STAT 700 Sample Final Exam Problems, Dec. 13, 2022

Instructions. You may use a single notebook page of memory aids for the exam. There will be 4 or 5 problems. The ones given here may be a little more detailed, in number of sub-parts, than the problems you will see on the final.

You will be asked to explain your reasoning in words, and it is important that you do so, referring to standard results whenever you can state them.

(1). In certain contexts, genetic data are observed in the form of discrete (multinomial) X_i (equal to the number out of two of genetic loci of a specified type A) with distinct possible values 0, 1, 2 occurring with respective probability masses

$$p_X(0) = (1-\theta)^2 \quad , \qquad p_X(1) = 2\theta\,(1-\theta) \quad , \qquad p_X(2) = \theta^2 \tag{1}$$

(a) If the dataset $\underline{X} = (X_1, ..., X_n)$ has jointly independent components with this distribution, then show that its probability mass function is of exponential family form with the scalar parameter $\theta \in (0,1)$. Give the Maximum Likelihood estimator of θ.

(b) Is this exponential family a curved expeonential family or one of full rank (in terms of its natural parameter of lowest possible dimension) ?

(c) Re-do parts (a)-(b) of this problem if the probability masses are instead

$$p_X(0) = \theta/2 \quad , \qquad p_X(1) = (1-\theta)(1+\theta/2) \quad , \qquad p_X(2) = \theta^2/2 \tag{1'}$$

Note: I would not ask you for the MLE in part (c) on the exam, but if you define $N_j = \sum_{i=1}^n I_{[X_i=j]}$ *for* $j = 0, 1, 2$*, then the MLE is given explicitly as the solution of a quadratic equation with coefficients involving the* N_j *variables.*

(d) Show that the sufficient statistic in (c) is not complete when $n \geq 2$.
Hint: evaluate $E(N_0^2 - (n-1)N_2/2 - N_0)$.

(2). (*Compare problems 2 and 3 from the Fall 2009 in-class test linked to the course web-page.*) Suppose that data $\underline{X} \sim f(\underline{x}, \theta)$ for $\theta \in \Theta \subset \mathbb{R}^d$ has the properties that:

(i). $T(\mathbf{X})$ is a sufficient statistic for θ, and
(ii). The parameter θ is identifiable from $\underline{X}$.

Then prove that θ is identifiable from $T(\underline{X})$.
Hint: you may do this either if $\underline{X}$ *is discrete or if there exists a smooth and smoothly invertible mapping* $\psi : \underline{x} \mapsto (T(\underline{x}), \underline{y})$.

(3). Suppose that X_i for $i = 1, \ldots, n$ are *iid* Geometric(θ) random variables, with probability mass function $p(x, \theta) = (1-\theta)^{x-1}\,\theta$ for $k = 1, 2, \ldots$ and $\theta \in (0, 1)$.

(a). Show that the data $\underline{X} = (X_1, \ldots, X_n)$ have exponential family form, and give the natural parameter η, natural parameter space $\mathcal{E}$, and function $B(\theta) = A(\eta)$.
Hint: $\eta = \log(1-\theta)$ and $B(\theta) = \log((1-\theta)/\theta) = \eta - \log(1-e^{\eta}) = A(\eta)$.

(b). Use (a) to exhibit the moment generating function of the natural sufficient statistic $T(\underline{X})$.

(c). Exhibit a family of conjugate prior densities for the natural parameter θ. (*Note: start by giving the conjugate prior family for η and transform it.*)

(d). Find the UMVUE of θ based on $\underline{X}$.

(e). Without calculating the variance of the estimator in (d), will it be equal to the lower bound for all unbiased estimators of θ ?

(f). What is the lower bound for the variance of all unbiased estimators of $E(X_1)$ based on $\underline{X}$? Will it be achieved by some estimator ?

(4). Suppose that X_i for $i = 1, \ldots, n$ is a sample of Poisson(λ) random variables on the parameter space $\lambda \in (0, \infty)$, and that we want to find Bayes optimal estimators of λ based on these data for the loss function $L(\lambda, a) = \lambda^{-2}\,(\lambda - a)^2$ using the prior density $\pi(\lambda) \sim$ Gamma(α, β) for $\alpha \geq 2$.

(a). Find the posterior density of λ, and explain why it can depend on $\underline{X}$ only through the sufficient statistic $\bar{X}$ for λ.

(b). Find the Bayes posterior risk function $r_\pi(g(\bar{X})) = E(L(\lambda, g(\bar{X})) \,|\, \underline{X})$ for the estimator $g(\bar{X})$.

(c). Using the same prior density as in (a), find the Bayes-optimal estimator of λ, i.e. the estimator $g(\bar{X})$ minimizing $r_\pi(g(\bar{X}))$.

(5). Suppose that predictor values X_i, $i = 1, \ldots, 10$, are fixed, and that $Y_i \sim$ Binom(1, $e^{\alpha+\beta X_i}/(1 + e^{\alpha+\beta X_i})$). Show that this is an exponential family dataset, and explain how you know that

(i) if $Y_i = 0$ for all $i = 1, \ldots 10$, the Maximum Likelihood estimator for (α, β) does not exist, and

(ii) if $Y_i = 0$ for $i = 1, \ldots 9$ and $Y_{10} = 1$, the MLE exists and is unique.

(6). Suppose that for $i = 1, \ldots, 50$, the random variables form a sample from the probability mass function $p(k) = 1/N$ for $k = 1, \ldots, N$ for some unknown N. Find and justify a UMP test of size exactly $\alpha = 0.05$ for the hypothesis $H_0 : \ N \leq 200$ versus $H_A : \ N > 200$. Find the power of your test against the parameter value $N = 250$, and the p-value of your test if the observed data sample of size 50 has $\max(X_1, \ldots, X_{50}) = 220$.