Statistics 581
Problem Set 5 Solutions
Wellner; 11/09/2001

1. Suppose that $\underline{N}_n = (N_{11}, N_{12}, N_{21}, N_{22}) \sim \text{Mult}_4(n, \underline{p})$ where $\underline{p} = (p_{11}, p_{12}, p_{21}, p_{22})$ where $\sum_{i=1}^2 \sum_{j=1}^2 p_{ij} = 1$. (Thus $\underline{N}_n$ is the sum of n independent $\text{Mult}_4(1, \underline{p})$ random vectors $\{\underline{Y}_i\}_{i=1}^n$.) Since there are really just three independently varying parameters for this problem, it is often useful to re-express the cell probabilities in terms of two marginal probabilities, say $p_{1\cdot} = p_{11} + p_{12}$ and $p_{\cdot 1} = p_{11} + p_{21}$, and ψ, the log of the odds-ratio, defined by

$$\psi \equiv \log \frac{p_{21}/p_{22}}{p_{11}/p_{12}} = \log \frac{p_{12}p_{21}}{p_{11}p_{22}} . \tag{0.1}$$

You may use the fact that $\psi = 0$ if and only if independence holds for the 2×2 table (i.e. $p_{ij} = p_{i\cdot}p_{\cdot j}$ for $i, j = 1, 2$).
(a) Suggest an estimator of ψ, say $\widehat{\psi}$.
(b) Show that the estimator you proposed in (a) is asymptotically normal and compute the asymptotic variance of your estimator.
Solution: Solution: (a) An obvious estimator of ψ is

$$\widehat{\psi} = \log \frac{\widehat{p}_{12}\widehat{p}_{21}}{\widehat{p}_{11}\widehat{p}_{22}}$$

where $\underline{\widehat{p}} = \underline{N}/n$.
(b) Now $\widehat{\psi} = g(\underline{\widehat{p}})$ where $g(\underline{p})$ is given in (0.1) and is differentiable with derivative

$$\nabla g(\underline{p}) = (-1/p_{11}, 1/p_{12}, 1/p_{21}, -1/p_{22})$$

and, by the multivariate CLT,

$$\sqrt{n}(\underline{\widehat{p}} - \underline{p}) \to_d Z \sim N_4(0, \Sigma)$$

where $\Sigma = \text{diag}(\underline{p}) - \underline{p}\underline{p}^T$. Thus the delta method (or g'-theorem) yields

$$\begin{aligned} \sqrt{n}(\widehat{\psi} - \psi) &= \sqrt{n}(g(\underline{\widehat{p}}) - g(\underline{p})) \\ &\to_d \nabla g(\underline{p}) Z \sim N(0, \nabla g^T \Sigma \nabla g) = N(0, V^2(\underline{p})) \end{aligned}$$

where

$$V^2(\underline{p}) = \frac{1}{p_{11}} + \frac{1}{p_{12}} + \frac{1}{p_{21}} + \frac{1}{p_{22}}.$$

2. Suppose that $Y_i = \alpha + \theta'(x_i - \overline{x}) + \epsilon_i$, $i = 1, \ldots, n$, where $\epsilon_i \sim (0, \sigma^2)$ are i.i.d. and the x_i's are known vectors in R^k. Equivalently, $\underline{Y} = X\underline{\beta} + \underline{\epsilon}$ where

$$X^T = \begin{pmatrix} 1 & 1 & \cdots & 1 \\ x_1 - \overline{x} & x_2 - \overline{x} & \cdots & x_n - \overline{x} \end{pmatrix}$$

so that X is an $n \times (k+1)$ matrix. Let $\underline{\hat{\beta}}$ be the least squares estimator of $\underline{\beta} = (\alpha, \theta')'$; i.e. $\underline{\hat{\beta}} = (X^T X)^{-1} X^T \underline{Y}$. Suppose that $n^{-1}(X^T X) \to D$ where D is positive definite.
(a) What additional condition(s) do you need to impose to prove that

$$\sqrt{n}(\hat{\beta}_n - \beta) \to_d N_{k+1}(0, \text{“}\textit{something}\text{”})\,?$$

(b) Find “something” in part (a).
Solution: (a) Let $\underline{a} \in I\!R^{k+1}$. Now

$$\begin{aligned} \hat{\beta} &= (X^T X)^{-1} X^T Y \\ &= (X^T X)^{-1} X^T (X\beta + \epsilon) \\ &= \beta + (X^T X)^{-1} X^T \epsilon\,, \end{aligned}$$

so

$$\sqrt{n}(\hat{\beta} - \beta) = \sqrt{n}(X^T X)^{-1} X^T \epsilon \equiv B_n \epsilon$$

where $B_n \equiv \sqrt{n}(X^T X)^{-1} X^T$ is a $(k+1) \times n$ matrix. Thus

$$\begin{aligned} a^T(\sqrt{n}(\hat{\beta} - \beta)) &= a^T B_n \epsilon \equiv b_n^T \epsilon \\ &= \sum_{i=1}^n b_{ni}\epsilon_i \equiv \sum_{i=1}^n X_{ni} \end{aligned}$$

where $b_n^T \equiv a^T B_n$ is an $1 \times n$ vector. Now we compute

$$\mu_{ni} = E(X_{ni}) = b_{ni} \cdot 0, \qquad \sigma_{ni}^2 = Var(X_{ni}) = b_{ni}^2 \sigma^2\,,$$

and hence, using the hypothesized convergence of $n^{-1}X^TX \to D$ in the last line,

$$\begin{aligned}
\sigma_n^2 &= \sigma^2 \sum_{i=1}^n b_{ni}^2 = \sigma^2 b_n^T b_n \\
&= \sigma^2 a^T B_n B_n^T a = n\sigma^2 a^T (X^TX)^{-1}(X^TX)(X^TX)^{-1} a \\
&= \sigma^2 a^T (n^{-1}X^TX)^{-1} a \to \sigma^2 a^T D^{-1} a \equiv V^2(a) > 0
\end{aligned}$$

since D is nonsingular. To establish asymptotic normality of $a^T(\sqrt{n}(\hat{\beta} - \beta))/\sigma_n$, it remains to verify the Lindeberg condition: namely

$$\frac{1}{\sigma_n} \sum_{i=1}^n E\left\{ |X_{ni}|^2 1_{[|X_{ni}| > \delta\sigma_n]} \right\} \to 0 \tag{0.2}$$

for every $\delta > 0$. But, as we have seen before, this holds if

$$\max_{1 \le i \le n} |b_{ni}| \to 0 \qquad \text{as} \qquad n \to \infty \ : \tag{0.3}$$

the left side of (0.2) is bounded as follows:

$$\begin{aligned}
&\frac{1}{\sigma_n^2} \sum_{i=1}^n b_{ni}^2 E\left\{ \epsilon_1^2 1_{[|\epsilon_1| > \delta\sigma_n/|b_{ni}|]} \right\} \\
&\le \frac{1}{\sigma^2} E\left\{ \epsilon_1^2 1_{[|\epsilon_1| > \delta\sigma_n/\max_{1\le i \le n}|b_{ni}|]} \right\} \\
&\to \frac{1}{\sigma^2} \cdot 0 = 0
\end{aligned}$$

by (0.3), $E(\epsilon_1^2) < \infty$, and the dominated convergence theorem. Thus it follows from the Lindeberg-Feller CLT that

$$a^T(\sqrt{n}(\hat{\beta} - \beta))/\sigma_n \to_d N(0,1),$$

and since $\sigma_n^2 \to \sigma^2 a^T D^{-1} a$, this implies that

$$a^T(\sqrt{n}(\hat{\beta} - \beta)) \to_d N(0, a^T(\sigma^2 D^{-1})a)\,,$$

which in turn, via the Cramér-Wold device, implies

$$\sqrt{n}(\hat{\beta} - \beta) \to_d N_{k+1}(0, \sigma^2 D^{-1})\,.$$

3. Suppose that $X_1, \ldots, X_n$ are i.i.d. F on R. Let $\mathbb{F}_n(x) = n^{-1} \sum_{i=1}^n 1_{(-\infty,x]}(X_i)$ be the empirical d.f. of the sample. Let $\alpha \in (0,1)$. The goal of the following problem is to find a number c so that

$$P_F\{c\mathbb{F}_n(x) \le F(x) \ \text{ for all } \infty < x < \infty\} = 1 - \alpha\,, \tag{0.4}$$

i.e. so that $c\mathbb{F}_n(x)$ is a lower $1-\alpha$ confidence bound for F. Let

$$A_n(c,F) = \{\mathbb{F}_n(x) \le F(x)/c \ \text{ for all } -\infty < x < \infty\}\ .$$

(a) Show that $P_F(A_n(c,F)) = P(B_n(c))$ where, for $\mathbb{G}_n$ the empirical distribution function of $U_1, \ldots, U_n$ i.i.d. Uniform$(0,1)$ random variables,

$$B_n(c) = \{\mathbb{G}_n(x) \le x/c \ \text{ for all } 0 < x \le 1\}\ .$$

(b) Re-express the event $B_n(c)$ in terms of the order statistics $0 \le U_{(1)} \le \ldots \le U_{(n)} \le 1$ of the Uniform$(0,1)$ sample. [Hint: draw a picture first!]
(c) Compute $P(B_n(c))$ using the re-expression of the event $B_n(c)$ you found in (b) and the joint density of the uniform order statistics for $n=1$, $n=2$, and $n=3$.
(d) Extend the calculations in (b) to a general n.
(e) Find c explicitly as a function of α and give the resulting lower confidence bound.
Solution: (a) Note that for $\mathbb{F}_n^*$, the empirical d.f. of $X_i^* \equiv F^{-1}(\xi_i)$, $i = 1, \ldots, n$, we have

$$\begin{aligned}
P_F(A_n(c,F)) &= P_F\{\mathbb{F}_n(x) \le F(x)/c \ \text{ for all } \ -\infty < x < \infty\} \\
&= P\{\mathbb{F}_n^*(x) \le F(x)/c \ \text{ for all } \ -\infty < x < \infty\} \\
&= P\{\mathbb{G}_n(F(x)) \le F(x)/c \ \text{ for all } \ -\infty < x < \infty\} \\
&= P\{\mathbb{G}_n(t) \le t/c \ \text{ for all } \ 0 < t < 1\} = P(B_n(c))
\end{aligned}$$

where the next to last equality follows from Theorem 2.3.4, and the last equality follows from continuity of F.
(b) It is easily seen that

$$\begin{aligned}
B_n(c) &= \{\mathbb{G}_n(x) \le x/c \text{ for all } 0 < x \le 1\} \\
&= \left\{i/n = \mathbb{G}_n(U_{(i)}) \le U_{(i)}/c \ \text{ for all } \ i = 1, \ldots, n\right\} \\
&= \left\{U_{(i)} \ge ci/n \ \text{ for all } i = 1, \ldots, n\right\}\ .
\end{aligned}$$

(c) For $n = 1$ we find

$$P(B_1(c)) = \int_c^1 du_1 = 1 - c\,.$$

For $n = 2$ the density of the order statistics is $2! = 2$ on $0 \le u_1 \le u_2 \le 1$, and hence

$$\begin{aligned} P(B_2(c)) &= \int_c^1 \int_{c/2}^{u_2} 2du_1 du_2 = \int_c^1 2(u_2 - c/2)du_2 \\ &= (u_2^2 - cu_2|_c^1 = 1 - c - (c^2 - c^2) = 1 - c\,. \end{aligned}$$

For $n = 3$ the density of the order statistics is $3! = 6$ on $0 \le u_1 \le u_2 \le u_3 \le 1$, and hence

$$\begin{aligned} P(B_3(c)) &= \int_c^1 \int_{2c/3}^{u_3} \int_{c/3}^{u_2} 3! du_1 du_2 du_3 = \int_c^1 \int_{2c/3}^{u_3} 3!(u_2 - c/3) du_2 du_3 \\ &= \int_c^1 3!(\frac{1}{2}u_2^2 - (c/3)u_2|_{2c/3}^{u_3} du_3 \\ &= \int_c^1 3!(\frac{1}{2}u_3^2 - (c/3)u_3)du_3 \\ &= 3!(\frac{1}{3!}u_3^3 - \frac{c}{3!}u_3^2|_c^1 \\ &= 1 - c\,. \end{aligned}$$

(d) Now the density of the order statistics $U_{(1)}, \ldots, U_{(n)}$ is $n!$ on the set $0 \le u_1 \le \ldots \le u_n \le 1$, and hence it follows, continuing the pattern established for $n = 1, 2, 3$ in part (b), that

$$\begin{aligned} P(B_n(c)) &= P(U_{(i)} \ge ci/n \text{ for all } i = 1, \ldots, n) \\ &= n! \int \cdots \int_{0 \le u_1 \le \ldots \le u_n \le 1} \prod_{i=1}^n 1_{[ci/n, 1]}(u_i) du_1 \ldots du_n \\ &= n! \int_c^1 \int_{(n-1)c/n}^{u_n} \cdots \int_{2c/n}^{u_3} \int_{c/n}^{u_2} du_1 \ldots du_n \\ &= n! \int_c^1 \left\{ \frac{t^{n-1}}{(n-1)!} - \frac{c}{n(n-2)\cdots 1} t^{n-2} \right\} dt \end{aligned}$$

$$
\begin{aligned}
&= n!\left\{\frac{t^n}{n!} - \frac{ct^{n-1}}{n!}\right\}\Big|_c^1 \\
&= 1 - c\,.
\end{aligned}
$$

(e) From (a) and (d) we have

$$P_F\{c\mathbb{F}_n(x) \le F(x) \text{ for all } -\infty < x < \infty\} = 1 - c$$

so choosing $c = \alpha$ yields a $1-\alpha$ lower confidence bound of the form $\alpha\mathbb{F}_n$.

4. Suppose that $X_1, \ldots, X_n$ are i.i.d. with continuous distribution function F. Let F_0 be a fixed, specified distribution function. Suppose we want to test $H: F = F_0$ versus $K: F \ne F_0$. 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

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

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 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 $F \ne F_0$. 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 $F \ne F_0$.

Soluton: (a) Since $\mathbb{F}_n(x) =_d \mathbb{F}_n^*(x) = \mathbb{G}_n(F_0(x))$ when $F = F_0$ and F_0 is continuous, it follows that

$$\begin{aligned} C_n^2 &=_d \int n(\mathbb{F}_n^*(x) - F_0(x))^2 dF_0(x) \\ &= \int n(\mathbb{G}_n(F_0(x)) - F_0(x))^2 dF_0(x) \\ &= \int_0^1 n(\mathbb{G}_n(t) - t)^2\, dt \end{aligned}$$

by using the change of variables $t = F_0(x)$ in the last line.
(b) Note that if $\{x_n\}$ is a sequence of functions in $D[0,1]$ satisfying $\|x_n - x\|_\infty \equiv \sup_{0\le t\le 1} |x_n(t) - x(t)| \to 0$, then with $g(x) \equiv \int_0^1 x^2(t)dt$, $g(x_n) \to g(x)$. It follows from (a) and the continuous mapping theorem that, under the null hypothesis,

$$C_n^2 = \int_0^1 [\mathbb{U}_n(t)]^2\, dt \to_d \int_0^1 [\mathbb{U}(t)]^2 dt \equiv C^2\,.$$

The distribution of C^2 is the same as that of $\sum_{j=1}^\infty Z_j^2/(\pi^2 j^2)$ where Z_j are i.i.d. $N(0,1)$, and tables of the d.f. are available; see e.g. Shorack and Wellner (1986), page 147.
(c) When the null hypothesis fails (so $F \ne F_0$),

$$\begin{aligned} n^{-1} C_n^2 &= \int [\mathbb{F}_n(x) - F_0(x)]^2 dF_0(x) \\ &=_d \int [\mathbb{G}_n(F(x)) - F_0(x)]^2 dF_0(x) \\ &\to_{a.s.} \int [F(x) - F_0(x)]^2 dF_0(x) \equiv c^2 \end{aligned}$$

since $\|\mathbb{G}_n - I\|_\infty \to_{a.s.} 0$.
(d) (Not assigned!) If $F = F_n$ satisfies $\sqrt{n}(F_n - F_0) \to g$, then

$$\begin{aligned} C_n^2 &=_d \int [\sqrt{(}\mathbb{F}_n^*(x) - F_n(x)) + \sqrt{n}(F_n(x) - F_0(x))]^2 dF_0(x) \\ &= \int [\mathbb{U}_n(F_n(x)) + \sqrt{n}(F_n - F_0)]^2 dF_0(x) \end{aligned}$$

$$\begin{aligned} \rightarrow_d &\quad \int [\mathbb{U}(F_0(x)) + g(x)]^2 dF_0(x) \\ = &\quad \int_0^1 [\mathbb{U}(t) + g(F_0^{-1}(t))]^2 \, dt \,. \end{aligned}$$