Stats 210A, Fall 2024
Homework 3

Due on: Wednesday, Sep. 25

Problem 1 (Multinomial subfamilies)**.** The multinomial family is a multi-category version of the binomial, it measures the number of times each category comes up if we sample a d-category random variable with distribution π on n independent trials. Throughout this problem assume $d \geq 3$.

If $X \sim \text{Multinom}(n, \pi)$, with all $\pi_j > 0$ and $\sum_j \pi_j = 1$, then X has density

$$p_\pi(x) = \pi_1^{x_1} \pi_2^{x_2} \cdots \pi_d^{x_d} \cdot \frac{n!}{x_1! x_2! \cdots x_d!}$$

Note: The coordinates of $X = (X_1, \ldots, X_d)$ are neither independent nor identically distributed.

(a) Rewrite the densities as a $(d-1)$-parameter exponential family, giving an explicit form for $T(x)$, $h(x)$, η, and $A(\eta)$. Show whether $X = (X_1, \ldots, X_d)$ is complete sufficient, minimal sufficient, or neither.

(b) Suppose a certain gene has two alleles **A** and **a**, and $\theta \in (0, 1)$ is the unknown prevalence of allele **a** in a well-mixed population. Then the proportion of people in the population with genotypes **aa**, **Aa**, and **AA** is θ^2, $2\theta(1-\theta)$, and $(1-\theta)^2$, respectively.

We can estimate θ by sampling n independent individuals from the population and counting the number who have each genotype. These counts will have a joint multinomial distribution with probability parameter

$$\pi(\theta) = (\theta^2, 2\theta(1-\theta), (1-\theta)^2).$$

Hence, scientific considerations might lead us to use the multinomial subfamily indexed by θ:

$$\mathcal{P} = \{\text{Multinom}(n, \pi(\theta)) : \ \theta \in (0, 1)\}.$$

Can $\mathcal{P}$ be written as a one-parameter exponential family? Find a minimal sufficient statistic for $\mathcal{P}$, and show whether or not it is complete.

(c) Now suppose our population is a mixture of two populations with different prevalences θ_1 and θ_2 for allele **a**. Define $\gamma \in (0, 1)$ as the proportion of individuals from population 1. Assume that θ_1, θ_2 are known and only γ is unknown. Since θ_1 and θ_2 are known it may be convenient to write the mixture probabilities as

$$\pi(\gamma) = \gamma \pi^{(1)} + (1-\gamma)\pi^{(2)}, \quad \text{for } \pi^{(k)} = (\theta_k^2, 2\theta_k(1-\theta_k), (1-\theta_k)^2), \ \ k = 1, 2.$$

Now suppose that we again sample n individuals form our unknown mixture, giving another one-parameter subfamily $\mathcal{Q}$ indexed by γ. Can $\mathcal{Q}$ be written as a one-parameter exponential family? Find a minimal sufficient statistic for $\mathcal{Q}$, and show whether or not it is complete.

Moral: The structure of the families and subfamilies determines the properties of the sufficient statistic.

Problem 2 (Gamma family)**.** The gamma family is a two-parameter family of distributions on $\mathbb{R}_+ = [0, \infty)$, with density

$$p_{k,\theta}(x) = \frac{x^{k-1}e^{-x/\theta}}{\Gamma(k)\theta^k}$$

with respect to the Lebesgue measure on $\mathbb{R}_+$. $k > 0$ and $\theta > 0$ are respectively called the shape and scale parameters, and $\Gamma(k)$ is the gamma function, defined as

$$\Gamma(k) = \int_0^\infty x^{k-1}e^{-x}\,dx.$$

The gamma distribution generalizes the exponential distribution

$$\text{Exp}(\theta) = \theta^{-1}e^{-x/\theta} = \text{Gamma}(1, \theta)$$

and the chi-squared distribution

$$\chi_d^2 = \frac{x^{d/2-1}e^{-x/2}}{\Gamma(d/2)2^{d/2}} = \text{Gamma}(d/2, 2).$$

(a) Show that the Gamma is a 2-parameter exponential family by putting it into its canonical form. Find the natural parameter, sufficient statistic, carrier density, and log-partition function (**Note**: there are multiple valid ways of doing this).

(b) Find the mean and variance of $X \sim \Gamma(k, \theta)$.

(c) Find the moment generating function of $X \sim \Gamma(k, \theta)$:

$$M_X(u) = \mathbb{E}_{k,\theta}[e^{uX}],$$

and use it to find the distribution of $X_+ = \sum_{i=1}^n X_i$ where $X_1, \ldots, X_n$ are mutually independent with $X_i \sim \text{Gamma}(k_i, \theta)$.

You may use without proof the following uniqueness result about MGFs: If Y and Z are two random variables whose MGFs coincide in a neighborhood of 0 ($\exists\delta > 0$ for which $M_Y(u) = M_Z(u) < \infty$ for all $u \in [-\delta, \delta]$), then Y and Z have the same distribution.

Problem 3 (Interpretation of completeness)**.** The concept of *completeness* for a family of measures was introduced in Lehmann and Scheffé (1950) as a precursor to their definition, in the same paper, of a complete statistic. The definition of a complete family did not stick, and lives on only in the (consequently confusingly named) idea of complete statistic (in particular it has nothing to do with the definition of a *complete measure* that you can find on Wikipedia).

If $\mathcal{P} = \{P_\theta :\ \theta \in \Theta\}$ is a family of measures on $\mathcal{X}$, we say that $\mathcal{P}$ is *complete* if

$$\int f(x)\,dP_\theta(x) = 0,\ \forall\theta \quad \Rightarrow \quad P_\theta(\{x :\ f(x) \neq 0\}) = 0,\ \forall\theta.$$

This can be interpreted as an inner product $\langle f, P_\theta\rangle = \int f\,dP_\theta$, where $f \perp P_\theta$ if $\langle f, P_\theta\rangle = 0$. Then, the family is **not** complete if there is some nonzero function f that is orthogonal to every P_θ. We will try to gain some intuition for this definition and, thereby, for the definition of a complete statistic.

For the following parts, let $\mathcal{P} = \{P_\theta :\ \theta \in \Theta\}$ be a family of probabilty measures on $\mathcal{X}$, assume $T(X)$ is a statistic, and let $\mathcal{T} = T(\mathcal{X})$ be the range of the statistic $T(X)$. Let $\mathcal{P}^T = \{P_\theta^T :\ \theta \in \Theta\}$ denote the induced model of push-forward probability measures on $\mathcal{T}$ denoting the possible distributions of $T(X)$:

$$P_\theta^T(B) = P_\theta(T^{-1}(B)) = \mathbb{P}_\theta(T(X) \in B).$$

(a) Show that $T(X)$ is a complete statistic for the family $\mathcal{P}$ if and only if $\mathcal{P}^T$ is a complete family.

(b) Assume (for this part only) that $\mathcal{X}$ is a finite set, i.e. $\mathcal{X} = \{x_1, \ldots, x_n\}$ for some $n < \infty$, and assume without loss of generality that every $x \in \mathcal{X}$ has $P_\theta(\{x\}) > 0$ for at least one value of θ (otherwise we could truncate the sample space).

Let $p_\theta(x) = \mathbb{P}_\theta(X = x) \geq 0$, and $v^\theta = (p_\theta(x_1), \ldots, p_\theta(x_n)) \in \mathbb{R}^n$. Show that $\mathcal{P}$ is complete if and only if $\text{Span}\{v^\theta : \theta \in \Theta\} = \mathbb{R}^n$.

(c) Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Pois}(\theta)$ for $\theta \in \Theta = \{\theta_1, \ldots, \theta_m\}$ with $2 \leq m < \infty$. Find a sufficient statistic that is minimal but not complete (prove both properties).

(d) **Optional:** (Not graded, no extra points) In the same scenario but with $\Theta = \pi\mathbb{Z}_+ = \{0, \pi, 2\pi, \ldots\}$, show that the same statistic is minimal but not complete.

Hint: Recall the Taylor series

$$\sin(\theta) = \theta - \frac{\theta^3}{3!} + \frac{\theta^5}{5!} - \frac{\theta^7}{7!} + \cdots.$$

(e) **Optional:** (Not graded, no extra points) Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Pois}(\theta)$ for $\theta \in \Theta$, and assume that Θ has an accumulation point at 0, i.e. Θ includes an infinite sequence of positive values $\theta_1, \theta_2, \ldots \in \Theta$ such that $\lim_{m\to\infty} \theta_m = 0$. Find a complete sufficient statistic and prove it is complete sufficient.

Hint: suppose f is a counterexample function; what is $f(0)$? It may be helpful to recall that $\int f \, d\mu$ is undefined unless either $\int \max(0, f(x)) \, d\mu(x)$ or $\int \max(0, -f(x)) \, d\mu(x)$ is finite; as a result $\int f \, d\mu = 0 \Rightarrow \int |f| \, d\mu < \infty$.

Moral 1: The definition of a complete statistic is easier to remember if we recall its interpretation as saying that the set of distributions P_θ^T "spans" a certain vector space, so that only the zero function is orthogonal to all P_θ^T.

Moral 2: If $\mathcal{P} = \{P_\eta : \eta \in \Xi\}$ is a full-rank exponential family with natural parameter η, meaning Ξ contains an open set, our result from class allows us to prove completeness of $T(X)$. But the converse is far from true: it is possible for T to be complete if Ξ is discrete, or even finite.

Problem 4 (Ancillarity in location-scale families)**.** In a parameterized family where $\theta = (\zeta, \lambda)$, we say a statistic T is *ancillary for* ζ if its distribution is independent of ζ; that is, if $T(X)$ is ancillary in the subfamily where λ is known, for each possible value of λ.

Suppose that $X_1, \ldots, X_n \in \mathcal{X} = \mathbb{R}$ are an i.i.d. sample from a *location-scale family* $\mathcal{P} = \{F_{a,b}(x) = F((x-a)/b) : a \in \mathbb{R}, b > 0\}$, where $F(\cdot)$ is a known cumulative distribution function. The real numbers a and b are called the *location* and *scale* parameters respectively.

Note: It is *not* enough to prove ancillarity of the coordinates; the joint distribution of the statistic shouldn't depend on the relevant parameter.

(a) Show that the vector of differences $(X_1 - X_i)_{i=2}^n$ is ancillary for a.

(b) Show that the vector of ratios $\left(\frac{X_1-a}{X_i-a}\right)_{i=2}^n$ is ancillary for b. (Note: this is only a statistic when a is known).

(c) **Optional:** (Not graded, no extra points) Show that the vector of difference ratios $\left(\frac{X_1-X_i}{X_2-X_i}\right)_{i=3}^n$ is ancillary for (a, b).

(d) Let $X_1, \ldots, X_n$ be mutually independent with $X_i \sim \text{Gamma}(k_i, \theta)$. Show that $X_+ = \sum_{i=1}^n X_i$ is independent of $(X_1, \ldots, X_n)/X_+$.

Moral: Location-scale families have common structure that we can exploit in some problems.

Problem 5 (Complete sufficient statistic for a nonparametric family)**.** Consider an i.i.d. sample from the nonparametric family of *all* distributions on $\mathbb{R}$:

$$X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} P,$$

Formally we can write this model as $\mathcal{P} = \{P^n : \; P \text{ is a probability measure on } \mathbb{R}\}$. Let $T(X) = (X_{(1)}, \dots, X_{(n)})$ denote the vector of order statistics.

(a) For a finite set of size m, $\mathcal{Y} = \{y_1, \dots, y_m\} \subseteq \mathbb{R}$, consider the subfamily $\mathcal{P}_{\mathcal{Y}}$ of distributions supported on $\mathcal{Y}$:

$$\mathcal{P}_{\mathcal{Y}} = \{P^n : \; P(\mathcal{Y}) = 1\} \subseteq \mathcal{P}.$$

Show that $T(X)$ is complete sufficient for this family.

Hint: It may help to review different ways to parameterize the multinomial family.

(b) Show that the vector of order statistics $T(X) = (X_{(1)}, \dots, X_{(n)})$ is a complete sufficient statistic for $\mathcal{P}$.

(c) Next, consider the restricted subfamily

$$\mathcal{Q}_k = \{P^n : \; \mathbb{E}_P[|X_1|^k] < \infty\} \subseteq \mathcal{P},$$

and define the sample mean and variance respectively as

$$\overline{X} = \frac{1}{n}\sum_{i=1}^{n} X_i, \quad S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X})^2.$$

Show that $\overline{X}$ is the UMVU estimator of $\mathbb{E}_P X_1$ in $\mathcal{Q}_1$, and S^2 is the UMVU estimator of $\text{Var}_P(X_1)$ in $\mathcal{Q}_2$.

Moral: Without any restrictions on the family $\mathcal{P}$, we can't do much better than estimating population quantities with sample quantities (when the sample quantities are unbiased). In the case of the mean, for examples, $\overline{X}$ is always available as an unbiased estimator of $\mathbb{E}X$, but if we impose additional assumptions on the family then we might be able to do better.

References

EL Lehmann and Henry Scheffé. Completeness, similar regions, and unbiased estimation: Part i. *Sankhyā: The Indian Journal of Statistics*, pages 305–340, 1950.