Statistics 583, Problem Set 7 Solutions

Wellner; 5/15/2006

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 4 (see page 9 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/sp06.probsets.html/nrvdat
and
http://www.stat.washington.edu/jaw/COURSES/580s/583/sp06.probsets.html/nerve.dat.

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*.

Postscript: 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 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*.

2. The expression for the jackknife variance estimator for the median, in the display (1) on page 11 (3rd line from the bottom) in chapter 8 was derived under the assumption $n = 2m$ and that $T(\mathbb{F}_n) = X_{(m)}$ if $n = 2m-1$, $T(\mathbb{F}_n) = (X_{(m)} + X_{(m+1)})/2$ if $n = 2m$.
(a) Derive the first equality in (1), page 11, using this definition of the sample median.
(b) Derive versions of (2.2) using $T(F) = F^{-1}(1/2)$ (strictly). Does the asymptotic result in (1) still hold? Here is some further explanation of what I mean by "strictly" here: let $T_1(\mathbb{F}_n) = X_m$ if $n = 2m-1$, $T_1(\mathbb{F}_n) = (X_{(m)} + X_{(m+1)})/2$ if $n = 2m$. This is one common definition of the median, and this is the definition used in (a). Let $T_2(\mathbb{F}_n) = \mathbb{F}_n^{-1}(1/2)$. This is my favorite definition of the median. Note that $T_2(\mathbb{F}_n) = T_1(\mathbb{F}_n)$ if $n = 2m-1$, but $T_2(\mathbb{F}_n) \neq T_1(\mathbb{F}_n)$ if $n = 2m$. (What is the value of $T_2(\mathbb{F}_n)$ in this case?) T_2 is the definition of the median to be considered in 2(b)!

Solution: (a). For $n = 2m$,

$$T_{n,i} = \begin{cases} X_{(m+1)} & \text{if} \quad i \leq m \\ X_{(m)} & \text{if} \quad i > m \end{cases}$$

and $T_{n,\cdot} = (X_{(m)} + X_{(m+1)})/2$. Hence

$$\begin{aligned} n\widehat{\mathrm{Var}}_n &= (n-1)\Big\{m(X_{(m+1)} - \frac{1}{2}(X_{(m)} + X_{(m+1)}))^2 \\ &\qquad + \; m(X_{(m)} - \frac{1}{2}(X_{(m)} + X_{(m+1)}))^2\Big\} \\ &= n(n-1)\{\frac{X_{(m+1)} - X_{(m)}}{2}\}^2 . \end{aligned} \tag{1}$$

(b). When $n = 2m$ and $T(F) = F^{-1}(1/2)$, we have $T(\mathbb{F}_n) = X_{(m)}$ and $T_{n,i}$ are exactly as in A above. Hence (1) continues to hold.
When $n = 2m-1$, then $T(\mathbb{F}_n) = X_{(m)}$,

$$T_{n,i} = \begin{cases} X_{(m)} & \text{if} \quad i \leq m-1 \\ X_{(m-1)} & \text{if} \quad i \geq m \end{cases} ,$$

and $T_{n,\cdot} = \{(m-1)X_{(m)} + mX_{(m-1)}\}/(2m-1)$. Therefore

$$\begin{aligned} n\widehat{\mathrm{Var}}_n &= (n-1)\left\{(m-1)\{X_{(m)} - \frac{1}{2m-1}[(m-1)X_{(m)} + mX_{(m-1)}]\}^2\right. \\ &\quad \left. + \; m\{X_{(m-1)} - \frac{1}{2m-1}[(m-1)X_{(m)} + mX_{(m-1)}]\}^2\right\} \\ &= \frac{(n-1)^2(n+1)}{n}\left\{\frac{X_{(m)} - X_{(m-1)}}{2}\right\}^2 \\ &\rightarrow_d \frac{1}{4f^2(F^{-1}(1/2))}\left(\frac{\chi_2^2}{2}\right)^2 \end{aligned}$$

just as before.

Remark: The only case left out in (a) and (b) is that of an odd sample size, $n = 2m-1$ in part (a). In this case,

$$T_{n,i} = \begin{cases} (X_{(m)} + X_{(m+1)})/2 & \text{if} \quad i \leq m-1 \\ (X_{(m-1)} + X_{(m+1)})/2 & \text{if} \quad i = m \\ (X_{(m-1)} + X_{(m)})/2 & \text{if} \quad i \geq m+1 \end{cases} .$$

Thus

$$\begin{aligned} T_{n,\cdot} &= \frac{1}{n}\left\{\frac{(m-1)}{2}(X_{(m)} + X_{(m+1)})\right. \\ &\quad \left. + \; \frac{1}{2}(X_{(m-1)} + X_{(m+1)}) + \frac{(m-1)}{2}(X_{(m-1)} + X_{(m)})\right\} . \end{aligned}$$

The analysis from this point proceeds not just by algebra, but by careful grouping of terms and observing which terms are negligible. I will not present a full analysis here, but will record the result:

$$\begin{aligned} n\widehat{\mathrm{Var}}_n &= \frac{(m-1)m^2}{2n^3}\left\{n(X_{(m+1)} - X_{(m-1)})\right\}^2 + o_p(1) \\ &\rightarrow_d \frac{1}{4f^2(F^{-1}(1/2))}(\frac{\chi_4^2}{4})^2 \end{aligned}$$

since, with $g \equiv F^{-1}$,

$$n(X_{(m+1)} - X_{(m-1)}) \rightarrow_d g'(1/2)W$$

where $W =_d Y_1 + Y_2 \sim \mathrm{Gamma}(2,1)$ for independent exponential rv's $Y_1,\ Y_2$, so that $2W \sim \chi_4^2$. Thus for this definition of the sample median, it is true that $n\widehat{\mathrm{Var}}_n = O_p(1)$ for the full sequence of nonnegative integers n but it converges in distribution to one limit as $n = 2m \rightarrow \infty$ and a different limit as $n = 2m-1 \rightarrow \infty$.