for many different possible values of μ^* and then identify the value of μ^* that gives us an expectation (or median) closest to the observed statistic (m, s).

Optimality Criteria for Estimators

The usual optimality criteria for estimators include

- Bias: is the expected value of the estimator equal to the true parameter value?

$$B(\hat{\mu}; \mu) = \mathrm{E}(\hat{\mu}) - \mu$$

- Mean-squared Error: What is the expected squared distance between the estimator and the true parameter value?

$$\mathrm{MSE}(\hat{\mu}; \mu) = \mathrm{E}\left[(\hat{\mu} - \mu)^2\right]$$

- Consistency: Does the estimator converge (in probability) to the true value as the sample size increases to infinity? This property is less emphasized for sequential designs, as we are primarily interested in the sample size for which the study is planned.
- Agreement with test decision: Is it possible that the estimate be in the null hypothesis region of the parameter space, but the decision based on the boundary is to reject the null?

Confidence Intervals

General $(1 - \alpha)100\%$ confidence intervals may be formed using the hypothesis testing-confidence region duality, giving:

$$\mathcal{I}_{(1-\alpha)} = \{\mu^* |\, H_0 : \mu = \mu^* \text{ would not be rejected at level } \alpha\}.$$

The sequential boundary was chosen to test the null hypothesis $H_0 : \mu = 0$ vs. a one-sided greater alternative, but we can still construct two-sided hypothesis tests for any null hypothesis $H_0 : \mu = \mu^*$ if we know the sampling distribution of the test statistic (sufficient statistic) and if we have an ordering of the outcome space to tell us which results are more extreme for the hypothesis under consideration.

We will consider a two-sided hypothesis test based on computing a two-sided p-value for the observed statistic under the hypothesized value of $\mu = \mu^*$, as follows:

$$\begin{aligned} &\text{(One-sided upper } p\text{-value)} \quad && \tilde{p}(m, s; \mu^*) = P\left((M, S) \succ (m, s); \mu^*\right) \\ &\text{(Two-sided } p\text{-value)} \quad && p(m, s; \mu^*) = 2 \min\left(\tilde{p}(m, s; \mu^*), 1 - \tilde{p}(m, s; \mu^*)\right) \end{aligned}$$

Then a two-sided hypothesis test of $H_0 : \mu = \mu^*$ may be performed by rejecting H_0 at level α if $p(m, s; \mu^*) < \alpha$. Therefore, a $(1 - \alpha)100\%$ confidence interval for μ is given by

$$\mathcal{I}_{(1-\alpha)} = \{\mu^* |\, p(m, s; \mu^*) > \alpha\}$$

To evaluate/construct a $(1 - \alpha)100\%$ confidence region for μ, we therefore need to be able to assess

$$P\left((M, S) \succ (m, s); \mu^*\right)$$

for any value of μ^* and an observed statistic (m, s), for the chosen ordering of the sample space. We can simulate to obtain these probabilities. A rough simulation outline follows:

- Construct a vector/grid of possible μ^* values, e.g. `muVec` $= (-1, -0.99, -0.98, ..., 0, 0.01, ..., 1.49, 1.5)$.

- For each value of possible μ^* values, generate `nsim` = 10,000 or 100,000 experiments, each experiment consisting of a sample of 300 observations, from which you compute the observed statistic $(M, S) =$ (observed stopping time, observed sample mean at chosen stopping time).
- Store a matrix of the stopping times for each simulated dataset, for each value of the mean μ^*, and another matrix of the sample means, so each matrix will be have dimensions `nsim` × `length(muVec)`.
- Using these stored matrices, you can then compute for a given value of an observed statistic (m^*, s^*) the associated p-value for each value of μ^* by comparing the value (m^*, s^*) to the set of simulated values using the ordering of choice. You would just have to count how many/what proportion of the simulated values for that particular μ^* are larger than (m^*, s^*) according to the chosen ordering.

Optimality Criteria for Confidence Intervals

- Coverage Probability: This should be the target level $(1 - \alpha)100\%$ by construction, but it is important to assess whether that is actually being achieved.
- Convexity: are the confidence regions true intervals? If $\mu_1 \in \mathcal{I}$ and $\mu_2 \in \mathcal{I}$, then we would like to have$\beta\mu_1 + (1 - \beta)\mu_2 \in \mathcal{I}$ for any $\beta \in (0, 1)$.
- Agreement with decision: If the study is stopped for efficacy (so $H_0 : \mu = 0$ is rejected in favor of the greater alternative), then μ_0 should not be in the $(1 - \alpha)100\%$ confidence region, where α is level for which the stopping boundaries were designed. If the study is stopped for futility (so $H_0 : \mu = 0$ is not rejected), then the design alternative at which the design has power $1 - \beta = 1 - \alpha$ should not be in the confidence region.
- Point estimates of μ should be in the confidence region. This is especially important for the well-behaved point estimates like the Bias-adjusted Mean. Some reasonable confidence regions may not contain the Maximum Likelihood Estimate, which is acceptable because of the bias of the MLE.
- Length of the confidence region. Methods that produce shorter intervals are preferred. Interval length may be compared on basis of average length or median or other quantile of the length of intervals produced by each method.