**Statistics 581**

**Problem Set 5**

Wellner; 10/22/2008

**Reading:** Ferguson, ACLST, Chapters 13 and 14, pages 87 - 100;
Wellner Notes, Chapter 2, sections 4 - 6.
**Due:** Wednesday, October 29, 2008.
**Reminder:** Midterm exam, Monday, November 3, 2008.

1. Verify the following claim made in our treatment of the asymptotic distribution of the sample correlation coefficient: if

$$(X,Y) \sim N_2\left(\underline{0}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right),$$

then

$$\begin{pmatrix} E(X^2Y^2)-\rho^2 & E(X^3Y)-\rho & E(XY^3)-\rho \\ E(X^3Y)-\rho & E(X^4)-1 & E(X^2Y^2)-1 \\ E(XY^3)-\rho & E(X^2Y^2)-1 & E(Y^4)-1 \end{pmatrix} = \begin{pmatrix} 1+\rho^2 & 2\rho & 2\rho \\ 2\rho & 2 & 2\rho^2 \\ 2\rho & 2\rho^2 & 2 \end{pmatrix}.$$

**Hint:** Compute conditionally and use Theorem 1.3.5, page 14, Chapter 1.

2. For a random variable $X$ taking values in $(0,\infty)$ define the *entropy* $H$ of $X$ by $H(X) \equiv E(X \log X) - E(X)\log(E(X))$.
(a) Show that $H(X) \ge 0$.
(b) Compute $H(X)$ and $H(X)/E(X)$ for $X \sim \text{Exp}(\lambda)$.
(c) Suppose that $X, X_1, \ldots, X_n$ are i.i.d. Propose an estimator of $H(X)$ and give conditions under which your estimator, say $\hat{H}_n$, satisfies $\hat{H}_n \to_p H(X)$.
(d) Give conditions under which the estimator you proposed in (c) satisfies $\sqrt{n}(\hat{H}_n - H(X)) \to_d N(0, V^2)$ and find $V^2$.

3. (a) Write out a proof of (10) on page 16 of the Chapter 2 notes.
(b) Write out a proof of the corresponding fact concerning the general empirical process $\mathbb{G}_n$: $\mathbb{G}_n \to_{f.d.} \mathbb{G}$ where $\mathbb{G}_n$ and $\mathbb{G}$ are as defined on page 21 of the chapter 2 notes; i.e. for any $f_1, \ldots, f_k \in L_2(P)$, $(\mathbb{G}_n(f_1), \ldots, \mathbb{G}_n(f_k)) \to_d (\mathbb{G}(f_1), \ldots, \mathbb{G}(f_k))$.

4. Ferguson, ACILST, problem 6, page 93.

5. Suppose that $X_1, \ldots, X_n$ are i.i.d. random vectors with values in $R^k$ with $E(X_1) = \mu$ and $E(X_1^T X_1) < \infty$ so that $\Sigma = E(X_1-\mu)(X_1-\mu)^T$ is well-defined. Thus

$$Z_n \equiv \sqrt{n}(\overline{X}_n - \mu) \to_d Z \sim N_k(0, \Sigma).$$

Suppose that $g : R^k \to R$ is a function, and suppose that $\nabla g = \dot{g}$ exists at $\mu$. Then the delta-method (or $g'$ theorem) tells us that

$$\sqrt{n}(g(\overline{X}_n) - g(\mu)) \to_d \nabla g(\mu)^T Z \sim N(0, \nabla g(\mu)^T \Sigma \nabla g(\mu)). \tag{1}$$

(a) Show that we can strengthen (1) as follows: Suppose that $\nabla g = \dot{g}$ is continuous at $\mu$. Then $\sqrt{n}(g(\overline{X}_n) - g(\mu))$ is *asymptotically linear* at $\mu$:

$$\begin{aligned}\sqrt{n}(g(\overline{X}_n) - g(\mu)) &= \nabla g(\mu)^T \sqrt{n}(\overline{X}_n - \mu) + o_p(1) \\ &= \frac{1}{\sqrt{n}} \sum_{i=1}^n \psi(X_i) + o_p(1)\end{aligned}$$

where

$$\psi(x) = \nabla g(\mu)^T (x - \mu)$$

which is called the *influence function* of $g(\overline{X}_n)$ as an estimator of $g(\mu)$, has mean $E\psi(X_i) = 0$ and $Var(\psi(X_i)) = \nabla g(\mu)^T \Sigma \nabla g(\mu)$.
(b) Does the result of (a) apply to the situation considered in problem 1(b) of problem set #3? If so, what is the resulting influence function?
(c) Does the result of (a) apply to the situation in Problem 4 above? If so, what is the resulting influence function?

6. Suppose that $X_i \sim \text{Bernoulli}(p_i)$, $i = 1, \ldots, n$ are independent. Show that if

$$\sum_{i=1}^n p_i(1 - p_i) \to \infty \,, \tag{2}$$

then

$$\frac{\sqrt{n}(\overline{X}_n - \overline{p}_n)}{\sqrt{n^{-1}\sum_{i=1}^n p_i(1 - p_i)}} \to_d N(0, 1) \,.$$

Give one example $\{p_i\}_{i \geq 1}$ for which (2) holds and another example for which it fails.

7. **Optional bonus problem 1:** Suppose that $X_1, \ldots, X_n$ are i.i.d. with continuous distribution function $F$. Let $F_0$ be a fixed, specified (continuous) distribution function. Suppose we want to test $H: \; F = F_0$ versus $K: \; F \neq F_0$. Consider the *Cramér - von Mises statistic* given by

$$C_n^2 \equiv \int_{-\infty}^{\infty} n(\mathbb{F}_n(x) - F_0(x))^2 dF_0(x) \,.$$

(a) Show that when the null hypothesis $H$ is true

$$C_n^2 =_d \int_0^1 n(\mathbb{G}_n(t) - t)^2 dt \,,$$

where $\mathbb{G}_n$ is the empirical d.f. of $n$ i.i.d. Uniform$(0, 1)$ rv's.
(b) Show that when the null hypothesis $H$ is true,

$$C_n^2 \to_d \int_0^1 \mathbb{U}(t)^2 dt$$

where $\mathbb{U}$ is a standard Brownian bridge process.
[Hint: Use the fact that $\mathbb{U}_n \Rightarrow \mathbb{U}$ in $(D[0, 1], \| \cdot \|_\infty)$ and the continuous mapping

theorem.]
(c) Suppose that the null hypothesis fails. Thus $F \neq F_0$. Show that in this case

$$n^{-1} C_n^2 \to_{a.s.} \int_{-\infty}^{\infty} (F(x) - F_0(x))^2 dF_0(x) > 0\,,$$

and hence the test based on $C_n^2$ is consistent for all $F \neq F_0$.

8. **Optional bonus problem 2:** This is a continuation of the previous problem, and should be thought of in analogy with our development for the Pearson chi-square statistic.
(a) Suppose that $F = F_n$ satisfies $\sqrt{n}(F_n(x) - F_0(x)) \to g(x)$ in $L_2(F_0)$; i.e.

$$\int [\sqrt{n}(F_n(x) - F_0(x)) - g(x)]^2 dF_0(x) \to 0\,.$$

Describe the limiting distribution of $C_n^2$ under the local alternatives $F_n$ in terms of a Brownian bridge process $\mathbb{U}$ and $g$.
(b) Let $c^2$ denote the constant on the right side in Problem 5(c) above. In the set-up of that problem, show that when $F \neq F_0$ it follows that

$$\sqrt{n}(n^{-1} C_n^2 - c^2) \to_d N(0, V^2)$$

and find $V^2$.
[Hint: Use $\sqrt{n}(\mathbb{F}_n - F) =_d \mathbb{U}_n(F)$, $\mathbb{U}_n \Rightarrow \mathbb{U}$, and the continuous mapping theorem.]