Statistics 583, Problem Set 6 Solutions

Wellner; 5/13/2009

1. Wasserman, example 3.10, page 29.
   (a) In the 3rd line of this example, Wasserman says that the jackknife estimate of the standard error of the skewness estimator for the nerve data is .17. Check this. (Show your code or method of calculation.)
   (b) I claimed in the solution to problem set #5, problem 1 (see page 3 of the solution set), that the estimated standard error using the delta method is .163 rather than .18 as Wasserman claimed. Check this. (Show your code or method of calculation.)
   (c) On page 31, Wasserman claims that the bootstrap based on $B = 10^3$ replications gives .16 as a bootstrap estimate of the standard error. Try it yourself to see what you get. (Show your code or method of calculation.)
   The data is posted in two forms at:
   http://www.stat.washington.edu/jaw/COURSES/580s/583/sp09.probsets.html/nrvdat
   and
   http://www.stat.washington.edu/jaw/COURSES/580s/583/sp09.probsets.html/nerve.dat.
   (d) Plot the empirical distribution function of this data together with the MLE of the distribution function assuming: (i) an exponential distribution; (ii) a Weibull distribution. What alternative methods for (a) and (b) does this suggest?

   Solution: (a) When I calculate the jackknife estimator of the standard error of the skewness estimator for the nerve data, I get .172; see the posted Mathematica notebook *NerveDatAnal-2.nb*.
   (b) As mentioned before, I get .163 for the delta-method estimator of the standard error; see the posted Mathematica notebook *NerveDatAnal-1.nb* which also computes the delta-method estimator as in Wasserman (without the additional term in the influence function) to get .179.
   (c) With a bootstrap sample size of 10^4 I get .1626 as the bootstrap estimator of standard error of the skewness estimator, almost exactly the same as the (corrected) delta method estimator; see the posted Mathematica notebook *NerveDatAnal-3.nb*.

(d) From the plot of the empirical distribution function of the nerve data on page 14 of Wasserman, one might guess that this data could be modeled by an exponential distribution. The skewness of the exponential distribution is 2, and it should be noted that all of the 95% confidence intervals for skewness given by Wasserman on page 34 include 2. On the other hand, a histogram of the nerve data (Figure 1) shows a slight dip in the frequency relative to exponential in the cell just to the right of zero, and it turns out that a Weibull distribution with (α, β) estimated by $(\hat{\alpha}, \hat{\beta}) = (.22557, 1.08181)$ via maximum likelihood gives an excellent fit. Under the Weibull model the estimated skewness is 1.7778... (to be compared to the sample skewness of 1.761.... See the posted Mathematica notebook *NerveDataAnal4ML.nb*. Thus parametric inference and parametric bootstrap methods based on the Weibull model should perform quite well here.

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:

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

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).

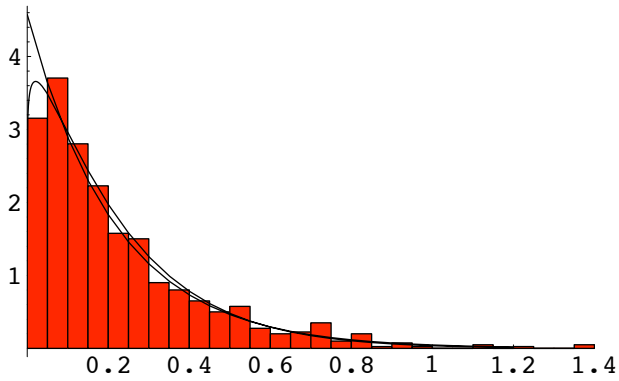

Figure 1: Histogram of the nerve data with exponential and Weibull fitted densities (via MLE).

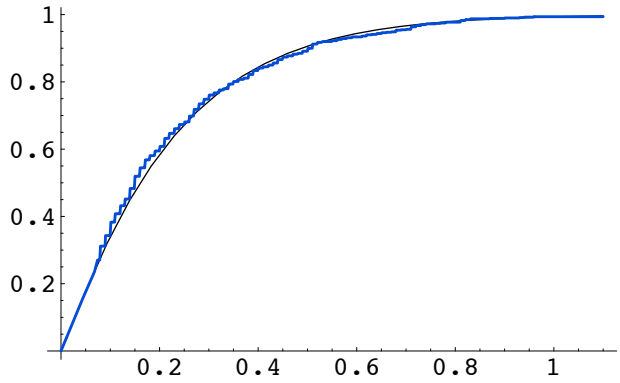

Figure 2: Empirical distribution function (blue) and fitted Weibull c.d.f. (via MLE).