Math 541b: Introduction to Mathematical Statistics (Fall 2022)
Homework 1

Due: **Mon 10/03**

1^o: *Estimating the support of* $\mathsf{Unif}([0, \theta])$.

Let $X_1, X_2, ..., X_n$ be an i.i.d. sample from the uniform density $f(x|\theta) = \frac{1}{\theta}\mathbb{1}\{x \in [0, \theta]\}$ with $\theta > 0$. The Cramér-Rao theorem would seem to imply that the variance of any unbiased estimator of θ is lower-bounded with $\frac{\theta^2}{n}$. Why cannot we apply it here?

- Construct an unbiased estimator that "violates" the Cramér-Rao bound. To this end, start with the MLE, compute its bias, and "correct" it, so that the resulting estimator is unbiased.
- Show that the Cramér-Rao bound is "violated" by computing the variance of this estimator.

2^o: *Estimating the median in a location family.*

For any $\mu \in \mathbb{R}$, let $\mathbb{P}_\mu$ be the distribution of $\mu^* + Z$, where Z has the median 0 and some (known) p.d.f. f such that $f(0) > 0$. We estimate μ^* from i.i.d. sample $X_{1:n} = (X_1, ..., X_n) \sim \mathbb{P}_{\mu^*}^{\otimes n}$ with odd n, by the *sample median* of $X_{1:n}$, defined as

$$\widehat{\mathrm{Med}}_n := X_{(\frac{n+1}{2})}$$

where $X_{(k)}$ is the k-th order statistic, i.e. k^{th} largest among $X_{1:n}$.
(Note that we do not have to care about possible ties, since all X_i's are different with probability 1).

(a) Show that $\widehat{\mathrm{Med}}_n$ is unbiased when f is symmetric, i.e. when $f(-x) = f(x)$ for all $x \in \mathbb{R}$.

Hint. *Use the tower rule:* $\mathbb{E}[\widehat{\mathrm{Med}}_n] = \mathbb{E}[\mathbb{E}[\widehat{\mathrm{Med}}_n|Y]]$ *for any random variable* Y*. Try to find the right random variable* Y *—supported on* $\{1, ...n\}$*—for which* $\mathbb{E}[\widehat{\mathrm{Med}}_n|Y] = 0$ *a.s.*

(b) Show that $\widehat{\mathrm{Med}}_n$ is the MLE when Z has the standard Laplace distribution: $f(u) = \frac{1}{2}e^{-|u|}$.
Hint: *what is the derivative of* $\ell(u) := \log f(u)$*? Does it matter that* $\ell'(0)$ *is not defined?*

~~(c) Compute the Fisher information and $\mathrm{Var}(\widehat{\mathrm{Med}}_n)$ in this case, and verify that $\widehat{\mathrm{Med}}_n$ here achieves the Cramér-Rao bound for *any n*.~~

(c$'$) Compute the variance of $\widehat{\mathrm{Med}}_n$ in this situation (i.e. Laplace distribution) in the cases $n = 1$ and $n = 3$. Compare with the Cramér-Rao bound.

Hint. *Use the following fact: if* $X_{1:n}$ *is an i.i.d. sample from a law with c.d.f.* $F_X(x)$*, then the c.d.f. of its* k^{th} *order statistic is*

$$F_{X_{(k)}}(x) = \sum_{j=0}^{k-1} \binom{n}{j} F_X(x)^{n-j}(1 - F_X(x))^j$$

Another useful fact is that $\widehat{\mathrm{Med}}_n$ is shift-equivariant*: if we shift the distribution by a constant $a \in \mathbb{R}$, then the distribution of $\widehat{\mathrm{Med}}_n$ will simply be shifted in the same way. This allows to assume w.l.o.g. that $\mu_* = 0$ in your calculations.*

(d) Find the *asymptotic variance* of $\widehat{\mathrm{Med}}_n$ in the general situation, i.e. only assuming that $f(0) > 0$.

Hint. *Use the so-called "delta-method:" if $\widehat{\theta}_n$ estimates $\theta \in \mathbb{R}$ in such a way that*

$$\sqrt{n}(\widehat{\theta}_n - \theta) \underset{n\to\infty}{\rightsquigarrow} \mathcal{N}(0, \sigma^2),$$

and $g(\cdot)$ is differentiable at θ, then $\sqrt{n}[g(\widehat{\theta}_n) - g(\theta)] \underset{n\to\infty}{\rightsquigarrow} \mathcal{N}(0, \sigma_g^2)$ with $\sigma_g^2 = \sigma^2(g'(\theta))^2$.

What you can say about this in the light of the previous example (with the Laplacian density)?

(e) Give another example of a symmetric $f(u)$ such that $\widehat{\mathrm{Med}}_n$ does not attain the Cramér-Rao bound in the corresponding family.

3^o: *MLE for the ratio of two independent exponential distributions.*
Let $X \sim \mathsf{Exp}(\lambda)$ and $Y \sim \mathsf{Exp}(\mu)$ be independent, where $\mathsf{Exp}(\lambda)$ is the distribution with p.d.f.

$$\lambda e^{-\lambda x}, \quad x > 0.$$

Let $Z = X/Y$, and consider an i.i.d. sample $(Z_1, ..., Z_n)$ with each Z_i being distributed as Z.

(a) Without deriving the distribution of Z explicitly, argue that it depends only on $\theta := \mu/\lambda$, not on λ, μ separately.

Hint: for $\alpha > 0$, what is the distribution of αX?

(b) Show that the p.d.f. of Z is

$$f(z|\theta) = \frac{\theta}{(z+\theta)^2}, \quad z > 0.$$

Does this distribution have an expectation?

Hint: *you might want to start with the c.d.f.*

(c) Show that $\widehat{\theta}_n$, the MLE of θ from $Z_1, ..., Z_n$, satisfies the following equation:

$$\sum_{i=1}^{n} F(Z_i|\widehat{\theta}_n) = \frac{n}{2}$$

where $F(z|\theta)$ is the c.d.f. of Z. Argue that the solution always exists and is unique.

(d) Comment on the above equation, explaining why the right-hand side has the factor $\frac{1}{2}$. To this end, show that for <u>any continuous distribution</u> $\mathbb{P}_\theta$ with c.d.f. $H(t;\theta)$, it holds that

$$\mathbb{E}_{T\sim\mathbb{P}_\theta}[H(T;\theta)] = \frac{1}{2}.$$

Then explain the MLE equation in this context.

Hint: *You may draw an analogy with the method of moments.*

(e) Show that for the distribution whose p.d.f. you found in (b), θ also happens to be the median.

~~(f)* As such, in the setup of (a)-(b) we can also estimate ρ with the sample median $\widehat{\text{Med}}_n$ whose properties we have studied in **2°**, and compare it with $\widehat{\rho}$. Recalling the result of **2°**, part (d), show that here $\text{Var}(\widehat{\rho}_n) \le \text{Var}(\widehat{\text{Med}}_n)$. ***Hint:*** *Use Jensen's inequality.*~~

4°: *Tail bounds for the Gaussian distribution.*

Let $\phi(\cdot)$ be the p.d.f. of $\mathcal{N}(0,1)$, i.e. $\phi(t) = \frac{1}{\sqrt{2\pi}} e^{-\frac{t^2}{2}}$. For any $u \ge 0$, let $\Phi(u) := \int_{t \ge u} \phi(t) dt$.

(a) Prove the following bounds (holding for all $u \ge 0$):

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

Hint 1: *Try to prove the upper bound first.*

Hint 2: *Integration by parts is the way here; use it first to prove the upper bound, and then for the lower bound.*

(b) Capitalizing on the trick you have just figured out to get the lower bound from the upper bound, prove a new upper bound:

$$\Phi(u) \le \left(\frac{1}{u} - \frac{1}{u^3} + \frac{3}{u^5}\right) \phi(u).$$

Note that this bound is sharper than the previous one for large enough u.

(c)* If we continue applying this approach iteratively, what bounds shall we get after k such "iterations?"

5°: *Bias-variance tradeoff and the James-Stein estimator.* Consider the problem of estimating the mean μ in the multivariate Gaussian location family

$$\mathbb{P}_\mu = \mathcal{N}(\mu, I), \quad \mu \in \mathbb{R}^d, \tag{1}$$

from a single observation $X \sim \mathbb{P}_\mu$. Note that here, X itself is the maximum likelihood estimator (MLE) for μ. Defining for any estimator $\hat{\mu} = \hat{\mu}(X)$ of μ the variance

$$\text{Var}_\mu[\hat{\mu}] := \mathbb{E}_\mu[\|\hat{\mu} - \mathbb{E}[\hat{\mu}]\|^2]$$

and the quadratic risk

$$\text{Risk}_\mu[\hat{\mu}] := \mathbb{E}_\mu[\|\hat{\mu} - \mu\|^2],$$

where $\|x\| := (\sum_i x_i^2)^{1/2}$ is the Euclidean norm of $x = (x_1, ..., x_d) \in \mathbb{R}^d$, we see that for any $\mu \in \mathbb{R}^d$,

$$\text{Risk}_\mu[X] = \text{Var}_\mu[X] = d.$$

Intuitively, it is hard to suspect that one can find a more reasonable estimator of μ than X. Yet, this turns out to be the case: one may improve over the MLE uniformly on the family (1) when $d > 2$. This celebrated result was established by James and Stein in 1976, and our goal is to reproduce it.

But first, let us establish the terminology.

Definition 1. *An estimator* $\hat{\mu}$ *is* dominated *by some other estimator* $\hat{\mu}'$ *if* $\text{Risk}_\mu[\hat{\mu}'] \le \text{Risk}_\mu[\hat{\mu}]$ *for any* μ*, and there exists a parameter value* $\bar{\mu}$ *such that* $\text{Risk}_{\bar{\mu}}[\hat{\mu}'] < \text{Risk}_{\bar{\mu}}[\hat{\mu}]$.

Definition 2. *An estimator* $\hat{\mu}$ *is called* admissible *if it is not dominated by any other estimator. Otherwise, it is called* inadmissible.

As statisticians, ideally we would like to compare two estimators over the whole family at once, without specifying a value of μ. Two admissible estimators cannot be compared this way, but at the very least we can rule out any *inadmissible* estimator, as for it there exists a uniformly better one.

You will show that the MLE is inadmissible when $d \ge 3$, by constructing a dominating estimator.

(a) Consider *shrinkage estimators* $\hat{\mu} = sX$ with $s \in \mathbb{R}$, and compute their risks for any s. Show that one can restrict attention to $s \in [0,1]$ (hence "shrinkage") by finding a dominating estimator for $\hat{\mu}$ with $s < 0$ or $s > 1$.

Hint: *look for an estimator* $\hat{\mu}' = s'X$ *with* $s' \in [0,1]$.

(b) Show that, for given μ, the best value of s—i.e., the one minimizing the risk—is given by

$$s^* = \frac{\|\mu\|^2}{d + \|\mu\|^2} = 1 - \frac{d}{d + \|\mu\|^2}.$$

(c) Unfortunately, $\hat{\mu}^* = s^* X$ is not a proper estimator. (Why?) Instead of it, one may consider

$$\left(1 - \frac{d}{\|X\|^2}\right) X,$$

which is an actual estimator. Can you explain the heuristic motivation behind this estimator?

(d)* **(Bonus.)** Assuming that $d \ge 2$, derive the *James-Stein estimator*

$$\hat{\mu}^{JS} = \left(1 - \frac{d-2}{\|X\|^2}\right) X \tag{2}$$

by minimizing over $\delta \in \mathbb{R}$ the risk of the estimator

$$\hat{\mu}^{\delta} = \left(1 - \frac{\delta}{\|X\|^2}\right) X$$

for a fixed μ. In order to show that $R(\delta) = \text{Risk}_\mu[\hat{\mu}^\delta]$ is minimized at $d-2$, use Stein's lemma:

Lemma 1. *Let* $X \sim \mathcal{N}(\mu, I)$ *and* $g(x)$ *be a function on* $\mathbb{R}^d$ *differentiable almost everywhere, and such that* $\mathbb{E}_\mu\left[\left|\frac{\partial}{\partial x_i} g(X)\right|\right] < \infty$ *and* $\mathbb{E}_\mu[|(X_i - \mu_i) g(X)|] < \infty$ *for any* $i \in [d] := \{1, 2, ..., d\}$. *Then*

$$\mathbb{E}_\mu[(X_i - \mu_i) g(X)] = \mathbb{E}_\mu\left[\frac{\partial}{\partial x_i} g(X)\right], \quad i \in [d].$$

When applying Stein's lemma to the right function $g(X)$, please do check the absolute integrability conditions in its premise, and explain why the argument does not work for $d = 1$. Finally, verify that $R(\delta)$ is strictly convex when $d \ge 3$ (thus $\hat{\mu}^{JS}$ indeed dominates the MLE).