Semester 1 2012 Lecturer: Michael Stewart

Computer Exercise Week 8

Nonparametric bootstrap (i.e. simulation-from-best-guess) standard errors

Last week we obtained bootstrap (simulation-from-best-guess) standard errors in a situation where we were modelling the data as IID RVs from a *parametric* distribution, a distribution fully determined up to a single parameter θ, in that case a binomial "success probability" parameter. The simulation-from-best-guess was performed by first estimating the parameter and then simulating from the "estimated distribution".

In some cases, we are not able to assume that the data will be well-modelled by a member from a so-called parametric family of distributions (e.g. the $Bin(2, \theta)$ from last week). Sometimes all we can say is that the data $x_1, x_2, \ldots, x_n$ are modelled as values taken by IID RVs $X_1, X_2, \ldots, X_n$ whose distribution is (more-or-less) otherwise completely general. At our current level of sophistication this translates to a random sample *with replacement* from some finite population of real numbers $\mathcal{Y} = \{y_1, y_2, \ldots, y_N\}$ (think of an urn model where we randomly sample with replacement from an urn containing N balls, the j-th ball having a number y_j written on it). Then each observation has probability distribution given by

$$P(X_i = y_j) = \frac{1}{N}\ , \quad \text{for } j = 1, 2, \ldots, N.$$

In particular, the expected value and variance of each observation are given by $E(X_i) = \mu = \bar{y} = \frac{1}{N}\sum_{j=1}^{N} y_j$ (the arithmetic mean of the y_j's) and $Var(X_i) = \sigma^2 = \frac{1}{N}\sum_{j=1}^{N}(y_j - \bar{y})^2$ (the "population variance" of the y_j's). Such a model is sometimes called a *nonparametric* model, to stress the fact it is much more general and flexible than a restricted, so-called "parametric" model.

Given any such "population" Suppose the parameter we wish to estimate is $\theta = \sqrt{\sigma^2} = \sqrt{\frac{1}{N}\sum_{j=1}^{N}(y_j - \bar{y})^2}$, the *population standard deviation*. We could estimate it firstly by estimating σ^2 using, e.g. the unbiased estimator $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \bar{X})^2$, and then taking the square root, i.e. using S. However, what would be a standard error for the resulting estimate $s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \bar{x})^2}$?

If we *knew* $\mathcal{Y} = \{y_1, y_2, \ldots, y_N\}$ completely we could **either**

- analytically derive $\mathrm{MSE}(S) = E_{\mathcal{Y}}[(S - \sigma)^2]$ (this would be possible, although involved) **or**
- simulate B samples with replacement of size n from $\mathcal{Y}$, compute estimates $s_1^*, s_2^*, \ldots, s_B^*$, and then use the average squared error $\frac{1}{B}\sum_{k=1}^{B}(s_k - \sigma)^2$ to approximate the MSE;

once the MSE (or an approximation) is obtained, its square root would be the standard error. However, we don't know $\mathcal{Y}$ completely!

Again, the answer is to use a "best guess" for $\mathcal{Y}$, namely the **observed data itself**. Even for a moderately-sized sample the histogram of the data should look similar to the histogram for $\mathcal{Y}$. To implement the nonparametric bootstrap procedure therefore, we

1. obtain the estimate s from the original data
2. obtain B samples of size n *with replacement* from the original data; from each obtain a sample sd, yielding simulated estimates $s_1^*, s_2^*, \ldots, s_B^*$;
3. the final step is the compute square-root of the the average squared error, but the big question is **what are the "errors"?**
4. Answer: $\sqrt{\frac{1}{B}\sum_{k=1}^{B}(s_k^* - \sigma_x)^2}$ where $\sigma_x^2 = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x})^2$. Why? Because σ_x is "population standard deviation" of the population being sampled from here, which we could write as $\mathcal{X} = \{x_1, x_2, \ldots, x_n\}$.

To assess the performance of such a procedure, we shall use a "known population" and a sample from it, but for part of the exercise act as if the population was *not* known. This way we will be able to see how well this procedure actually performs.

1. Obtain the "population" using `pop=scan(url("http://www.maths.usyd.edu.au/stat2011/r/pop.txt"))`.

2. Set your random-number generator to always use the same random numbers:

   ```
   sid=...               ## put your 9-digit SID here
   set.seed(sid)
   ```

3. Compute the population standard devation: `sig.pop=sqrt(mean((pop-mean(pop))^2))`.

4. Obtain a random sample **with replacement** of size 50 from `pop`, calling it `samp`.

5. Supposing that you didn't know the population, only the sample, estimate the population standard deviation using the square-root of the (unbiased) sample variance: `est = sd(samp)`.

6. What is a standard error for this estimate? Since we know the population, we can actually approximate $\sqrt{MSE}$ arbitrarily accurately via simulation:

   ```
   B=10000
   errs=0
   for(j in 1:B){
    sim.samp=sample(pop,size=50,replace=T)
    errs[j]=sd(sim.samp)-sig.pop
   }
   sqrt(mean(errs^2))
   ```

   However in practice we would not know the true population. But can we *approximate* what we have done here?

7. The answer is: use a best guess to the population, and proceed in otherwise the same way. That best guess is the observed sample itself. The "population standard deviation" for this "population" is

   ```
   sig.samp=sqrt(mean((samp-mean(samp))^2))
   ```

   Repeat the above question but replacing `pop` with `samp` to get a *computable* (simulation-based) standard error for your estimate. **Comment** on the effectiveness of this procedure.