Statistics 583, Problem Set 6 Solutions

Wellner; 5/3/2006

1. Show that for linear statistics the jackknife and bootstrap estimates of bias are zero. (A linear statistic corresponds to a functional $T(F)$ of the form $T(F) = \int \psi dF$ for some function ψ for a distribution function F on R, or $T(P) = \int \psi dP$ for a probability distribution P on a general sample space $\mathcal{X}$.)

   Solution: First note that for a linear statistic, assuming that $E_F|\psi(X)| = \int |\psi(x)|dF(x) < \infty$, we have

$$E_F T(\mathbb{F}_n) = E\left(\frac{1}{n}\sum_{i=1}^n \psi(X_i)\right) = E\psi(X_1) = \int \psi dF = T(F)\,,$$

   so, with $T_n \equiv T(\mathbb{F}_n)$,

$$\text{bias}(F) \equiv E_F(T_n) - T(F) = 0\,.$$

   The bootstrap estimator of bias(F) is

$$\begin{aligned}
\text{bias}(\mathbb{F}_n) &= E_{\mathbb{F}_n}(T_n^*) - T(\mathbb{F}_n) \\
&= E_{\mathbb{F}_n}\left(\frac{1}{n}\sum_{i=1}^n \psi(X_i^*)\right) - T(\mathbb{F}_n) \\
&= \frac{1}{n}\sum_{i=1}^n E_{\mathbb{F}_n}\{\psi(X_i^*)\} - T(\mathbb{F}_n) \\
&= \frac{1}{n}\sum_{i=1}^n \frac{1}{n}\sum_{j=1}^n \psi(X_j) - T(\mathbb{F}_n) \\
&= T(\mathbb{F}_n) - T(\mathbb{F}_n) = 0\,,
\end{aligned}$$

   so the bootstrap estimator of bias is correct. For the jackknife, we first note that the pseudo-values, $T_{n,i}^*$ are given by

$$T_{n,i}^* \equiv nT_n - (n-1)T_{n,i} = \psi(X_i)$$

   and hence

$$\overline{T}_n^* \equiv n^{-1}\sum_{i=1}^n T_{n,i}^* = n^{-1}\sum_{i=1}^n \psi(X_i) = T_n \equiv T(\mathbb{F}_n)\,.$$

   Thus the jackknife estimate of bias is

$$\widehat{\text{bias}}_n \equiv T_n - \overline{T}_n^* = 0\,.$$

   Thus the jackknife estimate of bias is also correct.

2. (a) Given n distinct data items, show that the probability that a given data item does not appear in a bootstrap sample is $e_n = (1 - 1/n)^n$
(b) Show that $e_n \to e^{-1} \approx .368$ as $n \to \infty$.
(c) Hence show that the probability that each of B bootstrap samples contains an item i is $(1 - e_n)^B$. Evaluate this quantity for $n = 10, 20, 50, 100$ and $B = 10, 20, 50, 100$.
(d) Let $N_n \equiv \sum_{j=1}^{n} 1_{[M_j=0]}$ where $\underline{M} \equiv (M_1, \ldots, M_n) \sim \text{Mult}_n(n, \underline{1}/n)$. Show that $E(n^{-1}N_n) = e_n$ as computed in (a).

Solution: (a) The probability that X_i does not appear in a bootstrap sample $X_1^*, \ldots, X_n^*$ from $\mathbb{F}_n$ is just $e_n = P(M_i = 0)$ where $M_i \sim \text{Binomial}(n, 1/n)$. Thus we have $e_n = P(M_i = 0) = \binom{n}{0}(1/n)^0(1 - 1/n)^n = (1 - 1/n)^n$.
(b) Since $(1+x/n)^n \to e^x$ for any x, it follows immediately that $e_n \to e^{-1} \approx .368$.
(c) The probabality that each of B bootstrap samples contains X_i is clearly $(1 - e_n)^B$. The following table gives values of this for $n = 10, 20, 50, 100$ and $B = 10, 20, 50, 100$.

B/n	10	20	50	100
10	.0137	.0118	.0108	.0105
20	.000189	.000139	.000117	.000110
50	4.89×10^{-10}	2.29×10^{-10}	1.47×10^{-10}	1.27×10^{-10}
100	2.39×10^{-19}	5.26×10^{-20}	2.16×10^{-20}	1.61×10^{-20}

(d) N_n/n is the proportion of the original sample not appearing in the bootstrap sample. Since $N_n \equiv \sum_{j=1}^{n} 1_{[M_j=0]}$ where each M_j is marginally Binomial$(n, 1/n)$, it follows immediately that

$$E(N_n/n) = P(M_1 = 0) = (1 - 1/n)^n \to e^{-1}\,.$$

Furthermore, from occupancy theory for urn models,

$$\sqrt{n}(n^{-1}N_n - (1 - 1/n)^n) \to_d N(0, e^{-1}(1 - 2e^{-1}))\,;$$

see e.g. Johnson and Kotz (1977), page 317.

3. Suppose that $T(F) = Var_F(X)$ so that $T_n \equiv T(\mathbb{F}_n) = n^{-1}\sum_{i=1}^{n}(X_i - \overline{X})^2$. Show that the jackknife estimate of the variance $\sigma_n^2(F) \equiv Var_F(T_n)$ is

$$\widehat{Var} = \frac{n^2}{(n-1)^3}(\widehat{\mu}_4 - \widehat{\mu}_2^2)$$

where $\widehat{\mu}_k \equiv n^{-1}\sum_{i=1}^n (X_i - \overline{X})^k$ for $k = 1, 2, \ldots$. Hence, assuming that $EX^4 < \infty$, the jackknife estimate of variance is consistent for this T:

$$n\widehat{Var} \to_p \mu_4 - \mu_2^2 = \mu_2^2\{2 + \frac{\mu_4}{\mu_2^2} - 3\} = T_2(F)(2 + \gamma_2)\,.$$

Solution: If $T_n = n^{-1}\sum_{j=1}^n (X_j - \overline{X})^2$, then some algebra yields

$$T_{n,i}^* = nT_n - (n-1)T_{n,i} = \frac{n}{n-1}(X_i - \overline{X})^2$$

and hence

$$\overline{T}_n^* = \frac{n}{n-1}\widehat{\mu}_2\,.$$

Furthermore,

$$\begin{aligned}
\widehat{\mathrm{Var}}_n &= \frac{1}{n(n-1)}\sum_{i=1}^n (T_{n,i}^* - \overline{T}_n^*)^2 \\
&= \frac{1}{n(n-1)}\sum_{i=1}^n T_{n,i}^{*2} - \frac{1}{n-1}(\overline{T}_n^*)^2 \\
&= \frac{1}{n(n-1)}\sum_{i=1}^n (\frac{n}{n-1}(X_i - \overline{X})^2)^2 - \frac{1}{n-1}(\frac{n}{n-1}\widehat{\mu}_2)^2 \\
&= \frac{1}{n-1}\frac{n^2}{(n-1)^2}\frac{1}{n}\sum_{i=1}^n (X_i - \overline{X})^4 - \frac{1}{n-1}\frac{n^2}{(n-1)^2}\widehat{\mu}_2^2 \\
&= \frac{n^2}{(n-1)^3}(\widehat{\mu}_4 - \widehat{\mu}_2^2)\,.
\end{aligned}$$

Thus we have

$$n\widehat{\mathrm{Var}}_n = \frac{n^3}{(n-1)^3}(\widehat{\mu}_4 - \widehat{\mu}_2^2)\,. \to_p \mu_4 - \mu_2^2 = \mu_2^2\{2 + \frac{\mu_4}{\mu_2^2} - 3\} = T_2(F)(2 + \gamma_2)\,.$$

where the (excess of) kurtosis is

$$\gamma_2 \equiv \frac{\mu_4}{\mu_2^2} - 3\,.$$

4. (a) Wasserman, problem 3.8.9, page 40: Let $X_1, \ldots, X_n$ be distinct observations (no ties). Let $X_1^*, \ldots, X_n^*$ denote a bootstrap sample and let $\overline{X}_n^* = n^{-1}\sum_{i=1}^n X_i^*$. Find $E(\overline{X}_n^*|X_1, \ldots, X_n)$, $Var(\overline{X}_n^*|X_1, \ldots, X_n)$, $E(\overline{X}_n^*)$, and $Var(\overline{X}_n^*)$.
(b) Wasserman, problem 3.8.13, page 41: let $X_1, \ldots, X_n$ be i.i.d. with distribution function F and empirical d.f. $\mathbb{F}_n$. Let $X_1^*, \ldots, X_n^*$ denote a bootstrap

sample, i.e. i.i.d. from $\mathbb{F}_n$. Let G denote the marginal distribution of X_i^*. Note that $G(x) = P(X_i^* \leq x) = E\{P(X_i^* \leq x|X_1, \ldots, X_n)\} = E\{\mathbb{F}_n(x)\} = F(x)$. So it appears that X_i^* and X_i have the same distribution. But in part (a) above we showed that $Var(\overline{X}_n) \neq Var(\overline{X}_n^*)$. This appears to be a contradiction. Explain.

Solution: (a) First, as we computed in class,

$$E(\overline{X}_n^*|X_1, \ldots, X_n) = n^{-1}\sum_{i=1}^n E(X_i^*|X_1, \ldots, X_n) = n^{-1}\sum_{i=1}^n \overline{X}_n = \overline{X}_n.$$

Similarly,

$$\begin{aligned} Var(\overline{X}_n^*|X_1, \ldots, X_n) &= n^{-2}\sum_{i=1}^n Var(X_i^*|X_1, \ldots, X_n) = n^{-2}\sum_{i=1}^n n^{-1}\sum_{j=1}^n (X_i - \overline{X}_n)^2 \\ &= n^{-1}S_n^2 \end{aligned}$$

where $S_n^2 = n^{-1}\sum_1^n (X_i - \overline{X}_n)^2$. Then it follows that

$$E(\overline{X}_n^*) = E\{E(\overline{X}_n^*|X_1, \ldots, X_n)\} = E\{\overline{X}_n\} = \mu(F),$$

and

$$\begin{aligned} Var(\overline{X}_n^*) &= E\{Var(\overline{X}_n^*|X_1, \ldots, X_n)\} + Var(E(\overline{X}_n^*|X_1, \ldots, X_n)) \\ &= E\{n^{-1}S_n^2\} + Var(\overline{X}_n) \\ &= n^{-1}\frac{n-1}{n}\sigma^2(F) + n^{-1}\sigma^2(F) \\ &= n^{-1}\sigma^2(F)\left\{\frac{n-1}{n} + 1\right\} \\ &= n^{-1}\sigma^2(F)\frac{2n-1}{n} = \frac{2n-1}{n}Var(\overline{X}_n) \approx 2Var(\overline{X}_n). \end{aligned}$$

Another way to organize this calculation is as follows:

$$\begin{aligned} Var(\overline{X}_n^*) &= Var(\overline{X}_n^* - \overline{X}_n + \overline{X}_n - \mu + \mu) \\ &= Var(\overline{X}_n^* - \overline{X}_n) + Var(\overline{X}_n - \mu) + 2Cov(\overline{X}_n^* - \overline{X}_n, \overline{X}_n - \mu) \\ &= E\{Var(\overline{X}_n^* - \overline{X}_n)|X_1, \ldots, X_n)\} + Var\{E(\overline{X}_n^* - \overline{X}_n)|X_1, \ldots, X_n)\} \\ &\quad + n^{-1}\sigma^2(F) \\ &\quad + 2E\{(\overline{X}_n^* - \overline{X}_n)(\overline{X}_n - \mu)\} \\ &= E\{n^{-1}S_n^2\} + 0 + n^{-1}\sigma^2(F) + 0 \\ &= \frac{n-1}{n}n^{-1}\sigma^2(F) + \sigma^2(F), \end{aligned}$$

so that the contributions of $Var(\overline{X}^*_n - \overline{X}_n)$ and $Var(\overline{X}_n - \mu)$ are approximately equal. This is important, especially since we want the first of these to estimate the second! [The marginal behavior of bootstrap estimators is largely irrelevant, since this accounts for both the deviations $T(\mathbb{F}^*_n) - T(\mathbb{F}_n)$ and $T(\mathbb{F}_n) - T(F)$ via $T(\mathbb{F}_n) - T(F) = T(\mathbb{F}^*_n) - T(\mathbb{F}_n) + T(\mathbb{F}_n) - T(F)$. What is important is that $T(\mathbb{F}^*_n) - T(\mathbb{F}_n)$ mimics (or estimates) $T(\mathbb{F}_n) - T(F)$!

(b) The marginal distribution of the X^*_i's (separately, or marginally) agrees with the marginal distribution of the (separate) X_i's, but the *joint distribution* of the X^*_i's is dependent. For example,

$$\begin{aligned}
G_2(x_1, x_2) &\equiv P(X^*_1 \le x_1, X^*_2 \le x_2) = E\{P(X^*_1 \le x_1, X^*_2 \le x_2 | X_1, \ldots, X_n)\} \\
&= E\{\mathbb{F}_n(x_1)\mathbb{F}_n(x_2)\} = E\{n^{-2}\sum_{i=1}^n \sum_{j=1}^n 1\{X_i \le x_1\}1\{X_j \le x_2\} \\
&= E\{n^{-2}[\sum_{i=1}^n 1\{X_i \le x_1\}1\{X_i \le x_2\} + \sum_{i \ne j} 1\{X_i \le x_1\}1\{X_i \le x_2\}]\} \\
&= n^{-2}\{nF(x_1 \wedge x_2) + n(n-1)F(x_1)F(x_2)\} \\
&= (1 - n^{-1})F(x_1)F(x_2) + n^{-1}F(x_1 \wedge x_2).
\end{aligned}$$

Thus $G_2(x_1, x_2)$ is a mixture of the independence distribution $F(x_1)F(x_2)$ and the (Fréchet bound) distribution concentrated on the diagonal, $F(x_1 \wedge x_2)$, with mixing probabilities $(1 - n^{-1})$ and n^{-1}. More generally,

$$\begin{aligned}
G(x_1, \ldots, x_n) &\equiv P(X^*_1 \le x_1, \ldots, X^*_n \le x_n) \\
&= E\{P(X^*_1 \le x_1, \ldots, X^*_n \le x_n | X_1, \ldots, X_n)\} \\
&= E\{\mathbb{F}_n(x_1) \cdots \mathbb{F}_n(x_n)\} = E\{\prod_{k=1}^n \mathbb{F}_n(x_k)\} \\
&= E\left\{ n^{-n} \sum_{j_1=1}^n \cdots \sum_{j_n=1}^n \prod_{k=1}^n 1\{X_{j_k} \le x_k\} \right\} \\
&= n^{-n} \sum_{j_1=1}^n \cdots \sum_{j_n=1}^n E\left\{ \prod_{k=1}^n 1\{X_{j_k} \le x_k\} \right\}
\end{aligned}$$

where now the computation becomes somewhat complicated. In particular, though X^*_1 and X^*_2 are correlated: from our formula (1.4.14) on page 19 of Chapter 1

(with $G(x) = H(x) = x$, $F = G_2$, and $F_X = F_Y = F$), it follows that

$$\begin{aligned} Cov(X_1^*, X_2^*) &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\{G_2(x,y) - F(x)F(y)\}dxdy \\ &= \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\{(1-n^{-1})F(x)F(y) + n^{-1}F(x\wedge y) - F(x)F(y)\}dxdy \\ &= n^{-1}\int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\{F(x\wedge y) - F(x)F(y)\}dxdy \\ &= n^{-1}\sigma^2(F) \end{aligned}$$

where the last equality follows from (1.14.16)-(1.14.17) on page 19, chapter 1. Thus

$$Var(\overline{X}_n^*) = n^{-2}\{n\sigma^2(F) + n(n-1)(n^{-1}\sigma^2(F))\} = n^{-1}\sigma^2(F)\{1 + \frac{n-1}{n}\}$$

in agreement with our calculation in (a).