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

The Monte Carlo approximation is constructed using model M_1 as defined by

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

with the two samples independent.

A single Monte Carlo draw consists of the full dataset

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

generated according to the joint distribution implied by M_1. Denote this joint distribution of X by f.

The approximate confidence interval is based on expression (11.14),

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

which is equivalent to the pair of inequalities

$$\left|\bar{Y}_1 - \bar{Y}_2 - (\mu_1 - \mu_2)\right| \le z_{1-\alpha/2}\left[\frac{S_1^2}{n_1} + \frac{S_2^2}{n_2}\right]^{1/2}.$$

Define the function

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

which is the indicator that the approximate confidence interval in (11.14), computed from the dataset X, contains the true parameter value $\mu_1 - \mu_2$. In particular, for the mth Monte Carlo replicate,

$$g(X_m^*) = \mathbf{1}\{\mu_1 - \mu_2 \in CI(X_m^*)\}.$$

Under model M_1, the exact coverage probability of the interval (11.14) is

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

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

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

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

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 want 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), we begin with what is desired: find L and U such that

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

Equivalently,

$$\Pr(L \leq \theta \leq U) = 1 - \alpha.$$

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

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

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

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

Let $q_{M,\nu}$ denote the νth quantile of this empirical distribution, meaning

$$L_{M,b}(q_{M,\nu}) = \nu \qquad \text{(equivalently, } q_{M,\nu} \text{ is the } \nu\text{th sample quantile of the above set).}$$

Using Hint (a), we take as given that the distribution function of $T_M = \tau_M(\hat{\theta}_M - \theta)$ is approximated by $L_{M,b}(y)$. Therefore the central $(1-\alpha)$ probability region for T_M is approximated by the corresponding central region under $L_{M,b}$, namely

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

Substituting $T_M = \tau_M(\hat{\theta}_M - \theta)$ 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 $\tau_M > 0$, the inequality

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

is equivalent to

$$\hat{\theta}_M - \theta \le \tau_M^{-1} q_{M,1-\alpha/2} \qquad \text{and} \qquad \hat{\theta}_M - \theta \ge \tau_M^{-1} q_{M,\alpha/2}.$$

Rearranging for θ yields

$$\theta \ge \hat{\theta}_M - \tau_M^{-1} q_{M,1-\alpha/2} \qquad \text{and} \qquad \theta \le \hat{\theta}_M - \tau_M^{-1} q_{M,\alpha/2}.$$

Thus the event

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

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 $L_{M,b}$, and Hint (a) says $L_{M,b}$ approximates the distribution of T_M, we have the tail approximations

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

The inequality $T_M \le q_{M,\alpha/2}$ is $\tau_M(\hat{\theta}_M - \theta) \le q_{M,\alpha/2}$, which is equivalent (since $\tau_M > 0$) to $\theta \ge \hat{\theta}_M - \tau_M^{-1} q_{M,\alpha/2} = U$, hence

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

Similarly, $T_M \ge q_{M,1-\alpha/2}$ is equivalent to $\theta \le \hat{\theta}_M - \tau_M^{-1} q_{M,1-\alpha/2} = L$, hence

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

This matches the target tail conditions from Hint (b) and confirms

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