Statistics 523, Problem Set 1 Solutions

Wellner; 4/8/2020

1. Consider the bivariate distribution function H defined in the proof of Skorokhod's theorem in the course notes (and PfS, 2017 page 316). Show that

$$H(a,b) \equiv \int_{[0,a]} \int_{(0,b]} \frac{u+v}{EX^+} dF(-u) dF(v)$$

is in fact a bivariate d.f. on $[0,\infty) \times (0,\infty)$.

Solution: First note that $H(u,0) = 0$ for each $u \in [0,\infty)$, $H(0,v) = 0$, $H(0,v) = 0$ for each $v \in (0,\infty)$, and $H(0,0) = 0$. On the other hand, since $0 = E(X) = EX^+ - EX^-$,

$$\begin{aligned} H(\infty,\infty) &= \int_{[0,\infty)} \int_{(0,\infty)} \frac{u+v}{EX^+} dF(-u) dF(v) \\ &= \frac{1}{EX^+} \int_{[0,\infty)} u dF(-u) \int_{(0,\infty)} dF(v) \\ &\quad + \frac{1}{EX^+} \int_{(0,\infty)} v dF(v) \int_{[0,\infty)} dF(-u) \\ &= 1 \cdot (F(\infty) - F(0)) \ + \ 1 \cdot (F(0) - F(-\infty)) = 1. \end{aligned}$$

Since the integrand $(u+v)/EX^+$ is non-negative it follows that $u \mapsto H(u,v)$ is a non-decreasing function of u for each fixed v and similarly for $v \mapsto H(u,v)$. It remains to show that the two-dimensional differences of H around rectangles are always non-negative. But this holds since $(u+v)/EX^+$ is non-negative and differences of F are non-negative since F is a distribution function.

Here is a bit more concerning evaluation of the integrals involving $F(-a)$ in this proof and the proof on PfS pages 316-317. Note that $X^- = -X1_{[X \le 0]}$, and hence

$$E(X^-) = -E(X1_{[X \le 0]}) = -\int_{(-\infty,0]} x dF(x) = \int_{[0,\infty)} y dF(-y)$$

by the change of variable $y = -x$. Similarly

$$F(-t) = F(-t) - F(-\infty) = \int_{(-\infty,-t]} dF(x) = \int_{[t,\infty)} dF(-y),$$

again by the change of variable $x = -y$.

2. Consider sampling n balls from an urn with R red balls and W white balls.
(a) Show that if the sampling is carried out with replacement, the process which counts the number of red balls drawn is a Markov process.
(b) Is the process defined in (a) a strong Markov process?
(c) Now suppose that the sampling is without replacement. Is the process which counts the number of red balls in the sample a Markov process? Justify your answer.

Solution: (a) If we sample with replacement, then, letting $X_j = 1$ if the ball drawn is red and $X_j = 0$ if the ball drawn is white, for $j = 1, \ldots, n$, X_j's are i.i.d. Bernoulli(p) with $p = R/(R+W)$, and the number of red balls drawn after k draws is $S_k = \sum_{j=1}^{k} X_j$. Let $\mathcal{A}_k = \sigma[S_1, \ldots, S_k]$ denote the sigma-field generated by the process $\{S_1, S_2, \ldots, S_k\}$. Then

$$\begin{aligned}
&P(S_{k+1} = j|\mathcal{A}_k) \\
&\quad = P(S_{k+1} = j|S_1 = s_1, \ldots, S_k = s_k) \\
&\quad = P(S_{k+1} - S_k = j - s_k|S_1 = s_1, \ldots, S_k = s_k) \\
&\quad = P(X_{k+1} = j - s_k, S_1 = s_1, \ldots, S_k = s_k)/P(S_1 = s_1, \ldots, S_k = s_k) \\
&\quad = P(X_{k+1} = j - s_k|S_k = s_k) = P(S_{k+1} = j|S_k = s_k).
\end{aligned}$$

Thus $\{S_k, \mathcal{A}_k\}$ is a Markov process.
(b) Since $\{S_k\}$ is the partial sum process of i.i.d. rv's X_k, it is a strong Markov process by PfS, Theorem 8.7.1, page 179.
(c) My initial solution for this part of the problem was completely incorrect. (Everybody gets a perfect score on this one!) Thanks to Yunkyu Song and Peter Liu for pointing out several of my errors.
First a couple of remarks on queries I received:

(a) **Query 1:** The filtration $\{\mathcal{A}_k : 1 \le k \le n\}$? Whereas the solution of part (a) is given in terms of $\mathcal{A}_k = \sigma[S_1, \ldots, S_k]$, the solution of

(c) is given in terms of $\widetilde{\mathcal{A}}_k \equiv \sigma[X_1, \ldots, X_k]$? But $\widetilde{\mathcal{A}}_k = \mathcal{A}_k$ since $X_j = S_j - S_{j-1}$ for $j \in \{1, \ldots, n$ (with $S_0 \equiv 0$. Thus this is not the key difficulty.

(b) **Query 2:** While the problem is stated in terms of $\{(S_k, \mathcal{A}_k) : \ 0 \leq k \leq n\}$, the solution of (2c) is given in terms of $\{(X_k, \mathcal{A}_k) : \ 0 \leq k \leq n\}$? This is the real problem in terms of the Markov property. We really want to show that the process $\{(S_k, \mathcal{A}_k) : \ 0 \leq k \leq n\}$ is Markov. In this notation we want to show that

$$P(S_{k+1} = m|\mathcal{A}_k) = P(S_{k+1} = m|S_k) \tag{1}$$

a.s. for all $0 \leq m \leq R \wedge (k+1)$.

The following proof of (1) is adapted from the solution given by Yunkyu Song.

Let $N \equiv R + W$. Note that after k draws, the total number of balls left in the urn is $N - k$. On the event $[S_k = m - 1]$, the number of red balls remaining in the urn is $R - (m-1)$, and on the event $[S_k = m]$ the number of white balls remaining in the urn is $W - (k - m)$. We can write

$$\begin{aligned} P(S_{k+1} = m|S_k) &= \frac{P([X_{k+1} = 1] \cap [S_k = m-1])}{P(S_k = m-1)} 1_{[S_k = m-1]} \\ &\quad + \frac{P([X_{k+1} = 0] \cap [S_k = m])}{P(S_k = m)} 1_{[S_k = m]} \\ &= \frac{R - (m-1)}{N-k} 1_{[S_k = m-1]} + \frac{W - (k-m)}{N-k} 1_{[S_k = m]}. \end{aligned} \tag{2}$$

On the other hand,

$$\begin{aligned} &P(S_{k+1} = m|\mathcal{A}_k) \\ &= \sum_{a_1, \ldots, a_k \in \{0,1\} : \sum_1^k a_j = m-1} \frac{P([X_{k+1} = 1] \cap \cap_{j=1}^k [X_j = a_j])}{P(\cap_{j=1}^k [X_j = a_j])} \cdot 1\{\cap_{j=1}^k [X_j = a_j]\} \\ &\quad + \sum_{a_1, \ldots, a_k \in \{0,1\} : \sum_1^k a_j = m} \frac{P([X_{k+1} = 0] \cap \cap_{j=1}^k [X_j = a_j])}{P(\cap_{j=1}^k [X_j = a_j])} \cdot 1\{\cap_{j=1}^k [X_j = a_j]\} \\ &= \sum_{(m-1)} \frac{R - (m-1)}{N-k} 1\{\cap_{j=1}^k [X_j = a_j]\} + \sum_{(m)} \frac{W - (k-m)}{N-k} 1\{\cap_{j=1}^k [X_j = a_j]\} \\ &= \frac{R - (m-1)}{N-k} 1_{[S_k = m-1]} + \frac{W - (k-m)}{N-k} 1_{[S_k = m]}. \end{aligned} \tag{3}$$

where
$\sum_{(m-1)}$ denotes the sum over $\{a_1, \ldots, a_k \in \{0,1\} : \sum_1^k a_j = m-1\}$
and
$\sum_{(m)}$ denotes the sum over $\{a_1, \ldots, a_k \in \{0,1\} : \sum_1^k a_j = m\}$.
Since (3) equals (2), the process $\{(S_k, \mathcal{A}_k)\}$ is Markov. It would be useful to cross-check this as follows: since

$$P(S_k = m) = \frac{\binom{R}{m}\binom{W}{k-m}}{\binom{N}{k}},$$

the following identity should hold:

$$\begin{aligned} P(S_{k+1} = m) &= \frac{W-(k-m)}{N-k} P(S_k = m) + \frac{R-(m-1)}{N-k} P(S_k = m-1) \\ &= \frac{W-(k-m)}{N-k} \cdot \frac{\binom{R}{m}\binom{W}{k-m}}{\binom{N}{k}} + \frac{R-(m-1)}{N-k} \frac{\binom{R}{m-1}\binom{W}{k-(m-1)}}{\binom{N}{k}}. \end{aligned}$$

3. PfS Exercise 12.1.6, page 299, parts (a) and (b).

Solution: (a) To show that $C_\infty \equiv C[0,\infty)$ as a metric space with the metric $\rho_\infty(x,y) \equiv \sum_{k=1}^\infty 2^{-k}\rho_k(x,y)/(1+\rho_k(x,y))$ we need only to show that ρ_∞ is a metric: (i) $\rho_\infty(x,y) = 0$ if and only if $x = y$;
(ii) $\rho_\infty(x,y) = \rho_\infty(y,x)$;
(iii) $\rho_\infty(x,z) \le \rho_\infty(x,y) + \rho_\infty(y,z)$.
But (i) holds since it holds for ρ_k for every $k \ge 1$. Similarly, (ii) holds since it holds for ρ_k for each $k \ge 1$. Finally, (iii) holds since it holds for ρ_k for each $k \ge 1$:

$$\rho_k(x,z) \le \rho_k(x,y) + \rho_k(y,z),$$

(or $c \le a + b$), and then since $w \mapsto w/(1+w)$ is monotone increasing,

$$\begin{aligned} \frac{\rho_k(x,z)}{1+\rho_k(x,z)} &= \frac{c}{1+c} \le \frac{a+b}{1+a+b} \\ &= \frac{a}{1+a+b} + \frac{b}{1+a+b} \\ &\le \frac{a}{1+a} + \frac{b}{1+b} \quad \text{since } a, b \ge 0 \\ &= \frac{\rho_k(x,y)}{1+\rho_k(x,y)} + \frac{\rho_k(y,z)}{1+\rho_k(y,z)}. \end{aligned}$$

(b) Now to show $\rho_\infty(x,y) \to 0$ if and only if $\rho_k(x,y) \to 0$, first note that if $\rho_k(x,y) \to 0$, then by the dominated convergence theorem with integrable dominating function 2^{-k} (since $w/(1+w) \leq 1$ for $w \geq 0$), it follows that $\rho_\infty(x,y) \to 0$. On the other hand suppose that $\rho_\infty(x,y) \to 0$ and for some k_0 we have $\rho_{k_0}(x,y) \to c_0 > 0$. But then

$$\rho_\infty(x,y) \geq 2^{-k_0} \frac{\rho_{k_0}(x,y)}{1+\rho_{k_0}(x,y)} \to 2^{-k_0} \frac{c_0}{1+c_0},$$

which is a contradiction. Thus $\rho_k(x,y) \to 0$ for every k.

4. PfS (2017), Exercise 12.8.1, page 324: let $X_0 \equiv 0$ and let $X_1, X_2, \ldots$ be i.i.d. with mean zero and variance $\sigma^2 \equiv E(X_1^2) < \infty$. Let $S_k \equiv X_1 + \cdots + X_k$ for each integer $k \geq 0$.
(a) Find the asymptotic distribution of $S_1 + \cdots + S_n)/c_n$ for an appropriate sequence c_n.
(b) Determine a representation for the asymptotic distribution of the "absolute area" under the partial sum process as given by
$(|S_1| + \cdots + |S_n|)/c_n$.

Solution: (a) Without loss of generality, suppose that $E(X_1^2) = 1$; if not, replace X_j by X_j/σ for all j. Note that $S_k = X_1 + \cdots + X_k = \sqrt{n}\mathbb{S}_n(k/n)$, so

$$\sum_{k=1}^n S_k = \sum_{k=1}^n \sqrt{n}\mathbb{S}_n(k/n) = n^{3/2}n^{-1}\sum_{k=1}^n \mathbb{S}_n(k/n),$$

and hence

$$n^{-3/2}\sum_{k=1}^n S_k = n^{-1}\sum_{k=1}^n \mathbb{S}_n(k/n) = \int_0^1 \mathbb{S}_n(t)dt \equiv g(\mathbb{S}_n)$$

where, for $x \in C[0,1]$, $g(x) \equiv \int_0^1 x(t)dt$. Note that g is $\|\cdot\|-$continuous. Thus, by Donsker's theorem

$$g(\mathbb{S}_n) \to_d g(\mathbb{S}) = \int_0^1 \mathbb{S}(t)dt.$$

Note that $E[g(\mathbb{S})] = \int_0^1 E[\mathbb{S}(t)]dt = \int_0^1 0 \cdot dt = 0$. Moreover, by Fubini's theorem (justify?!)

$$\begin{aligned} E[g^2(\mathbb{S})] &= E\left(\int_0^1 \mathbb{S}(u)du\right)\left(\int_0^1 \mathbb{S}(v)dv\right)dudv \\ &= \int_0^1\int_0^1 E[\mathbb{S}(u)\mathbb{S}(v)]dudv = \int_0^1\int_0^1 u \wedge v\, dudv \\ &= 2\int_0^1 \left(\int_0^v udu\right)dv = \int_0^1 v^2 dv = 1/3. \end{aligned}$$

Since $g(\mathbb{S})$ is a linear combination of Gaussian random variables and $Eg^2(\mathbb{S}) < \infty$ it follows that $g(\mathbb{S})$ is also Gaussian.

Thus $g(\mathbb{S}) \sim N(0, 1/3)$, and we have

$$g(\mathbb{S}_n) \to_d g(\mathbb{S}) \sim N(0, 1/3)$$

by Donsker's theorem and the variance computation above. Note that we also have $g(\mathbb{S}_n) \to_p g(\mathbb{S})$ for the special (Skorokhod) versions of $\mathbb{S}_n$ satisfying $\|\mathbb{S}_n - \mathbb{S}\| \to_p 0$.

An alternative approach to this problem would proceed via the Lindeberg-Feller CLT as follows: First note that

$$\begin{aligned} \sum_{k=1}^n S_k &= \sum_{k=1}^n\sum_{j=1}^k X_j = \sum_{k=1}^n\sum_{j=1}^n 1_{[j\le k]}X_j \\ &= \sum_{j=1}^n \left(\sum_{k=1}^n 1_{[j\le k]}\right)X_j \\ &= \sum_{j=1}^n (n-j+1)X_j \\ &= n^{3/2}n^{-1/2}\sum_{j=1}^n \left(1 - \frac{j-1}{n}\right)X_j, \end{aligned}$$

so

$$n^{-3/2}\sum_{k=1}^n S_k = \sum_{j=1}^n a_{n,j}X_j \equiv \sum_{j=1}^n Y_{n,j}$$

where $a_{n,j} \equiv n^{-1/2}(1-(j-1)/n)$ for $1 \le j \le n$. Then apply the Lindeberg-Feller CLT. Note that

$$E\left(\sum_{j=1}^{n} a_{n,j}X_j\right) = 0,$$

and

$$\begin{aligned} Var(\sum_{j=1}^{n} a_{n,j}X_j) &= \sum_{j=1}^{n} a_{n,j}^2 = n^{-1}\sum_{j=1}^{n}(1-(j-1)/n)^2 \\ &\to \int_0^1 (1-t)^2 dt = 1/3. \end{aligned}$$

(b) To find an asymptotic representation for $(|S_1| + \cdots + |S_n|)/c_n$ we start again with $S_k = \sqrt{n}\mathbb{S}_n(k/n)$. Then, with the choice of $c_n = n^{3/2}$ as in (a),

$$n^{-3/2}\sum_{k=1}^{n}|S_k| = n^{-1}\sum_{k=1}^{n}|\mathbb{S}_n(k/n)| = \int_0^1 |\mathbb{S}_n(t)|dt = g(\mathbb{S}_n)$$

where now $g(x) \equiv \int_0^1 |x(t)|dt$ for $x \in C[0,1]$. This g is also $\|\cdot\|-$continuous, and hence it follows by Donsker's Theorem that

$$n^{-3/2}\sum_{k=1}^{n}|S_k| = g(\mathbb{S}_n) \to_d g(\mathbb{S}) = \int_0^1 |\mathbb{S}(t)|dt.$$

For more information about the distributions of this and other "Brownian areas" such as $\int_0^1 |\mathbb{U}(t)|dt$, see:

Perman, M. and Wellner, J.A. (1996).
On the distribution of Brownian areas.
Ann. Appl. Prob. **6**, 1091 - 1111.

and the references given there.