Inference and Uncertainty

Exercises

These exercises ask you to reproduce results similar to those shown in the lectures for the bootstrap, the M-out-of-N bootstrap, and subsampling. Here we will use two different data sets: data generated from a $\chi^2(3)$ distribution (as in the lectures), and a data set comprised of 1258 total column CO2 concentrations. The latter were produced by the Orbiting Carbon Observatory-2 Orbit Simulator in preparation for the operational mission, and represent synthetic measurements over North America during a generic June month. We will also use a different statistic here: the 95th percentile ($q_{.95}$), instead of the median ($q_{.50}$) so that we can see how well these methods work on extremes.

1. How well do the three sampling distributions corresponding to the ordinary bootstrap, the M-out-of-N bootstrap, and subsampling agree with (an approximation to) the true sampling distribution of $q_{.95}$ when the parent distribution is $\chi^2(3)$?

   (a) Simulate an approximation to the true sampling distribution of $\hat{q}_{.95}$.

      i. Write a function to draw 1000 samples of size $N = 1258$ from a $\chi^2(3)$ distribution. The reason for this choice of N will be apparent below.

      ii. For each sample, compute $\hat{q}_{.95}$.

   (b) Randomly generate a single sample of size $N = 1258$ from the $\chi^2(3)$ distribution. Write functions to simulate the sampling distribution of $\hat{q}_{.95}$ with the ordinary bootstrap, the M-out-of-N bootstrap, and subsampling from the sample you generated.

      i. These can all be done with `for`-loops.

      ii. Note that the resample/subsample sizes are different depending on which technique you are using. Use $M = N^{2/3}$ for the M-out-of-N bootstrap and subsampling.

      iii. Write these functions to be generic– you will need to use them below.

   (c) Make histograms of all four distributions. Make sure they are on the same scale.

   (d) How do the distributions differ? Which of the three resampling techniques produces a distribution most like the simulated true distribution?

2. Here we try out the three resampling techniques on real data, where we don't have the luxury of knowing the truth. We simply want to know whether the three methods produce similar results, and how sensitive they are to alternative parameter settings.

   (a) Load `XCO2.RData`. Make a histogram– this is our one sample. Compute $\hat{q}_{.95}$.

   (b) Use the functions you wrote above to simulate the sampling distribution of $\hat{q}_{.95}$ with the ordinary bootstrap, the M-out-of-N bootstrap, and subsampling. Note that $N = 1258$ as before. Use $M = N^{2/3}$ for the M-out-of-N bootstrap and subsampling, as before.

   (c) Make histograms all three distributions. Make sure they are on the same scale.

   (d) How do the distributions differ?

   (e) If there is time, repeat the analysis for other choices of M. Try anything you want, just be sure that $M/N \longrightarrow 0$ as $M, N \longrightarrow \infty$. An obvious choice might be $M = pN^{2/3}$ for various constants, p.