

## Statistics 583, Problem Set 7 Solutions

Wellner; 5/20/2009

1. 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 the development in (1), page 11, 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  $n = 2m$ ,

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

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

$$\begin{aligned} n\widehat{\text{Var}}_n &= (n-1) \left\{ m(X_{(m+1)} - \frac{1}{2}(X_{(m)} + X_{(m+1)}))^2 \right. \\ &\quad \left. + m(X_{(m)} - \frac{1}{2}(X_{(m)} + X_{(m+1)}))^2 \right\} \\ &= n(n-1) \left\{ \frac{X_{(m+1)} - X_{(m)}}{2} \right\}^2. \end{aligned} \tag{1}$$

(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  $T(\mathbb{F}_n) = X_{(m)}$ ,

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

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

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

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,

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

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:

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

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

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

where  $W =_d Y_1 + Y_2 \sim \text{Gamma}(2, 1)$  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  $n\widehat{\text{Var}}_n = O_p(1)$  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  $n = 2m-1 \rightarrow \infty$ .

2. (a) Wasserman, problem 3.8.3, page 39, modified. Show that the claimed expression for  $v_{boot}$  given in the display for this problem is incorrect and find the correct expression. Here  $v_{boot} = Var_{\mathbb{F}_n}(T_n)$  where  $T_n = \overline{X}_n^2$ . [Hint: see Dodd and Korn, *The American Statistician* **61** (2007), 127 - 131, and especially their appendix B, pages 130-131. Apparently the formula given by Wasserman in his problem is from Shao and Tu (1995), page 10; as noted by Dodd and Korn, the expression in Shao and Tu is incorrect.]
- (b) Explain how the resulting formulas relate to how you would estimate the variance of  $\overline{X}_n^2$  via the delta method.

**Solution:** (a) This is explained quite well in the appendix of the paper by Dodd and Korn (2007).

(b) The first term of the exact finite sample variance expression

$$Var(\overline{X}^2) = \frac{4\mu^2\sigma^2}{n} + \frac{2\sigma^4}{n^2} + \frac{4\mu\mu_3}{n^2} + \frac{\mu_4 - 3\sigma^4}{n^3}$$

corresponds exactly to what we would get from the delta method: with  $g(x) = x^2$  we have  $g'(x) = 2x$  and hence

$$\sqrt{n}(\overline{X}_n^2 - \mu^2) \rightarrow_d g'(\mu)\sigma Z \sim N(0, 4\mu^2\sigma^2)$$

where  $Z \sim N(0, 1)$ . Thus the delta-method estimator of  $Var(\overline{X}_n^2)$  is just  $4\overline{X}_n^2 S_n^2$  where  $S_n$  is the sample variance. The bootstrap estimator of variance refines this (as shown by Dodd and Korn) by correctly capturing the  $n^{-2}$  term when  $\mu \neq 0$ . When  $\mu = 0$ , then neither the (first order) delta method nor the (nonparametric) bootstrap tells the complete story.

3. Wasserman, problem 3.8.7, page 40: let  $X_1, \dots, X_n \sim t_3$  where  $n = 25$ . Let  $\theta = T(F) = (q_{.75} - q_{.25})/1.34$  where  $q_p$  denotes the  $p^{th}$  quantile. Do a simulation to compare the coverage and length of the following confidence intervals for  $\theta$ : (i) Normal interval with standard error from the bootstrap; (ii) Normal interval with standard error from the bootstrap; (iii) bootstrap percentile interval.

**Solution:** (i) The normal interval with standard error from the jackknife is given by

$$T(\mathbb{F}_n) \pm z_{\alpha/2} \widehat{se}_{jack}$$

where

$$\widehat{se}_{jack} = \sqrt{\frac{\widehat{\tilde{s}}_{jack}^2}{n}}$$

where

$$\hat{s}_{jack}^2 = \frac{1}{n-1} \sum_{i=1}^n \left( \tilde{T}_i - n^{-1} \sum_{j=1}^n \tilde{T}_j \right)^2$$

where  $\tilde{T}_i \equiv T_{n,i}^* \equiv nT_n - (n-1)T_{n-1,i}$ ,  $i = 1, \dots, n$  are the pseudo-values. See Wasserman page 28 and Chapter 8, section 3.

(ii) The normal interval with standard error from the bootstrap is given by

$$T(\mathbb{F}_n) \pm z_{\alpha/2} \hat{s}_{boot}$$

where

$$\hat{s}_{boot} = \sqrt{v_{boot}} = \sqrt{\frac{1}{B} \sum_{j=1}^B \left( T(\mathbb{F}_{n,j}^*) - \frac{1}{B} \sum_{k=1}^B T(\mathbb{F}_{n,k}^*) \right)^2}.$$

See Wasserman pages 30 and 32.

(iii) The percentile interval is given by the  $\alpha/2$  and  $(1-\alpha)/2$  quantiles of the bootstrap sample:

$$(T_{(B\alpha/2)}^*, T_{(B(1-\alpha)/2)}^*);$$

see Wasserman, page 34. The Mathematica code I wrote to do this problem is posted at

<http://www.stat.washington.edu/jaw/COURSES/580s/583/sp09.probsets.html>

When I use  $B = 1000$  and do  $MC = 500$ (try 1, try 2),  $MC = 1000$ (try 3) Monte-carlo replications of 95% confidence intervals, I get the following results for the three different confidence intervals when sampling from  $t_3$ , the Student  $t$ -distribution with 3 degrees of freedom. Note that if  $Y \sim t_3$ , then

Table 1:

population	Normal interval jackknife SE	Normal interval bootstrap SE	Bootstrap percentile
$t_3$ , try 1	0.84	0.956	0.978
$t_3$ , try 2	0.852	0.946	0.958
$t_3$ , try 3	0.851	0.958	0.973

$Var(Y) = 3/(3-2) = 3 < \infty$ . But if  $Y \sim t_2$ , then  $Var(Y) = \infty$ , and if  $Y \sim t_1$ , then both  $Var(Y) = \infty$  and  $E|Y| = \infty$ . In the following table we give coverages of the same jackknife and bootstrap confidence intervals as above, but for the situations when the population is changed to  $t_2$  or  $t_1$ .

Table 2:

population	Normal interval jackknife SE	Normal interval bootstrap SE	Bootstrap percentile
$t_1$ , try 1	0.768	0.974	0.976
$t_1$ , try 2	0.748	0.976	0.976
$t_2$ , try 1	0.864	0.970	0.988
$t_2$ , try 2	0.830	0.958	0.990

4. Wasserman, problem 3.8.11, page 41.

**Solution:** (a) Let  $F_\theta$  denote the Uniform(0,  $\theta$ ) distribution. The distribution of  $\hat{\theta}_n$  is just

$$\begin{aligned} P_\theta(\hat{\theta}_n \leq x) &= P_\theta(X_{(n)} \leq x) = P_\theta(X_1 \leq x, \dots, X_n \leq x) \\ &= P_\theta(X_1 \leq x)^n = (1 - x/\theta)^n, \quad 0 \leq x \leq \theta, \quad \text{so that} \\ f_{\hat{\theta}_n}(x) &= \frac{n}{\theta} \left(\frac{x}{\theta}\right)^{n-1} 1_{[0, \theta]}(x). \end{aligned}$$

Thus  $\hat{\theta}_n/\theta \stackrel{d}{=} \xi_{(n)}$ , the largest order statistic of a sample  $\xi_1, \dots, \xi_n$  of i.i.d. Uniform(0, 1) random variables, and  $n(1 - \hat{\theta}_n/\theta) \stackrel{d}{=} n\xi_{(1)} \rightarrow_d Y$  where  $Y \sim \text{exponential}(1)$ .

(b) The parametric bootstrap estimator of

$$K_n(x, F_\theta) \equiv P_\theta(\hat{\theta}_n \leq x) = (1 - x/\theta)^n, \quad 0 \leq x \leq \theta$$

is

$$K_n(x, F_{\hat{\theta}}) \equiv P_{\hat{\theta}}(\hat{\theta}_n^* \leq x) = (1 - x/\hat{\theta})^n, \quad 0 \leq x \leq \hat{\theta}.$$

The nonparametric bootstrap estimator is given by

$$\begin{aligned} K_n(x, \mathbb{F}_n) &= P_{\mathbb{F}_n}(\hat{\theta}_n^* \leq x) = P^*(X_{(n)}^* \leq x) \\ &= P^*(X_1^* \leq x, \dots, X_n^* \leq x) = P^*(X_1^* \leq x)^n = \mathbb{F}_n(x)^n. \end{aligned}$$

Thus

$$\begin{aligned} P_{\mathbb{F}_n}(\hat{\theta}_n^* = \hat{\theta}_n) &= P_{\mathbb{F}_n}(X_{(n)}^* = X_{(n)}) = \mathbb{F}_n(X_{(n)})^n - \mathbb{F}_n(X_{(n-1)})^n \\ &= 1 - \left(\frac{n-1}{n}\right)^n = 1 - (1 - 1/n)^n \rightarrow 1 - e^{-1}. \end{aligned}$$