18.465 PROBLEM SET 6 DUE FRIDAY, APRIL 3, 2015

The topic is the bootstrap. There is material about it in Venables and Ripley, pp. 133–138. There has been a handout, which may be revised.

To apply the bootstrap in R you can call library(boot).

1. Suppose given a sample $x = (X_1, X_2, \cdots, X_n)$ of n distinct numbers. For example, they might be i.i.d. from some continuous distribution. As by default, consider bootstrap samples $(X_1^B, \cdots, X_n^B)$ also of size n. Find the probability that the X_j^B are also all distinct for
(a) $n = 5$,
(b) $n = 10$,
(c) Asymptotically as a function of n for $n \to \infty$, using Stirling's formula $n! \sim (n/e)^n \sqrt{2\pi n}$. *Note*: $a_n \sim b_n$ means $a_n/b_n \to 1$.

2. Continuing in the same situation as in Problem 1, arrange the bootstrap sample in order to get its order statistics $X_{(1)}^B \le X_{(2)}^B \le \cdots \le X_{(n)}^B$. For the original sample we have the order statistics $X_{(1)} < X_{(2)} < \cdots < X_{(n)}$.
(a) For each i and j with $1 \le i \le n$ and $1 \le j \le n$ find the probability that $X_{(i)}^B \le X_{(j)}$ in terms of binomial probabilities. Namely, in n independent trials with probability p of success on each, let $B(k, n, p)$ be the probability of k or less successes, and $E(k, n, p)$ the probability of k or more.
(b) Find the probability that $X_{(i)}^B = X_{(j)}$ by taking a difference of two probabilities from part (a).

3. Suppose $n = 2m + 1$ is odd. Then an order statistic of special interest is the sample median, which is $X_{(m+1)}$ for the original sample and $X_{(m+1)}^B$ for a bootstrap sample.
(a) From Problem 2(b), give the exact distribution of the bootstrap sample median, in terms of the original sample.
(b) For $n = 9$, give the distribution numerically. To compute binomial probabilities, R finds $B(k, n, p)$ as pbinom(k,n,p). You can find $E(k, n, p)$ as $1 - B(k - 1, n, p)$.

Date: 18.465,

(c) Generate 9 i.i.d. $U[0,1]$ variables U_j to get a data vector y=runif(9). Sort them to find their order statistics. In the sequence of bootstrap commands given at the bottom of p. 134 of Venables and Ripley, namely, for the bootstrap of the sample median of the given sample y with $R = 1000$ replications:

```
> library(boot)
> set.seed(101)
> y.boot = boot(y, function(x,i) median(x[i]), R = 1000)
> y.boot
```

We are estimating the median of the distribution of the bootstrap sample median (as if we didn't know it). Does the estimate agree with the true value we know from part (b)?
(d) Then as on p. 135 of Venables and Ripley type the further command

```
> boot.ci(y.boot, conf = c(0.90,0.95), type = c("norm","basic","perc","bca"))
```

to get some 90% and 95% confidence intervals.

(For the "bca" i.e. "bias-corrected, adjusted" intervals we do not yet know an explicit definition.) Which of these intervals contain: $X_{(5)}$; also $X_{(4)}$; $X_{(6)}$; any other order statistics $X_{(j)}$?
(e) We know that the true median of $U[0,1]$ is $1/2$. Which if any of the confidence intervals is $1/2$ in? (If we had a distribution whose true median we didn't know, the bootstrap method would have given us a confidence interval for it.) In general, narrower intervals would be preferred, but we want them to contain the true value with high probability (about $1-\alpha$).

4. We can estimate the true unknown median m by the sample median. Suppose we want a confidence interval for m and that the true distribution is continuous. We can get such an interval from the original order statistics $X_{(i)}$ without the bootstrap. Namely, a $1-\alpha$ or $100(1-\alpha)\%$ confidence interval for m is $[X_{(j)}, X_{(k)}]$ for the largest j such that $P(X_{(j)} > m) \le \alpha/2$ and the smallest k such that $P(X_{(k)} < m) \le \alpha/2$.
(a) Express the two probabilities, for general j and k, as binomial probabilities, using the fact that the distribution is continuous and so for an individual X_i (not an order statistic) $P(X_i < m) = P(X_i \le m) = 1/2$.
(b) For reasonably large n, the binomial probabilities are approximately normal. The symmetry resulting from $p = 1/2$ is favorable to the normal approximation. Recall that for the binomial (n,p) distribution the mean is np and the standard deviation is $\sqrt{np(1-p)}$. For $\alpha = 0.05$, give approximations to j and k using normality.

(c) For $n = 99$ and $\alpha = 0.05$, find j and k exactly using binomial probabilities. (They may not be either of the nearest integers to those found by normal approximation, but they shouldn't be far off.)
(d) What we got was a "conservative" confidence interval. For the interval $[X_{(j+1)}, X_{(k-1)}]$ the probability that m is in the interval will be less than $1 - \alpha = 0.95$, but adopt this interval or the one from earlier parts, depending on which gives a probability of containing m closer to $1 - \alpha$. (One might come closer with $[X_{(j)}, X_{(k-1)}]$ or $[X_{(j+1)}, X_{(k)}]$. Check that the probabilities for these to contain m are identical, so it would be arbitrary which of them to choose.)

5. For any probability distribution P on the real line having a finite second moment $\int x^2 dP(x)$, let $T(P)$ be the variance of P. Recall that given a sample $X_1, ..., X_n$, the usual (unbiased) sample variance is $s_X^2 = \frac{1}{n-1}\sum_{j=1}^n (X_j - \overline{X})^2$ for $n \geq 2$ and another sample variance is $s_X'^2 = \frac{1}{n}\sum_{j=1}^n (X_j - \overline{X})^2$.
(a) For the empirical measure P_n from the given sample, find $T(P_n)$ in terms of the X_j. How does it relate to the sample variances?
(b) Let $n = 20$. Let $X_1, ..., X_n$ be i.i.d. $N(\mu, \sigma^2)$ for some unknown real μ and some $\sigma > 0$. One can get confidence intervals for μ and σ^2 by classical methods, without the bootstrap. For σ^2, the intervals are based on the fact that $\sum_{j=1}^n (X_j - \overline{X})^2/\sigma^2$ has a $\chi^2(n-1)$ distribution where in this case $n - 1 = 19$.

Let's see what happens when we look for confidence intervals for $\sigma^2 = T(P)$ from the bootstrap point of view. The probability p that $T(P_n)/T(P) < 1/2$ equals the probability that $\chi^2(19) < k$ for what k? (Since it's a continuous distribution, "$<$" and "$\leq$" give the same probability.)
(c) One can find $p = \Pr(\chi^2(d) \leq x)$ for any positive integer d and $x > 0$ in R as pchisq(x, d) in R. Do that for the probability p in part (b).
(d) If it does happen that $T(P_n) < T(P)/2$, and $0 < \alpha < 1/2$, then the probability that $T(P)$ is in the "basic" bootstrap $1 - \alpha$ confidence interval given P_n is nominally $1 - \alpha$, but what can one say about the actual value?
(e) Usually, as α decreases toward 0 and $1 - \alpha$ increases toward 1, the "coverage probability," namely the probability that the parameter being estimated (in this case $\sigma^2 = T(P)$) is in the confidence interval increases toward 1, but what happens in this case?
(f) (extra credit) Suppose instead of the basic intervals we consider the percentile intervals in parts (d) and (e), then what happens?