Statistics 523, Problem Set 3 Solutions

Wellner; 4/21/99

1. PfS, Exercise 17.1.3, page 383: Show that $\sqrt{n}(U_n - 1/2) \rightarrow_d N(0, 1/3)$ and $\sqrt{n}(V_n - 1/2) \rightarrow_d N(0, 1/3)$, and then establish Berry-Esseen bounds for W_n and V_n where

$$V_n = n^{-2} \sum_{i=1}^{n} \sum_{j=1}^{n} 1_{[X_i + X_j \geq 0]}; \qquad U_n = \binom{n}{2}^{-1} \sum_{i<j} 1_{[X_i + X_j \geq 0]};$$

and $W_n = \sum_{i \leq j} 1_{[X_i + X_j \geq 0]} = \sum_{i=1}^{n} R_i 1_{[X_i > 0]}$.

Solution: (a) By our results for $U-$statistics

$$\sqrt{n}(U_n - 1/2) = 2\sqrt{n}(\overline{Y}_n - 1/2) + \sqrt{n} U_{0n}$$

where

$$\begin{aligned} Y_i &= E[H(X_i, X_j)|X_i] = E[1_{[X_j \geq -X_i]}|X_i] \\ &= 1 - F(-X_i) = F(X_i) \qquad \text{since} \quad F(x) = 1 - F(-x)\,, \end{aligned}$$

is distributed as $U(0, 1)$, and hence $E(Y_i) = 1/2$, $\sigma_1^2 = Var(Y_i) = 1/12$. By the calculations in (17.1.8) - (17.1.11) we see that

$$\sqrt{n}(U_n - 1/2) \rightarrow_d N(0, 4\sigma_1^2) = N(0, 3)\,.$$

Note that with $R_i \equiv$ rank of X_i,

$$\begin{aligned} \sum \sum\nolimits_{i \leq j} 1_{[X_i + X_j \geq 0]} &= \sum_{i=1}^{n} \sum_{j: R_j \leq R_i} 1_{[X_i + X_j \geq 0]} \\ &= \sum_{i=1}^{n} \sum_{j: R_j \leq R_i} 1_{[X_i \geq 0]} \\ &= \sum_{i=1}^{n} R_i 1_{[X_i \geq 0]} \\ &= W_n \end{aligned}$$

with W_n as defined above. Therefore it follows that

$$\begin{aligned} n^2 V_n &= \sum_{i=1}^n \sum_{j=1}^n 1_{[X_i+X_j \geq 0]} \\ &= \sum_{i \neq j} 1_{[X_i+X_j \geq 0]} + \sum_{i=1}^n 1_{[X_i \geq 0]} \\ &= n(n-1)U_n + \sum_{i=1}^n 1_{[X_i \geq 0]} \\ &= 2\sum\sum_{i \leq j} 1_{[X_i+X_j \geq 0]} - \sum_{i=1}^n 1_{[X_i \geq 0]} \\ &= 2W_n - \sum_{i=1}^n 1_{[X_i \geq 0]}\,, \end{aligned}$$

and hence

$$\begin{aligned} \sqrt{n}(V_n - 1/2) &= \sqrt{n}\left(\frac{2\binom{n}{2}}{n^2} U_n + n^{-2} \sum_{i=1}^n 1_{[X_i \geq 0]} - 1/2 \right) \\ &= \frac{n(n-1)}{n^2}\sqrt{n}(U_n - 1/2) + \sqrt{n}\left(\frac{n(n-1)}{n^2} - 1 \right)(1/2) \\ &\quad + n^{-3/2} \sum_{i=1}^n (1_{[X_i \geq 0]} - 1/2) + (1/2)n^{-1/2} \\ &\to_d 1 \cdot N(0, 1/3) + 0 + 0 + 0 = N(0, 1/3)\,. \end{aligned}$$

Since $V_n = 2n^{-2}W_n - n^{-2}\sum_{i=1}^n 1_{[X_i \geq 0]}$ we also have

$$\sqrt{n}(n^{-2}W_n - 1/4) \to_d N(0, 1/12)\,.$$

Now for the Berry-Esseen type bounds. Note that we can write

$$\begin{aligned} \frac{\sqrt{n}(V_n - 1/2)}{2\sigma_1} &= \frac{\sqrt{n}(U_n - 1/2)}{2\sigma_1} \\ &\quad + \left\{ \frac{1}{4\sigma_1\sqrt{n}} + \frac{1}{2\sigma_1\sqrt{n}} \frac{1}{n} \sum_{i=1}^n 1_{[X_i \geq 0]} - \frac{1}{n} \frac{\sqrt{n}(U_n - 1/2)}{2\sigma_1} \right\} \\ &= \frac{\sqrt{n}(U_n - 1/2)}{2\sigma_1} + \Delta \end{aligned}$$

where we have a Berry-Esseen bound for the convergence of the $U-$statistic in the first term from (17.1.14). Moreover, upon using the bound $Var(A+B) = Var(A) + 2Cov(A,B) + Var(B) \le a^2 + 2\sqrt{a^2b^2} + b^2 = (a+b)^2$, we find that

$$
\begin{aligned}
E(\Delta^2) &= (E\Delta)^2 + Var(\Delta) \\
&= \left(\frac{1}{4\sigma_1\sqrt{n}} + \frac{1}{2\sigma_1\sqrt{n}} P(X_1 \ge 0)\right)^2 \\
&\quad + \left\{\left(\frac{1}{4\sigma_1^2 n}\frac{1/4}{n}\right)^{1/2} + \left(\frac{1}{n^2}\right)^{1/2}\right\}^2 \\
&\le \frac{1}{n}\left(2\frac{1}{4\sigma_1^2} + 2\frac{1}{n}\left[\frac{1}{16\sigma_1^2} + 1\right]\right).
\end{aligned}
$$

Hence we have, from (11.1.24) (the "Berry-Esseen potential" bound),

$$
\begin{aligned}
&\|F_{\sqrt{n}(V_n - 1/2)/(2\sigma_1)} - \Phi\| \\
&\le \|F_{\sqrt{n}(U_n - 1/2)/(2\sigma_1)} - \Phi\| + 4\sqrt{E\Delta^2}\{1 + \sqrt{E(W^2)}\} \\
&\le \frac{E|Y_1 - 1/2|^3}{\sqrt{n}\sigma_1^3} + 4\sqrt{2}\frac{\sqrt{\sigma_{12}/\sigma_1^2}}{\sqrt{n-1}} \\
&\quad + \frac{4}{\sqrt{n}}\left(2\frac{1}{4\sigma_1^2} + 2\frac{1}{n}\left[\frac{1}{16\sigma_1^2} + 1\right]\right)^{1/2} \cdot \left(1 + \left(\frac{\sigma_{12}/2\sigma_1^2}{n-1}\right)^{1/2}\right).
\end{aligned}
$$

For W_n, we rewrite

$$
\sqrt{n}(2n^{-2}W_n - 1/2) = \sqrt{n}(V_n - 1/2) + n^{-3/2}\sum_{i=1}^{n} 1_{[X_i \ge 0]}
$$

and hence

$$
\frac{\sqrt{n}(n^{-2}W_n - 1/4)}{\sigma_1} = \frac{\sqrt{n}(V_n - 1/2)}{2\sigma_1} + \frac{1}{2\sigma_1\sqrt{n}}\frac{1}{n}\sum_{i=1}^{n} 1_{[X_i \ge 0]}.
$$

Where we have a Berry-Esseen bound for the first term above. Hence taking $W = \sqrt{n}(V_n - 1/2)/(2\sigma_1)$ and

$$
\Delta = \frac{1}{2\sigma_1\sqrt{n}}\frac{1}{n}\sum_{i=1}^{n} 1_{[X_i \ge 0]}
$$

upon computing

$$\begin{aligned} E(\Delta^2) &= (E\Delta)^2 + Var(\Delta) \\ &= \left(\frac{1}{4\sigma_1\sqrt{n}}\right)^2 + \frac{1}{4\sigma_1^2 n}\frac{1/4}{n} \\ &= \left(\frac{1}{4\sigma_1\sqrt{n}}\right)^2 \left\{1 + \frac{1}{n}\right\} . \end{aligned}$$

Thus (11.1.24) (the "Berry-Esseen potential" bound) yields

$$\begin{aligned} &\|F_{\sqrt{n}(n^{-2}W_n - 1/4)/\sigma_1} - \Phi\| \\ &\le \|F_{\sqrt{n}(V_n - 1/2)/(2\sigma_1)} - \Phi\| + 4\sqrt{E\Delta^2}\{1 + \sqrt{E(W^2)}\} \\ &\le \frac{E|Y_1 - 1/2|^3}{\sqrt{n}\sigma_1^3} + 4\sqrt{2}\frac{\sqrt{\sigma_1 2/\sigma_1^2}}{\sqrt{n-1}} \\ &\qquad + \frac{4}{\sqrt{n}}\left(2\frac{1}{4\sigma_1^2} + 2\frac{1}{n}\left[\frac{1}{16\sigma_1^2} + 1\right]\right)^{1/2} (2 + o(1)) \\ &\qquad + 4\left(\frac{1}{4\sigma_1\sqrt{n}}\right)\left\{1 + \frac{1}{n}\right\}^{1/2} (2 + o(1)) \\ &= O(n^{-1/2}) \end{aligned}$$

where we have used $E(\sqrt{n}(V_n - 1/2)/(2\sigma_1))^2 = 1 + o(1)$ in the next to last line.

2. PfS, Exercise 17.1.4, page 383. Derive the asymptotic distribution of Gini's mean difference $G_n \equiv \sum_{i<j} |X_i - X_j|/\binom{n}{2}$, assuming that $E(X_1^2) < \infty$.

   Solution: From our theory of $U-$statistics, $G_n \to_{a.s.} E|X_1 - X_2| \equiv \theta$. Moreover with $h_1(x) \equiv E\{|X_1 - X_2||X_1 = x\} = E\{|x - X_2|\}$ and $Y_i \equiv h_1(X_i)$,

$$\sqrt{n}(G_n - \theta) = 2\sqrt{n}(\overline{Y}_n - \theta) + o_p(1) \to_d N(0, 4\sigma_1^2)$$

   where

$$\sigma_1^2 = Var(h_1(X_1)) = Var(\int |X_1 - y| dF(y)) .$$

Note that

$$\begin{aligned}
\sigma_1^2 \le E(H_1^2(X_1)) &= E\{E\{|X_1 - X_2||X_1\}^2\} \\
&\le E\{|X_1 - X_2|^2\} \\
&= 2Var(X_1) \le 2E(X_1^2) < \infty \,.
\end{aligned}$$

3. PfS, Exercise 17.2.4, page 392, reworded as follows:
(a) Verify (21). Show that the value of the variance simplifies to 1 in case X_i and Y_i are independent.
(b) Verify (24).

Solution: (a) To verify (21), note that with

$$\begin{aligned}
S_{X,Y} &= N^{-1}\sum_1^N (X_i - \overline{X})(Y_i - \overline{Y}) \,, \\
S_X^2 &= N^{-1}\sum_1^N (X_i - \overline{X})^2 \,, \\
S_Y^2 &= N^{-1}\sum_1^N (Y_i - \overline{Y})^2 \,,
\end{aligned}$$

and $X_i' \equiv (X_i - \mu_X)/\sigma_X$, $Y_i' \equiv (Y_i - \mu_Y)/\sigma_Y$, we have

$$\begin{aligned}
\sqrt{N}(S_{X,Y} - Cov(X,Y))/\sigma_X\sigma_Y &= \sqrt{N}(\overline{X'Y'} - \rho) - \sqrt{N}\,\overline{X'}\,\overline{Y'} \,, \\
\sqrt{N}(S_X^2 - \sigma_X^2)/\sigma_X^2 &= \sqrt{N}(\overline{X'^2} - 1) - \sqrt{N}(\overline{X'})^2 \,, \\
\sqrt{N}(S_Y^2 - \sigma_Y^2)/\sigma_Y^2 &= \sqrt{N}(\overline{Y'^2} - 1) - \sqrt{N}(\overline{Y'})^2 \,,
\end{aligned}$$

where the second terms are all $o_p(1)$ by the CLT and weak law of large numbers, and, by the multivariate CLT we have

$$\sqrt{N}\begin{pmatrix} \overline{X'Y'} - \rho \\ \overline{X'^2} - 1 \\ \overline{Y'^2} - 1 \end{pmatrix} \to_d \underline{Z} \sim N_3(\underline{0}, \Sigma)$$

with

$$\Sigma = E\left\{ \begin{pmatrix} X'Y' - \rho \\ X'^2 - 1 \\ Y'^2 - 1 \end{pmatrix} \begin{pmatrix} X'Y' - \rho \\ X'^2 - 1 \\ Y'^2 - 1 \end{pmatrix}^T \right\}.$$

Thus by the delta-method with $g(r,s,t) = r/\sqrt{st}$, so that $\nabla g(r,s,t) = (1, -r/(2s), -t/(2s))/\sqrt{st}$, and hence $\nabla g(\rho,1,1) = (1, -\rho/2, -\rho/2) \equiv \underline{c}$, we conclude that if $E|X|^4 < \infty$, $E|Y|^4 < \infty$, then (21) holds; i.e.

$$\sqrt{N}(\widehat{\rho}_N - \rho) \rightarrow N(0, v^2)$$

where $v^2 = \underline{c}^T \Sigma \underline{c}$.

(b) For the permutation test of independence with $T_N \equiv \sum_1^N a_i b_{\pi(i)}$ where

$$a_i \equiv (X_i - \overline{X})/S_X, \qquad b_i \equiv (Y_i - \overline{Y})/S_Y, \qquad i = 1, \ldots, N,$$

we have $E(T_N) = N a.b. = 0$, and

$$Var(T_N/N) = \frac{1}{N-1} A_2 B_2 = \frac{1}{N-1},$$

since $A_2 = 1 = B_2$. Thus $\overline{T}_N \equiv \sqrt{N-1}(T_N/N) \sim (0,1)$. Moreover, by the C_r inequality, since $E|X_1|^3 < \infty$,

$$\begin{aligned} A_3 &= N^{-1}\sum_1^N |a_i - a.|^3 = N^{-1}\sum_1^N |X_i - \overline{X}|^3 / S_X^3 \\ &\leq 2^2 \left\{ N^{-1}\sum_1^N |X_i|^3 + |\overline{X}|^3 \right\} / S_X^3 \\ &\rightarrow_{a.s.} 2^2 \left\{ E|X_1|^3 + |\mu_X|^3 \right\} / \sigma_X^3 < \infty, \end{aligned}$$

and similarly for B_3. Hence by Bolthausen's form of Hoeffding's combinatorial CLT it follows that

$$\|F_{\overline{T}_N} - \Phi\| \leq \frac{K}{\sqrt{N}} A_3 B_3 = O(N^{-1/2}) \qquad a.s.,$$

and, in particular, $\overline{T}_N = \sqrt{N-1}(T_N/N) \rightarrow N(0,1)$ a.s.

4. PfS, Exercise 17.2.7, page 392. Show that in the context of sampling without replacement from a finite population that $Cov[X_i, X_j] = -A_2/(N-1)$.

Solution:

Note that $\sum_1^N X_i = \sum_1^N a_{\pi(i)} = \sum_1^N a_j$. Thus we have

$$\begin{aligned} 0 &= Var(\sum_1^N a_j) = Var(\sum_{i=1}^N X_i) \\ &= \sum_1^N Var(X_i) + \sum\sum\nolimits_{i \neq j} Cov[X_i, X_j] \\ &= NVar(X_i) + N(N-1)Cov[X_i, X_j] \end{aligned}$$

for $i \neq j$. Since

$$Var(X_i) = Var(a_{\pi(i)}) = N^{-1} \sum_{j=1}^N (a_j - a.)^2 \equiv A_2,$$

it follows that for $i \neq j$

$$Cov[X_i, X_j] = -\frac{1}{N-1} Var(X_i) = -\frac{1}{N-1} A_2 \,.$$