Statistics GIDP
Ph.D. Qualifying Exam
Theory

May 27, 2015, 9:00am-1:00pm

Instructions: Provide answers on the supplied pads of paper; write on only one side of each sheet. Complete exactly 2 of the first 3 problems, and 2 of the last 3 problems. Turn in only those sheets you wish to have graded. Stay calm and do your best; good luck.

1. n people go to a conference where cell phones are not allowed. The cell phones are collected during the conference. Suppose that these people have their cell phones returned at random. Let $X_i = 1$ if the ith person gets his or her own cell phone back and 0 otherwise. Let $S_n = \sum_{i=1}^{n} X_i$. Then S_n is the total number of people who get their own cell phones back. Show that

   (a) $\mathrm{E}(X_i^2) = \frac{1}{n}$.

   (b) $\mathrm{E}(X_i X_j) = \frac{1}{n(n-1)}$ for $i \neq j$.

   (c) $\mathrm{E}(S_n^2) = 2$.

   (d) $\mathrm{Var}(S_n) = 1$.

2. The Multinomial distribution has probability mass function

$$P(\boldsymbol{X} = \mathbf{n}) = \frac{n!}{n_1! \cdots n_p!} \prod_{j=1}^{p} a_j^{n_j},$$

   where $\boldsymbol{X} = (X_1, ..., X_p)^\top$, $\mathbf{n} = (n_1, ..., n_p)^\top$ is a non-negative integer valued vector with $\sum_{j=1}^{p} n_j = n$ and $\mathbf{a} = (a_1, ..., a_p)^\top$ is a positive valued vector with $\sum_{j=1}^{p} a_j = 1$. $\mathbf{a}$ and n are parameters.

   (a) Show the marginal distribution of $X_j, j = 1, \cdots, n$, is Binomial. What are the values of the parameters of the Binomial distribution?

   (b) What is the distribution of $X_j + X_k$ for $j \neq k$? Specify the parameters.

   (c) Find the conditional expectation $\mathrm{E}[X_1 | X_2 + X_3]$.

   (d) Calculate the mean vector $\mathrm{E}(\boldsymbol{X})$ and var-cov matrix $\mathrm{Var}(\boldsymbol{X})$.

3. Let $X_1, X_2, \cdots$ be a sequence of random variables such that X_1 is uniform on $[0, 1]$. For $n = 2, \cdots$, the conditional distribution of X_{n+1} given $X_1, \cdots, X_n$, is uniform on $[0, cX_n]$ for some number c such that $\sqrt{3} < c < 2$.

   (a) Find the expectation of X_n^r for $r > 0$.

(b) Show that X_n converges to 0 in mean ($r = 1$), i.e., show that $\mathrm{E}|X_n - 0| \to 0$ as $n \to \infty$, but not in quadratic mean ($r = 2$), i.e., show that $\mathrm{E}|X_n - 0|^2 \not\to 0$ as $n \to \infty$.

(c) Does X_n converge to 0 almost surely? (Hint: recall the Borel-Cantelli Lemma: If for any $\epsilon > 0$, $\sum_{n=1}^{\infty} P(|X_n - X| > \epsilon) < \infty$, then $X_n \to X$ a.s.)

4. Let $X_1, \cdots, X_n$ be a sample of i.i.d. observations drawn from a distribution function F. The empirical distribution function is defined as

$$\hat{F}_n(t) = \frac{1}{n} \sum_{i=1}^{n} I(X_i \leq t), \quad \forall t \in (-\infty, \infty)$$

where $I(\cdot)$ is the indicator function: $I(U) = 1$ if the event U occurs and $= 0$ otherwise. Assume t is any fixed constant.

(a) Show that $\hat{F}_n(t)$ is an unbiased estimator of $F(t)$.

(b) Specify the distribution of $\hat{F}_n(t)$.

(c) For any fixed t, show that

$$\sqrt{n}\{\hat{F}_n(t) - F(t)\} \longrightarrow_d N(0, \nu(t))$$

as $n \longrightarrow \infty$, and determine the value of $\nu(t)$. Here $\longrightarrow_d$ represents convergence in distribution.

5. Data $(x_i, Y_i), i = 1, \ldots, n$ are modeled, assuming $x_1, \ldots, x_n$ are fixed and positive constants; $Y_1, \ldots, Y_n$ are independent; Y_i is distributed as Expoential(θx_i) with mean θx_i for each i; and $\theta > 0$ is an unknown parameter. Define $S_n = \sum_{i=1}^{n} \frac{Y_i}{x_i}$.

(a) Specify the likelihood for θ in terms of S_n, the x_i's, and θ.

(b) Show that the sum S_n is complete and sufficient for θ.

(c) Find the uniform minimum variance unbiased estimator (UMVUE) of θ. Justfiy your answer.

(d) Compute the variance of the UMVUE estimator.

(e) Suppose θ has an inverse Gamma IG(α, β) prior distribution and $\alpha > 1, \beta > 0$. Find the Bayes estimator of θ.

Fact: If $W \sim IG(\alpha, \beta)$, then its pdf is given by

$$f_W(w) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^{\alpha}} \frac{1}{w^{\alpha+1}} e^{-\frac{1}{\beta w}}, & \text{if } w > 0, \\ 0 & \text{otherwise,} \end{cases},$$

where $\alpha > 1, \beta > 0$, and $\Gamma(\cdot)$ denotes the gamma function.

6. Let $X_1, \cdots, X_n$ be an i.i.d. sample with common pdf

$$f_X(x|\theta) = \frac{3x^2}{\theta^3} I_{[0,\theta]}(x),$$

where $\theta > 0$ is an unknown parameter.

(a) Find the MLE of θ and verify that it is also sufficient for θ.

(b) Let T be the MLE obtained in part (a). Show that the pdf of T is

$$f_T(t|\theta) = \frac{3nt^{3n-1}}{\theta^{3n}} I_{[0,\theta]}(t).$$

(c) Does this family of distributions have a montone likelihood ratio (MLR) in T? Justify your answer.

(d) Construct the UMP test of

$$H_0 : \theta \leq 1 \quad \text{vs} \quad H_1 : \theta > 1.$$

and find the critical value when $\alpha = 0.05$ and $n = 2$.

(e) Use T to construct the shortest 95% confidence interval for θ.

Statistics GIDP Ph.D. Qualifying Exam Theory Solution 2015, May

1. This is the classic Envelop matching problem

   (a) $P(X_i = 1) = (n-1)!/n! = 1/n$. Thus $\mathrm{E}(X_i^2) = 1/n$.

   (b) For $i \neq j$, $P(X_i = 1, X_j = 1) = (n-2)!/n! = 1/n(n-1)$, thus $\mathrm{E}(X_i X_j) = \frac{1}{n(n-1)}$.

   (c) Using results from parts (a) and (b), we can easily get $\mathrm{E}(S_n^2) = 1/n \cdot n + n(n-1) \cdot \frac{1}{n(n-1)} = 2$.

   (d) $\mathrm{Var}(S_n) = \mathrm{E}(S_n^2) - [\mathrm{E}(S_n)]^2 = 2 - 1 = 1$.

2. (a) Taking the summation w.r.t. $X_2, \cdots, X_n$ we can find $X_i \sim Bin(n, a_i)$.

   (b) Similarly we can obtain the distribution of $X_j + X_k$ as $Bin(n, a_j + a_k)$.

   (c) According to part (a) and (b), we can find $X_1 | X_2 + X_3 \sim Bin(n - X_2 - X_3, \frac{a_1}{1-(a_2+a_3)})$. Therefore the conditional expectation is $(n - X_2 - X_3)\frac{a_1}{1-(a_2+a_3)}$.

   (d) From the binomial distribution we have $\mathrm{E}(X_i) = na_i$ and $\mathrm{Var}(X_i) = na_i(1 - a_i)$. Using part (a) and part (b) we can get $\mathrm{Cov}(X_i, X_j) = -na_i a_j$ for all $i \neq j$.

3. (a) Consider a r.v. $Y \sim Unif[0, \theta]$, then $\mathrm{E}Y^r = \frac{\theta^r}{r+1}$. We first obtain $\mathrm{E}[X_2^r | X_1] = (cX_1)^r/(r+1)$, then $\mathrm{E}X_2^r = \frac{c^r}{(r+1)^2}$. By keeping doing this, we can find $\mathrm{E}X_n^r = \frac{c^{r(n-1)}}{(r+1)^n}$.

   (b) Using the result from part (a), $r = 1$, $\mathrm{E}|X_n - 0| = c^{n-1}/2^n \to 0$ as $n \to \infty$ since $c < 2$. However $\mathrm{E}|X_n - 0|^2 = c^{2(n-1)}/3^n \not\to 0$ because $c^2 > 3$.

   (c) By Markov's inequality $P(X_n > \epsilon) \leq \frac{\mathrm{E}X_n}{\epsilon} = \frac{c^{n-1}}{\epsilon 2^n}$. Applying the Borel-Cantelli lemma we have $\sum_{n=1}^{\infty} P(X_n > \epsilon) = \sum \frac{1}{\epsilon c}(\frac{c}{2})^n < \infty$ when $c < 2$. Therefore $X_n \to 0, a.s.$.

4. (a) For any t, since X_i's are iid, we have

$$E[\hat{F}_n(t)] = E[I(X_1 \leq t)] = P(X_1 \leq t) = F(t).$$

   So $\hat{F}_n(t)$ is an unbiased estimator of $F(t)$ for any fixed t.

   (b) For any t, define $W_i = I(X_i \leq t)$ for $i = 1, \cdots, n$. Note that W_i takes value 1 with probability $P(X_i \leq t) = F(t)$, and 0 with probability $1 - F(t)$, so W_i follows $Bin(1, F(t))$. The data are iid, so

$$n\hat{F}_n(t) = \sum_{i=1}^{n} W_i \sim Bin(n, F(t)).$$

   (c) Then $W_i \sim Bin(1, F(t))$ with $Var(W_i) = F(t)[1 - F(t)] < \infty$. By CLT, we have

$$\sqrt{n}\{\hat{F}_n(t) - F(t)\} \longrightarrow_d N(0, \nu(t)),$$

   as $n \longrightarrow \infty$ with $\nu(t) = F(t)[1 - F(t)]$.

5. (a) The likelihood for θ is

$$L(\theta) = \prod_{i=1}^{n} \frac{1}{\theta x_i} e^{-y_i/(x_i\theta)} = \frac{1}{\theta^n \prod_{i=1}^n x_i} e^{-S_n/\theta}.$$

(b) One-parameter full-rank exponential family. So the sum $S_n = \sum_{i=1}^n \frac{Y_i}{x_i}$ is complete and sufficient for θ.

(c) Since $E(S_n/n) = \frac{1}{n}\sum_{i=1}^n \frac{E(Y_i)}{x_i} = \theta$, using Rao-Blackwell Theorem, the UMVUE for θ is $\bar{S}$.

(d) $Var(S_n/n) = \frac{1}{n^2}\sum_{i=1}^n \frac{\text{Var}(Y_i)}{x_i^2} = \frac{1}{n^2}\sum_{i=1}^n \frac{\theta x_i^2}{x_i^2} = \frac{\theta^2}{n}$.

(e) The posterior density of θ given data is given by

$$\pi(\theta|y_1,\ldots,y_n) \propto \pi(\theta)L(\theta) \propto \frac{1}{\theta^{\alpha+1}} e^{-\frac{1}{\beta\theta}} \frac{1}{\theta^n} e^{-S_n/\theta} = \frac{1}{\theta^{n+\alpha+1}} e^{-\frac{1}{\theta}(S_n+\frac{1}{\beta})},$$

which is IG$(n+\alpha, \frac{1}{S_n+\frac{1}{\beta}})$. The Bayes estimator is $E(\theta|S_n) = \frac{S_n+\frac{1}{\beta}}{n+\alpha-1}$.

Suppose θ has an inverse Gamma IG(α,β) prior distribution and $\alpha > 1, \beta > 0$. Find the Bayes estimator of θ.

Fact: If $W \sim IG(\alpha,\beta)$, then its pdf is given by

$$f_W(w) = \begin{cases} \frac{1}{\Gamma(\alpha)\beta^\alpha} \frac{1}{w^{\alpha+1}} e^{-\frac{1}{\beta w}}, & \text{if } w > 0, \\ 0 & \text{otherwise,} \end{cases}$$

and $E(W) = \frac{1}{\beta(\alpha-1)}$ and $Var(W) = \frac{1}{\beta^2(\alpha-1)^2(\alpha-2)}$.

6. (a) The MLE $\hat{\theta}_{MLE} = X_{(n)}$.

(b) Define $T = \hat{\theta}_{MLE}$. Note that $F_T(t|\theta) = [F_X(x|\theta)]^n = \left[\frac{t^3}{\theta^3}\right]^n = \frac{t^{3n}}{\theta^{3n}}$ if $t \in [0,\theta]$. Then $f_T(t) = F_T'(t) = \frac{3nt^{3n-1}}{\theta^{3n}} I_{[0,\theta]}(t)$.

(c) Yes. Use the MLR definition.

(d) The rejection region is

$$P(X_n > c|\theta = 1) = \int_c^1 3nt^{3n-1}dt = 1 - c^{3n} = 0.05,$$

then $c = 0.95^{1/6}$ if $n = 2$.

(e) Choose $\frac{T}{\theta}$ as a pivot quantity, since $W = \frac{T}{\theta} \sim Beta(3n, 1)$ with the pdf $f_W(w) = 3nw^{3n-1}I(0 \le w \le 1)$. We need to choose a and b such that

$$P(a \le \frac{T}{\theta} \le b) = P(\frac{T}{b} \le \theta \le \frac{T}{a}) = 0.95, \quad 0 \le a < b \le 1,$$

which is equivalent to $\int_a^b 3nw^{3n-1}dw = b^{3n} - a^{3n} = 0.95$. The 95% confidence interval is $[\frac{T}{b}, \frac{T}{a}]$, whose expected length is $(\frac{1}{a} - \frac{1}{b})E(T) = \frac{b-a}{ab}E(T)$. Since $f_W(w)$ is increasing, $b-a$ is minimized and ab is maximized if $b = 1$. Then the shortest interval will have $b = 1$ and $a = (1-0.95)^{1/(3n)} = 0.05^{1/(3n)}$. So the shortest 95% is $[X_{(n)}, X_{(n)}/(0.05^{\frac{1}{3n}})]$.