Statistics 322, Statistical Computing, Home Work 6

Posted: Oct. 1, 1996 Due: Oct. 10, 1996

The goal of this is a simulation study to investigate numerically some of the issues about bootstrapping that have been developed theoretically in class. Several methods were discussed for the construction of 1-sided confidence intervals for the mean μ of a distribution F, based on a sample $X_1, \ldots, X_n$ from F. Here we will study:

CI1: Ordinary Central Limit Thereom based:

$$[\overline{X} - z_\alpha \hat{\sigma} n^{-1/2}, \infty),$$

where z_α is the α quantile of the $N(0, 1)$ distribution, and where $\hat{\sigma}^2$ is the "maximum likelihood estimate" $\hat{\sigma}^2 = \frac{1}{n} \Sigma_{i=1}^n \left(X_i - \overline{X}\right)^2$

CI2: Student's t distribution based:

$$\left[\overline{X} - t_\alpha s n^{-1/2}, \infty\right\},$$

where t_α is the α quantile of Student's t distribution with $n-1$ degrees of freedom, and where s^2 is the "usual unbiased estimate" $s = \frac{1}{n-1} \Sigma_{i=1}^n \left(X_i - \overline{X}\right)^2$ (note `std` gives this).

CI3: Direct Bootstrap:

$$\left[\overline{X} - q_\alpha^* n^{-1/2}, \infty\right),$$

where q_α^* is the α quantile of the bootstrap distribution of $n^{1/2}\left(\overline{X}^* - \overline{X}\right)$.

CI4: Studentized Bootstrap:

$$\left[\overline{X} - \tilde{q}_\alpha^* \hat{\sigma} n^{-1/2}, \infty\right),$$

where $\tilde{q}_\alpha^*$ is the α quantile of the bootstrap distribution of $n^{1/2}\left(\frac{\overline{X}^* - \overline{X}}{\hat{\sigma}^*}\right)$,

for $\alpha = .90$, and for the underlying distributions:

F1: $X_1 = Z \sim N(0, 1)$, i.e. $F_1(x) = \Phi(x)$.

F2: $X_2 \sim exp\ (1)$, i.e. $F_2(x) = (1 - e^{-x}) 1_{\{x \geq 0\}}$.

F3: $X_3 \sim \chi_1^2$, i.e. $X_3 = Z^2$, i.e. $F_3(x) = 2\Phi\left(\sqrt{x}\right) - 1$.

1. Verify that these distributions satisfy:

F1: $\mu_1 = 0$, $\sigma_1^2 = 1$, $\kappa_{3,1} = 0$.

F2: $\mu_2 = 1, \ \sigma_2^2 = 1, \ \kappa_{3,2} = 2.$

F3: $\mu_3 = 1, \ \sigma_3^2 = 2, \ \kappa_{3,3} = 2\sqrt{2}.$

2. Write a MATLAB program which estimates coverage probabilities for:

(a) CI1, CI2, CI3, CI4,

using:

(b) 1000 pseudo data sets, for sample sizes $n = 10, \ 25, \ 100$, and each distribution, F1, F2, F3.

(c) $B = 499$ bootstrap replications for CI3 and CI4.

Construct a table with the estimated coverage probabilities of the mean by each of the above one sided Confidence Intervals. Do some type of accounting for the Monte Carlo variability. Comment on how these compare, with respect to: (i) the usual notions about the t vs. normal theory, (ii) the theory (and discussion) discussed in class about the bootstrap, (iii) how well the "asymptotics have kicked in" in this setup.

Notes:

i. For each simulation run, one way to generate a matrix of bootstrap data, is to first generate a matrix (e.g. each column corresponds to one bootstrap replication) of "random indices", say `randind`. Then a very long vector of bootstrap samples is calculated in one line as `bootsamp = vdata(randind)`, where `vdata` contains the current data set. Use `reshape` to get a matrix of bootstrap replications.

ii. The reason for 1000 pseudo data sets (as opposed to 999), is that after the simulation runs we want probabilities, based on $F_n(x) = \frac{1}{n} \cdot \#(coverages)$, while for each bootstrap run we want quantiles.

iii. This will take a nonnegligible amount of time to run, so develop everything first with a smaller number of data sets, and bootstrap sample.

3. Analyze the bootstrap performances, in the case $n = 10$, by constructing histograms (for each of F1, F2, F3) of $n^{1/2}\left(\frac{\bar{X}^* - \bar{X}}{\hat{\sigma}}\right)$, $n^{1/2}\left(\frac{\bar{X}^* - \bar{X}}{\hat{\sigma}^*}\right)$, and $\hat{\sigma}^*$ for the 499 bootstrap replications of a single simulated data set (all on one page), and discuss lessons about the results in (2). Overlay the $N(0,1)$ density on each histogram, for easier comparison. Try changing the seed to see that this analysis feels the original strongly. I chose a seed that gave an "interesting", *but not valid*, result (serious further analysis requires simulation).