Stats 210A, Fall 2023
Homework 3

Due date: Wednesday, Sep. 20

You may disregard measure-theoretic niceties about conditioning on measure-zero sets, almost-sure equality vs. actual equality, "all functions" vs. "all measurable functions," etc. (unless the problem is explicitly asking about such issues).

1. 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

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

This can be interpreted as an inner product $\langle f, P_\theta\rangle = \int f\, \mathrm{d}P_\theta$, 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 $\mathrm{Span}\{v^\theta : \theta \in \Theta\} = \mathbb{R}^n$.

(c) Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Pois}(\theta)$ 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) 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 $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \mathrm{Pois}(\theta)$ 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 $\int f \, \mathrm{d}\mu$ is undefined unless either $\int \max(0, f(x)) \, \mathrm{d}\mu(x)$ or $\int \max(0, -f(x)) \, \mathrm{d}\mu(x)$ is finite; as a result $\int f \, \mathrm{d}\mu = 0 \Rightarrow \int |f| \, \mathrm{d}\mu < \infty$.

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.

2. 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:** recall it is *not* enough to prove ancillarity of the coordinates.)

(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) 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.

3. Unbiased estimation in replicated studies

One focal issue in the ongoing scientific replication crisis is the "file drawer problem," i.e. the tendency of researchers to report findings (or of journals to publish them) only if they have a p-value less than 0.05. Replication studies typically represent cleaner estimates of the results under study, since they are reported regardless of whether they are statistically significant. This is one of the reasons that replication studies often find much smaller effect size estimates than the original studies: if the original study had gotten a good estimate of the (small) true effect, we wouldn't have heard about it.

We can introduce a toy model for a replicated study where the original study is $X_1 \sim N(\mu, 1)$ and the replication study is $X_2 \sim N(\mu, 1)$, but we only observe the study pair given that $X_1 > c$ for some significance cutoff $c \in \mathbb{R}$, e.g. $c = 1.96$. In other words, the distribution for a study pair conditional on our observing it is

$$\begin{aligned} p_\mu(x_1, x_2) &= \mathbb{P}_\mu(X_1 = x_1, X_2 = x_2 \mid X_1 > c) \\ &= \frac{\phi(x_1 - \mu) 1\{x_1 > c\}}{1 - \Phi(c - \mu)} \phi(x_2 - \mu), \end{aligned}$$

where $\phi(x) = \frac{1}{\sqrt{2\pi}} e^{-x^2/2}$ is the standard normal pdf and $\Phi(x) = \int_{-\infty}^x \phi(u) \, \mathrm{d}u$ is the standard normal cdf. We will consider the problem of estimating μ after observing a study pair.

Arguably, we should only care about the *conditional* bias or risk of an estimator, given that we actually get to see the data, since the conditional distribution more accurately describes the set of published results. Thus, all questions below about bias, admissibility, UMVU, etc. should be answered in terms of the conditional distribution given that $X_1 > c$ (i.e., with densities $p_\mu(x_1, x_2)$ above), *not* in terms of the marginal distribution (whose densities would be $\phi(x_1 - \mu)\phi(x_2 - \mu)$.) For example, in part (a) it would not be true to say that $\overline{X}$ is marginally biased, but I want you to show it is conditionally biased given that it is observed.

(a) Show that $\overline{X} = (X_1 + X_2)/2$ is an upwardly biased estimator of μ (we can call this the *naive* estimator since it ignores the selection bias).

(b) Show that X_2 is unbiased for μ, but it is inadmissible under any strictly convex loss function (we can call this the *data splitting* estimator since we ignore X_1, which was used for selection, and use the fresh data X_2.)

(c) Show that the UMVU estimator for μ is

$$\delta(\overline{X}) = \overline{X} - \frac{1}{\sqrt{2}}\,\zeta\left(\sqrt{2}(c - \overline{X})\right),$$

where

$$\zeta(x) = \mathbb{E}_{Z\sim N(0,1)}[Z \mid Z > x] = \frac{\int_x^\infty u\phi(u)\,\mathrm{d}u}{1 - \Phi(x)}.$$

Hint: It may help to note that $X_1 + X_2$ is marginally independent of $X_1 - X_2$ (but note they are **not** conditionally independent given $X_1 > c$.)

(d) Show that

$$\lim_{\overline{X}\to\infty} \delta(\overline{X}) - \overline{X} = 0.$$

In other words, if $\overline{X} \gg c$, then $\delta(\overline{X}) \approx \overline{X}$, the naive estimator. Can you give any intuition for why this limit makes sense?

(e) **Optional:** (not graded, no extra points). Show that

$$\lim_{\overline{X}\to-\infty} \delta(\overline{X}) - (X_2 + (X_1 - c)) = 0,$$

and furthermore that for any $\varepsilon > 0$, we have

$$\lim_{\overline{X}\to-\infty} \mathbb{P}(X_1 - c > \varepsilon \mid \overline{X}, X_1 > c) \to 0.$$

In other words, if $\overline{X} \ll c$, we have $\delta(\overline{X}) \approx X_2 + (X_1 - c) \approx X_2$, the data splitting estimator. Can you give any intuition for why this limit makes sense?

Hint: It may be helpful to use the tail inequality

$$\left(\frac{1}{x} - \frac{1}{x^3}\right)\phi(x) \le 1 - \Phi(x) \le \frac{1}{x}\phi(x),$$

for $x > 0$.

Moral: This is a nice estimator that transitions adaptively between the data splitting estimator (when X_1 is subject to extreme selection bias) and the unadjusted sample mean (when X_1 is nearly unaffected by selection bias). It manages to do this even though we don't know how bad the selection bias is, since that depends on μ. It would be difficult to come up with an estimator like this without the theory of exponential families and UMVU estimators, specifically the idea of Rao-Blackwellization.

4. Poisson UMVU estimation

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Pois}(\theta)$ and consider estimating

$$g(\theta) = e^{-\theta} = \mathbb{P}_\theta(X_1 = 0)$$

(a) Find the UMVU estimator for $g(\theta)$ by Rao-Blackwellizing a simple unbiased estimator. You may use without proof the fact that $(X_1, \ldots, X_n) \sim \text{Multinom}(t, (n^{-1}, \ldots, n^{-1}))$ given $\sum_{i=1}^n X_i = t$.

(b) Find the UMVU estimator for $g(\theta)$ directly, using the power series method from class.

Moral: This problem is for practice deriving UMVU estimators using the two methods from class.

5. Complete sufficient statistic for a nonparametric family

Consider an i.i.d. sample from the nonparametric family of *all* distributions on $\mathbb{R}$:

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

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

(a) For a finite set of size m, $\mathcal{Y} = \{y_1, \ldots, y_m\} \subseteq \mathbb{R}$, 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 $T(X) = (X_{(1)}, \ldots, X_{(n)})$ 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 $\mathcal{Q}_2$.

(d) In the original family $\mathcal{P}$, find the UMVU estimator of the probability

$$\pi_c = \mathbb{P}_P(X \le c).$$

Note: If we come up with estimators for every c we can "assemble" them all into an estimator for the CDF of P.

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.