PS1

Sam Olson

1.

In Chapter 11.1.2 of the Stat 520 notes we gave the basic form of a Monte Carlo approximation as

$$E_M\{g(X_m)\} = \frac{1}{M}\sum_{m=1}^{M} g(X_m^*), \tag{1}$$

where X_m^* were values sampled from the distribution of X, call it f.

Chapter 11.2 contained an example which compared coverage rates of approximate and exact confidence intervals for the difference in means. For this question, consider only the approximate interval given in expression (11.14) of the Stat 520 notes, and use M_1 from expression (11.13) as the pertinent model. Table 11.1 on page 476 of the Stat 520 notes reports observed coverage rates under the column headed *MC Approx* for various Monte Carlo sample sizes M.

The coverage rates of concern are called Monte Carlo Approximations, and hence must have the form of (1) above. Explicitly identify $g(X_m)$ and f for this use of Monte Carlo.

Answer

As given, a Monte Carlo approximation to an expectation is given by

$$E_M\{g(X)\} = \frac{1}{M}\sum_{m=1}^{M} g(X_m^*), \tag{1}$$

where $X_1^*, \ldots, X_M^*$ are independent draws from the distribution of a random vector X, denoted by f.

Consider model M_1 as defined in (11.13),

$$M_1: \quad Y_{1,i} \overset{iid}{\sim} N(\mu_1, \sigma_1^2), \qquad Y_{2,i} \overset{iid}{\sim} N(\mu_2, \sigma_2^2), \tag{11.13}$$

with the two samples independent.

We seek the Monte Carlo approximation to the coverage probability of the approximate confidence interval for $\mu_1 - \mu_2$ under this model.

To that end, a single Monte Carlo draw consists of the full dataset

$$X = \big(Y_{1,1}, \ldots, Y_{1,n_1}, Y_{2,1}, \ldots, Y_{2,n_2}\big).$$

Under model M_1, the joint distribution of X has density f given by

$$f(x) = \prod_{i=1}^{n_1} \phi(y_{1,i}; \mu_1, \sigma_1^2) \prod_{i=1}^{n_2} \phi(y_{2,i}; \mu_2, \sigma_2^2)$$
$$= \prod_{i=1}^{n_1} \frac{1}{\sqrt{2\pi\sigma_1^2}} \exp\left\{ -\frac{(y_{1,i} - \mu_1)^2}{2\sigma_1^2} \right\} \prod_{i=1}^{n_2} \frac{1}{\sqrt{2\pi\sigma_2^2}} \exp\left\{ -\frac{(y_{2,i} - \mu_2)^2}{2\sigma_2^2} \right\},$$

where $\phi(\cdot; \mu, \sigma^2)$ denotes the $N(\mu, \sigma^2)$ density.

The approximate confidence interval for $\mu_1 - \mu_2$ is given by

$$\bar{Y}_1 - \bar{Y}_2 \pm z_{1-\alpha/2} \left[\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right]^{1/2}. \tag{11.14}$$

Define the function

$$g(X) = \mathbf{1}\left\{ \left| \bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2) \right| \leq z_{1-\alpha/2} \left[\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2} \right]^{1/2} \right\},$$

where $\mathbf{1}\{\cdot\}$ denotes the indicator function. Note: $g(X) \in \{0, 1\}$ is a Bernoulli random variable under f.

Under the data-generating distribution f, the exact coverage probability of the interval is

$$E_f[g(X)] = P_f\{\mu_1 - \mu_2 \text{ is covered by the interval}\}.$$

Let $X_1^*, \ldots, X_M^* \overset{iid}{\sim} f$ be independent Monte Carlo samples generated under model M_1. The Monte Carlo approximation to the coverage probability is

$$E_M\{g(X)\} = \frac{1}{M} \sum_{m=1}^{M} g(X_m^*),$$

which is precisely of the form given in (1).

2.

In Chapter 11.3.2 of the Stat 520 notes, a $(1-\alpha)100\%$ confidence interval for a parameter θ based on subsampling is represented as (L, U) where these quantities are given in expression (11.25) as,

$$L = \hat{\theta}_M - \tau_M^{-1} q_{M,\,1-\alpha/2}$$

$$U = \hat{\theta}_M - \tau_M^{-1} q_{M,\,\alpha/2}.$$

Here, $q_{M,\nu}$ is the νth quantile of $L_{M,b}(y)$, the empirical distribution function of $\tau_b(\hat{\theta}_{b,j} - \hat{\theta}_M)$ as in expression (11.24) of the Stat 520 notes.

Verify that the intervals given in (11.25) of the Stat 520 notes are correct.

Hint:

(a) The subsampling principle is that the distribution function of $\tau_M(\hat{\theta}_M - \theta)$ is approximated by $L_{M,b}(y)$, the empirical distribution function of

$$\{\tau_b(\hat{\theta}_{b,j} - \hat{\theta}_M) : j = 1, \ldots, k\}.$$

Take this as being true. That is, the difference between the sampling distribution of $\tau_M(\hat{\theta}_M - \theta)$ and the empirical distribution of $\{\tau_b(\hat{\theta}_{b,j} - \hat{\theta}_M) : j = 1, \ldots, k\}$ is not the key for this question.

(b) Begin with what is desired, finding quantities L and U such that

$$\Pr(\theta < L) = \Pr(\theta > U) = \alpha/2.$$

Answer

We seek to verify that the confidence interval endpoints in (11.25),

$$L = \hat{\theta}_M - \tau_M^{-1} q_{M,\,1-\alpha/2}, \qquad U = \hat{\theta}_M - \tau_M^{-1} q_{M,\,\alpha/2}, \tag{11.25}$$

form a $(1-\alpha)100\%$ confidence interval for θ in the subsampling framework.

Using Hint (b), begin with what is desired: find L and U such that

$$\Pr(\theta < L) = \alpha/2, \qquad \Pr(\theta > U) = \alpha/2.$$

These equal-tail conditions motivate the construction of a confidence interval with approximate central coverage.

In particular, if we choose L and U so that each tail probability is approximately $\alpha/2$, then

$$\Pr(L \le \theta \le U) \approx 1 - \alpha.$$

Define the centered-and-scaled statistic

$$T_M = \tau_M(\hat{\theta}_M - \theta)$$

We now make explicit use of (11.24). The subsampling empirical distribution function is

$$L_{M,b}(y) = \frac{1}{M-b+1} \sum_{j=1}^{M-b+1} I\Big\{\tau_b(\hat{\theta}_{b,j} - \hat{\theta}_M) \le y\Big\}, \tag{11.24}$$

where $L_{M,b}(y)$ is the empirical CDF of the subsampling values

$$\Big\{\tau_b(\hat{\theta}_{b,j} - \hat{\theta}_M) : j = 1, \ldots, M-b+1\Big\}.$$

Then, let $q_{M,\nu}$ denote the νth sample quantile of the empirical distribution $L_{M,b}$, defined as

$$q_{M,\nu} = \inf\{y : L_{M,b}(y) \ge \nu\}.$$

Using Hint (a), we take as given that the distribution function of $T_M = \tau_M(\hat{\theta}_M - \theta)$ is approximated by $L_{M,b}(y)$, i.e., the difference between the sampling distribution of $\tau_M(\hat{\theta}_M - \theta)$ and the empirical distribution is negligible.

Therefore the central $(1-\alpha)$ probability region for T_M is approximated by the corresponding central region under $L_{M,b}$,

$$\Pr\big(q_{M,\alpha/2} \le T_M \le q_{M,1-\alpha/2}\big) \approx 1 - \alpha.$$

Substitution of $T_M = \tau_M(\hat{\theta}_M - \theta)$ then gives

$$\Pr\Big(q_{M,\alpha/2} \le \tau_M(\hat{\theta}_M - \theta) \le q_{M,1-\alpha/2}\Big) \approx 1 - \alpha.$$

Assuming $\tau_M > 0$ (to avoid division by zero), the inequality

$$q_{M,\alpha/2} \le \tau_M(\hat{\theta}_M - \theta) \le q_{M,1-\alpha/2}$$

is equivalent to

$$\tau_M^{-1} q_{M,\alpha/2} \le \hat{\theta}_M - \theta \le \tau_M^{-1} q_{M,1-\alpha/2}$$

which is equivalent to

$$\hat{\theta}_M - \tau_M^{-1} q_{M,1-\alpha/2} \le \theta \le \hat{\theta}_M - \tau_M^{-1} q_{M,\alpha/2}$$

Therefore, defining

$$L = \hat{\theta}_M - \tau_M^{-1} q_{M,1-\alpha/2}, \qquad U = \hat{\theta}_M - \tau_M^{-1} q_{M,\alpha/2},$$

we obtain

$$\Pr(L \le \theta \le U) \approx 1 - \alpha,$$

which verifies the endpoints given in (11.25).

Finally, we connect directly back to Hint (b).

Because $q_{M,\alpha/2}$ and $q_{M,1-\alpha/2}$ are the $\alpha/2$ and $1-\alpha/2$ quantiles of the empirical distribution $L_{M,b}$, and Hint (a) states that $L_{M,b}$ approximates the distribution of

$$T_M = \tau_M(\hat{\theta}_M - \theta),$$

we have the approximate tail probabilities

$$\Pr(T_M \le q_{M,\alpha/2}) \approx \alpha/2, \qquad \Pr(T_M \ge q_{M,1-\alpha/2}) \approx \alpha/2.$$

Since $\tau_M > 0$, the inequality

$$T_M \le q_{M,\alpha/2}$$

is equivalent to

$$\tau_M(\hat{\theta}_M - \theta) \le q_{M,\alpha/2} \quad \Longleftrightarrow \quad \theta \ge \hat{\theta}_M - \tau_M^{-1} q_{M,\alpha/2} = U,$$

which implies

$$\Pr(\theta \ge U) \approx \alpha/2.$$

Similarly,

$$T_M \ge q_{M,1-\alpha/2}$$

is equivalent to

$$\tau_M(\hat{\theta}_M - \theta) \ge q_{M,1-\alpha/2} \quad \Longleftrightarrow \quad \theta \le \hat{\theta}_M - \tau_M^{-1} q_{M,1-\alpha/2} = L,$$

which implies

$$\Pr(\theta \le L) \approx \alpha/2.$$

Assuming boundary probabilities are negligible, these conditions match the target equal-tail requirements from Hint (b) and confirm that the subsampling confidence interval is given by

$$L = \hat{\theta}_M - \tau_M^{-1} q_{M,\,1-\alpha/2}, \qquad U = \hat{\theta}_M - \tau_M^{-1} q_{M,\,\alpha/2}. \tag{11.25}$$