Statistics 583, Problem Set 5 - revised

Wellner; 5/13/2015

Reading: Wasserman, Chapter 4, pages 43-60; Chapter 5, pages 61-80.
Van der Vaart, Chapter 24, pages 341-348.

Due: Wednesday, May 27, 2015

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)!

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.

3. Suppose that $X_1, \ldots, X_n$ are i.i.d. with density p on $\mathbb{R}$. The kernel density estimator $\hat{p}_n$ of p using the kernel k and bandwidth b is defined by

$$\hat{p}_n(x;b) = \frac{1}{b}\int k\left(\frac{x-y}{b}\right) d\mathbb{F}_n(y) = \frac{1}{nb}\sum_{i=1}^n k\left(\frac{x-X_i}{b}\right).$$

   We assume that $\int k(v)dv = 1$ and $k(v) \geq 0$.
   (a) For a fixed point $x \in \mathbb{R}$ calculate $E\hat{p}_n(x;b)$ and give an expression for the

bias, $E\hat{p}_n(x;b) - p(x)$.
(b) Again for a fixed x, calculate $Var(\hat{p}_n(x;b))$.
(c) Discuss how the bias and variance change as the bandwidth b increases.
(d) Assuming that $\int vk(v)dv = 0$ and $\int v^2k(v)dv = 1$, compute

$$\hat{\mu}_n = \int x\hat{p}_n(x)dx \quad \text{and} \quad \hat{\sigma}_n^2 \equiv \int (x - \hat{\mu}_n)^2\hat{p}_n(x)dx.$$

How do $\hat{\mu}_n$ and $\hat{\sigma}_n^2$ compare to $\overline{X}_n$ and S_n^2?

4. Silverman (1981, 1983; see the course handout page for copies of these two papers) proposed a test for the number of modes of a density. This test is discussed on pages 227-232 of Efron and Tibshirani (1998), *An Introduction to the Bootstrap*. Silverman's proposed test goes as follows: let $\hat{p}_n(\cdot;b)$ be the kernel estimator of p based on a standard Gaussian kernel k; i.e. $k(v) = \phi(v)$ where ϕ is the standard Gaussian density $\phi(v) = (2\pi)^{-1/2}\exp(-v^2/2)$. [Use of a Gaussian kernel is crucial in Silverman's test!] As b increases the density estimator $\hat{p}_n(\cdot;b)$ becomes smoother and has fewer modes. In fact for a Gaussian kernel, the number of modes is a monotone non-increasing function of the bandwidth b. See Figure 16.2 on page 228 of Efron and Tibshirani (1998) for an illustration of this. Consider testing H_0 : p has one mode versus H_1 : p has 2 or more modes. Since the number of modes decreases as b increases, there is a smallest value of b such that $\hat{p}(\cdot;b)$ has one mode. Call this $\hat{b}_1$. Now we use $\hat{p}_n(\cdot;\hat{b}_1)$ as the estimated null distribution for our test of H_0 versus H_1. As noted in Efron and Tibshirani, it seems reasonable to adjust $\hat{p}_n(\cdot;\hat{b}_1)$ slightly to adjust for the fact noted in problem 3(d) above that the variance under $\hat{p}_n(\cdot;\hat{b}_1)$ is somewhat larger than the sample variance. We call the resulting estimator $\hat{q}_n(\cdot;\hat{b}_1)$. A reasonable test statistic is $\hat{b}_1$: if this is large, then a greater amount of smoothing is required to obtain one mode, and this supports the alternative hypothesis. Now the test is carried out via bootstrap resampling from the fitted model under the null hypothesis; see Efron and Tibshirani (1998) for details.
(a) Describe this bootstrap testing procedure from the perspective of estimation of some functional of the true distribution and our discussion in sections 8.2 and 8.3, distinguishing carefully between the ideal bootstrap and the Monte-Carlo implementation of the bootstrap.
(b) Verify that the resampling scheme outlined on page 232 of Efron and Tibshirani accomplishes the desired adjustment of $\hat{p}_n(\cdot;\hat{b}_1)$ so that the resulting $\hat{q}_n(\cdot;\hat{b}_1)$ has variance very nearly equal to the sample variance.
(c) Find at least one alternative test of multimodality of a univariate density p that has been proposed since (1983).

5. **Optional bonus problem:** Consider a two-sample testing problem with $X_1, \ldots, X_m$ i.i.d. F and $Y_1, \ldots, Y_n$ i.i.d. G. Consider testing:
(1) $H : F = G$ versus $K_{Gaussian,1}$: $F(x) = \Phi((x-\mu)/\sigma)$, $G(y) = \Phi((y-\nu)/\sigma)$, $\mu \neq \nu$.
(2) $H : F = G$ versus $K_{Gaussian,2}$: $F(x) = \Phi((x-\mu)/\sigma)$, $G(y) = \Phi((y-\nu)/\tau)$, $\mu \neq \nu$, $\sigma \neq \tau$.
(3) $H : \mu_F = \mu_G$, F, G otherwise unknown, versus $K_{Gaussian,2}$ as in (2).

(a) Discuss appropriate bootstrap testing procedures for these three testing problems, including identification of which (estimated) distribution is involved for the resampling procedure.
(b) In which problems is a permutation test appropriate?
(c) Which of the three problems involves the largest null hypothesis?
Hint: see Efron and Tibshirani (1998), sections 16.1-16.3.

6. **Optional bonus problem:** (Hard!) On page 12, line 4 of Chapter 8 of the lecture notes, it is claimed that if $E_F|X|^r < \infty$ for some $r > 0$ and $f(F^{-1}(1/2)) > 0$, then for the median function $T(F) = F^{-1}(1/2)$ we have

$$nVar_F(T(\mathbb{F}_n)) \to \frac{1/4}{f^2(F^{-1}(1/2))}.$$

Prove (or disprove) this claim.