TOPICS IN STATISTICAL THEORY Part III.

Example Sheet 1 (of 3) YZ/Lent 2020

Comments and corrections to zemel@statslab.cam.ac.uk. Starred questions will be marked for the examples class.*

[*Notation: For a square-integrable function* $g : \mathbb{R} \to \mathbb{R}$, *define* $R(g) = \int_{-\infty}^{\infty} g(x)^2 \, dx$; *for a kernel* K, *define* $\mu_2(K) = \int_{-\infty}^{\infty} x^2 K(x) \, dx$.]

1. Let $U_1, \ldots, U_n \overset{iid}{\sim} U(0,1)$, and let $Y_1, \ldots, Y_{n+1} \overset{iid}{\sim} \mathrm{Exp}(1)$. Writing $S_j = \sum_{i=1}^{j} Y_i$ for $j = 1, \ldots, n+1$, show that

$$U_{(j)} \overset{d}{=} \frac{S_j}{S_{n+1}} \sim \mathrm{Beta}(j, n-j+1),$$

for $j = 1, \ldots, n$.

2.* (Hoeffding's inequality) (a) Let Y be a random variable with mean zero and $a \le Y \le b$. Use convexity to show that for every $t \in \mathbb{R}$, we have

$$\log \mathbb{E}(e^{tY}) \le -\alpha u + \log(\beta + \alpha e^u),$$

where $u = t(b-a)$ and $\alpha = 1-\beta = -a/(b-a)$. Using a second-order Taylor expansion about the origin, deduce that $\log \mathbb{E}(e^{tY}) \le t^2(b-a)^2/8$.

(b) Now let $Y_1, \ldots, Y_n$ be independent with $\mathbb{E}(Y_i) = 0$ and $a_i \le Y_i \le b_i$ for $i = 1, \ldots, n$. Use Markov's inequality to show that, for every $\epsilon > 0$, we have

$$\mathbb{P}\left(\left| \sum_{i=1}^{n} Y_i \right| > \epsilon \right) \le 2 \exp\left(-\frac{2\epsilon^2}{\sum_{i=1}^{n} (b_i - a_i)^2} \right).$$

3. Let $X_1, \ldots, X_n$ be independent with distribution P on a measurable space $(\mathcal{X}, \mathcal{A})$, and let $\hat{P}_n$ be the empirical measure of $X_1, \ldots, X_n$; thus $\hat{P}_n(A) = n^{-1} \sum_{i=1}^{n} \mathbb{1}_{\{X_i \in A\}}$ for $A \in \mathcal{A}$. Show that, for all $\epsilon > 0$ and $A \in \mathcal{A}$, we have

$$\mathbb{P}(|\hat{P}_n(A) - P(A)| > \epsilon) \le 2e^{-2n\epsilon^2}.$$

4. For $x, y \in \mathbb{R}^d$ write $x \le y$ if and only if $x_i \le y_i$ for all $1 \le i \le d$. Let $X_1, \ldots, X_n$ be independent d-dimensional random vectors with distribution function

$F(x) = P(X \leq x)$, $x \in \mathbb{R}^d$. Let $\widehat{F}_n(x) = n^{-1}\sum_{i=1}^n 1(X_i \leq x)$ be the empirical distribution function. Show that

$$\sup_{x\in\mathbb{R}^d} |\widehat{F}_n(x) - F(x)| \to 0 \qquad \text{almost surely.}$$

5. Let $X_1, \ldots, X_n$ be independent with density function $f(x) = |x|$ on $[-1, 1]$ and distribution function $F(x) = [1 + x^2 sgn(x)]/2$ on $[-1, 1]$. Show that $\sqrt{n}(X_{([n/2])} - F^{-1}(1/2))$ is not asymptotically normal. **Hint:** Take $n = 2k + 1$ and show that $P(\sqrt{n}(X_{([n/2])} - F^{-1}(1/2)) > t) \to 1/2$ for all $t > 0$.

6. (a) Verify the algebraic identity

$$\phi_\sigma(x-\mu)\phi_{\sigma'}(x-\mu') = \phi_{\sigma\sigma'/(\sigma^2+\sigma'^2)^{1/2}}(x-\mu^*)\phi_{(\sigma^2+\sigma'^2)^{1/2}}(\mu-\mu'),$$

where $\mu^* = (\sigma'^2\mu + \sigma^2\mu')/(\sigma^2 + \sigma'^2)$, and $\phi_\sigma(x)$ is the $N(0, \sigma^2)$ density.

(b) Let $X_1, \ldots, X_n$ be independent $N(0, \sigma^2)$ random variables. Taking K to be the $N(0, 1)$ density, show that the mean integrated squared error of the kernel density estimate $\hat{f}_h$ with kernel K and bandwidth h can be expressed exactly as

$$\mathrm{MISE}(\hat{f}_h) = \frac{1}{2\pi^{1/2}}\left\{\frac{1}{nh} + \left(1-\frac{1}{n}\right)\frac{1}{(h^2+\sigma^2)^{1/2}} - \frac{2^{3/2}}{(h^2+2\sigma^2)^{1/2}} + \frac{1}{\sigma}\right\}.$$

7. (Continuation) Now suppose that $h = h_n$ satisfies $h \to 0$ as $n \to \infty$ and $nh \to \infty$ as $n \to \infty$. Derive an appropriate asymptotic expansion of the $MISE$ computed above, and deduce that the asymptotically optimal bandwidth with respect to the $MISE$ criterion is given by

$$h_{AMISE} = \left(\frac{4}{3n}\right)^{1/5}\sigma.$$

Check that the same expression is obtained from the general formula for the asymptotically optimal bandwidth for a second-order kernel.

8. Let $X_1, \ldots, X_n \overset{iid}{\sim} f$, where f'' is bounded. Write $\tilde{f}_b$ for the histogram estimator of f with binwidth b. Assume $b = b_n \to 0$ and $nb \to \infty$ as $n \to \infty$. For $x \in \mathbb{R}$, let $I_b(x)$ denote the bin containing x and $p_b(x) = \mathbb{P}\{X_1 \in I_b(x)\}$ denote the bin probability. Show that

$$p_b(x) = bf(x) + \frac{1}{2}f'(x)[b^2 - 2b\{x - t_b(x)\}] + O(b^3)$$

as $n \to \infty$, where $t_b(x)$ is the left-hand endpoint of $I_b(x)$. Deduce that

$$\mathrm{MSE}\{\tilde{f}_b(x)\} = \frac{f(x)}{nb} + \frac{1}{4}b^2 f'(x)^2 + f'(x)^2\{x - t_b(x)\}^2 - bf'(x)^2\{x - t_b(x)\} + O\left(\frac{1}{n} + b^3\right).$$

9. (Continuation) Assuming in addition that $R(f') < \infty$, argue informally that

$$\mathrm{MISE}(\tilde{f}_b) = \frac{1}{nb} + \frac{1}{12}b^2 R(f') + o\Big(\frac{1}{nb} + b^2\Big).$$

Hence derive the AMISE optimal binwidth b_{AMISE} and find $\mathrm{AMISE}(\tilde{f}_{b_{AMISE}})$.

10. (Scheffé's theorem) Let (f_n) be a sequence of densities and f be another density such that $f_n \to f$ almost everywhere. By integrating $g_n = f - f_n$ separately over $\{x : g_n(x) > 0\}$ and $\{x : g_n(x) \leq 0\}$ and using dominated convergence, show that

$$\int_{-\infty}^{\infty} |f_n(x) - f(x)|\, dx \to 0.$$

11.* Assume the standard conditions on f, h and K from lectures, and also that f'' is continuous with $R(f'') < \infty$. Use Fubini's theorem to show that $h \int_{-\infty}^{\infty} (K_h^2 * f)(x)\, dx = R(K)$.

Use the dominated convergence theorem to show that $(K_h * f)(x) \to f(x)$ for each $x \in \mathbb{R}$, and show that $\sup_{n\in\mathbb{N}} \sup_{x\in\mathbb{R}} (K_h * f)(x) < \infty$. Apply Scheffé's theorem to deduce that $\int_{-\infty}^{\infty} (K_h * f)^2(x)\, dx \to \int_{-\infty}^{\infty} f(x)^2\, dx$.

Finally, deduce that

$$\int_{-\infty}^{\infty} \mathrm{Var}\{\hat{f}_h(x)\}\, dx = \frac{1}{nh} R(K) + O(n^{-1}).$$

12. (Continuation) Show that $\int_{-\infty}^{\infty} \big[\mathbb{E}\{\hat{f}_h(x)\} - f(x)\big]^2\, dx = h^4 \int_{-\infty}^{\infty} A_n^2(x)\, dx$, where

$$A_n(x) = \int_{-\infty}^{\infty} \int_0^1 (1-t) f''(x - thz) z^2 K(z)\, dt\, dz.$$

Apply Cauchy-Schwarz twice, firstly to the innermost integral with $(1-t)^{1/2}|z|K^{1/2}(z)$ as one term of the product, and secondly to the middle integral, and then use Fubini's theorem to evaluate the x-integral first, to show that

$$\int_{-\infty}^{\infty} A_n^2(x)\, dx \leq \frac{1}{4} R(f'') \mu_2^2(K)$$

for all n. Use dominated convergence to show that $A_n(x) \to \frac{1}{2} f''(x) \mu_2(K)$ for each $x \in \mathbb{R}$. Apply Fatou's lemma and combine the previous results to conclude that

$$\mathrm{MISE}(\hat{f}_h) = \frac{1}{nh} R(K) + \frac{1}{4} h^4 R(f'') \mu_2^2(K) + o\Big(\frac{1}{nh} + h^4\Big).$$