Stats 210A, Fall 2025
Homework 4

Due on: Wednesday, Oct. 1

Instructions: 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).

Problem 1 (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 availableas an unbiased estimator of $\mathbb{E}X$, but if we impose additional assumptions on the family then we might be able to do better.

Problem 2 (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 $\Phi(x) = \int_{-\infty}^{x} \phi(u)\, du$ 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

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

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

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

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. You can read more about problems like this in Hung and Fithian (2020).

Problem 3 (Bayesian law of large numbers)**.** Let $p(x)$ and $q(x)$ denote two strictly positive probability densities with respect to a common dominating measure μ. The *Kullback–Leibler divergence* between p and q is defined as

$$D(p\|q) = \int_{\mathcal{X}} p(x) \log \frac{p(x)}{q(x)}\, d\mu(x).$$

(a) Show that $D(p\|q) \geq 0$, with equality only in the case that $p(X) = q(X)$ almost surely

Hint: recall that $\log(1 + x) \leq x$ for all $x > -1$.

(b) Consider a dominated likelihood model $\mathcal{P} = \{p_\theta(x) : \ \theta \in \Theta\}$, where the parameter space Θ is a finite set, and the densities are strictly positive on $\mathcal{X}$. Let λ denote a prior density w.r.t. the counting measure on Θ, and consider the Bayes posterior after observing a sample $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta_0}(x)$ for some *fixed* value θ_0 (that is, we are examining the *frequency* properties of the *Bayesian* posterior distribution). Assume that all the densities are distinct; that is, $p_{\theta_1}(X) = p_{\theta_2}(X)$ almost surely if and only if $\theta_1 = \theta_2$.

If the prior λ puts positive mass on all values in Θ, show that as $n \to \infty$, the posterior density eventually concentrates nearly all its mass on the true value θ_0. That is,

$$\mathbb{P}_{\theta_0}\left[\lambda(\theta_0 \mid X_1, \ldots, X_n) \geq 1 - \varepsilon\right] \to 1, \quad \text{for all } \varepsilon > 0.$$

Hint: apply the law of large numbers and see if you can find a way to use part (a).

Moral: At least for a finite parameter space, the Bayes estimator always converges to the right answer as long as we put positive mass on the right answer. This result can be generalized with more effort to continuous parameter spaces under some regularity conditions on the likelihood function, similar to the types of conditions we will use to guarantee the MLE is consistent.

The requirement that the prior density should be nonzero everywhere is sometimes called Cromwell's Rule, after Oliver Cromwell's famous plea to the Church of Scotland: "I beseech you, in the bowels of Christ, think it possible that you may be mistaken."

Problem 4 (Fisher information for location and scale families)**.** This problem considers the Fisher information for families with location or scale structure. Your verbal explanations for each part will be graded leniently.

(a) Consider a location family

$$p_\theta(x) = p_0(x - \theta), \quad \text{for } \theta \in \mathbb{R},$$

where p_0 is some fixed probability density function with respect to the Lebesgue measure.

Show that the Fisher information for a single observation X is given by

$$J(\theta) = \int_{-\infty}^{\infty} \frac{\dot{p}_0(u)^2}{p_0(u)}\, du.$$

Explain in your own words why it makes sense that there should be no dependence on θ.

(b) Consider a scale family

$$p_\theta(x) = \frac{1}{\theta} p_0\left(\frac{x}{\theta}\right), \quad \theta > 0.$$

where p_0 is some fixed probability density function with respect to the Lebesgue measure.

Show that the Fisher information of a single observation X is given by

$$J(\theta) = \frac{1}{\theta^2} \int_{-\infty}^{\infty} \left[\frac{u\dot{p}_0(u)}{p_0(u)} + 1\right]^2 p_0(u)\, du.$$

Try to explain in your own words why it makes sense that the Fisher information should be proportional to θ^{-2}.

(c) If we instead parameterize the scale family using $\zeta = \log\theta$, show that the Fisher information $J(\zeta)$ of a single observation X does not depend on ζ. Explain in your own words why this makes sense.

References

Kenneth Hung and William Fithian. Statistical methods for replicability assessment. *The Annals of Applied Statistics*, 14(3):1063–1087, 2020.