# Stats 210A, Fall 2023
# **Optional** Homework 0

***Not* Due on**: Wednesday, Aug. 30

Lecture 1 included a "whirlwind tour" of measure theory at the heuristic level that we'll be using in class. Problem 1 is meant to give a little more intuition about densities and the others are meant to motivate measure-theoretic probability a bit.

Note this problem set has three problems; a typical problem set will have 5.

## 1. Densities

For a given point $x \in \mathcal{X}$, the *Dirac measure* is defined as

$$\delta_x(A) = 1\{x \in A\} = \begin{cases} 1 & \text{if } x \in A \\ 0 & \text{otherwise} \end{cases}.$$

Essentially, $\delta_x$ is the measure that puts a unit of mass on $x$ and none anywhere else.[1] Integrals wrt $\delta_x$ are defined as $\int f(u)\,\mathrm{d}\delta_x(u) = f(x)$.

Furthermore, suppose $\mu_1$ and $\mu_2$ are both measures on $\mathcal{X}$, and $a_1, a_2 \geq 0$. You may use without proof that the sum $\nu = a_1\mu_1 + a_2\mu_2$ is also a measure, and that for "nice enough" functions,

$$\int f(x)\,\mathrm{d}\nu(x) = a_1 \int f(x)\,\mathrm{d}\mu_1(x) + a_2 \int f(x)\,\mathrm{d}\mu_2(x).$$

(a) Let $x_1, x_2, \ldots, x_n$ be integers (not necessarily all distinct), and define two measures on the set $\mathbb{Z}$ of all integers: the counting measure $\#$ from class, and the *empirical distribution*

$$\widehat{P}_n(A) = \frac{1}{n}\sum_{i=1}^{n} \delta_{x_i}(A).$$

That is, $\widehat{P}_n(A)$ is the fraction of points that fall into the set $A$.

**Note:** if $x_1, \ldots, x_n$ are sampled from some distribution $P$ then $\widehat{P}_n$ is a natural nonparametric estimator of the measure $P$.

Show that $\widehat{P}_n$ is absolutely continuous with respect to $\#$ but not the other way around. What is the density of $\widehat{P}_n$ with respect to $\#$? Is it possible to define a density of $\#$ with respect to $\widehat{P}_n$?

(b) For $\mathcal{X} = [0, \infty)$, define the measure $\mu(A) = \lambda(A) + \delta_0(A)$, where $\lambda$ represents the Lebesgue measure. For fixed $\theta \in \mathbb{R}$, define the random variable

$$X = \max(0, Z) \text{ where } Z \sim N(\theta, 1),$$

what is the density of $X$'s distribution with respect to $\mu$?

(c) Consider two densities $p_1$ and $p_2$ with respect to some common measure $\mu$ on a space $\mathcal{X}$ (not necessarily the same $\mu$ from part (b)). Suppose $p_1$ and $p_2$ both result in the same measure $P$ defined by $P(A) = \int 1_A(x) p_i(x)\,\mathrm{d}\mu(x)$.

[1] In a sense this is defined identically to the indicator function $1_A(x)$, but we think of one as being a function of $x$ with $A$ fixed, and the other as a function of $A$ (a measure) with $x$ fixed.

Define the set $A = \{x : \ p_1(x) \neq p_2(x)\}$, and show that $\mu(A) = 0$ (**Hint:** consider sets like

$$A_n = \left\{ x : \ p_1(x) - p_2(x) \in \left[ \frac{1}{n+1}, \ \frac{1}{n} \right) \right\}$$

for $n = 1, 2, \ldots$)

Don't worry about whether the measure is well-defined for $A_n$ (i.e., whether these sets are measurable). They are in the sense we need them to be.

**2. A conditional probability paradox**

Let $X, Y \overset{\text{i.i.d.}}{\sim} N(0,1)$. This problem is meant to show that by carelessly conditioning on probability-zero events we can get ourselves into trouble. It is directly inspired by a calculation I personally flubbed a few years ago.

(a) Defining $S = X + Y$ and $D = X - Y$, show $S$ and $D$ are independent and conclude that

$$\mathbb{E}[X^2 + Y^2 \mid D] = D^2/2 + 1$$

(b) Now define the polar parameterization $(R, \Theta)$ with $R = \sqrt{X^2 + Y^2}$ and $\Theta \in [0, 2\pi)$ such that $X = R\cos\Theta$ and $Y = R\sin\Theta$. Show that $R$ is independent of $\Theta$ and conclude that

$$\mathbb{E}[X^2 + Y^2 \mid \Theta] = 2$$

(c) Use (a) and then (b) to find the expectation of $X^2 + Y^2$ conditional on the event $X = Y$. Can you come up with an intuitive explanation for how we could have arrived at two different answers?

**Moral:** Intuition may fail us when we condition on a measure-zero event, and in cases like this the meaning can be ambiguous and give different answers. Conditioning on a random variable, on the other hand, tends to give less ambiguous answers (there are still some ambiguities, similar to those we encounter in defining densities, but they don't really matter).

**3. Non-measurable sets**

This problem goes through a construction of a non-measurable set, meant to motivate measure theory from a real analysis perspective. It concerns the impossibility of defining "volume" for every subset of the unit interval $U = [0, 1)$.

For $x, y \in \mathbb{R}$ define the "wraparound addition" (modulo 1) as the fractional part of their sum:

$$x \oplus y = x + y - \lfloor x + y \rfloor.$$

Recall that for $x \in \mathbb{R}$ and $A \subseteq \mathbb{R}$ we define the set $x + A = \{x + a : \ a \in A\}$. Analogously, we can define

$$x \oplus A = \{x \oplus a : \ a \in A\} \subseteq U$$

Any reasonable definition of "volume" on the interval should have several properties:

(i) Additivity: $\lambda(\bigcup_{i=1}^{\infty} A_i) = \sum_{i=1}^{\infty} \lambda(A_i)$ if all $A_i \subseteq U$ and $A_i \cap A_j = \emptyset$ for all $i \neq j$.

(ii) Translation invariance: $\lambda(x \oplus A) = \lambda(A), \ \forall x \in U, A \subseteq U$.

(iii) Interval length: $\lambda([x, y)) = y - x, \ \forall 0 \leq x \leq y \leq 1$.

Assume that some measure $\lambda$ exists which satisfies (i)–(iii) and which is defined for all subsets of $U$. We will go through several steps to derive a contradiction.

(a) Define the function $A(x)$ mapping elements of $U$ to subsets of $U$, via $A(x) = x \oplus \mathbb{Q}$, where $\mathbb{Q}$ is the set of rational numbers. Show that $\lambda(A(x)) = 0$ for any $x$.

(b) Consider the range $\mathcal{R}_A = \{A(x) : \ x \in U\}$. Show that $\mathcal{R}_A$ is a collection of uncountably many subsets of $U$, all of which are disjoint from each other. That is, show that for any $x, y \in U$, we have either $A(x) = A(y)$ or $A(x) \cap A(y) = \emptyset$.

(c) Now, let $B \subseteq U$ denote a new set, which we construct by selecting a *single element* from each set $R \in \mathcal{R}_A$ (it doesn't matter which element; note this step uses the axiom of choice.)

Define a new function $C(x) = x \oplus B$ and define $\mathcal{R}_C = \{C(x) : \ x \in \mathbb{Q}\}$. Show that $\mathcal{R}_C$ is a collection of *countably* many subsets of $U$, all of which are disjoint from each other, and whose union is $U$.

(d) Show that no matter what value $\lambda(B)$ takes, $\lambda$ will have to violate one of the properties (i)–(iii) (**Hint:** what does the value of $\lambda(B)$ imply about $\lambda(U)$?)

Because the Lebesgue measure satisfies properties (i)–(iii), it follows that $\lambda$ must not be defined for every subset of $U$.

**Moral:** One motivation (but not the only motivation) for the idea of a $\sigma$-field is to exclude pathological counterexamples like this.

# Stats 210A, Fall 2023
# Homework 1

**Due date**: Wednesday, Sep. 6

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. Bias-Variance Tradeoff**

Consider a generic estimation setting where we observe $X \sim P_\theta$, for a model $\mathcal{P} = \{P_\theta : \theta \in \Theta \subseteq \Theta\}$, and we want to estimate $\theta$ using some estimator $\delta(X) \in \mathbb{R}^d$. The *bias* of $\delta$ (under sampling from $P_\theta$) is defined as

$$\text{Bias}_\theta(\delta(X)) = \mathbb{E}_\theta[\delta(X)] - \theta.$$

For $d = 1$, it is well-known that the mean squared error $\text{MSE}(\theta; \delta)$ can be decomposed as the sum of the squared bias of $\delta$ and its variance:

$$\text{MSE}(\theta; \delta) = \text{Bias}_\theta(\delta)^2 + \text{Var}_\theta(\delta). \tag{1}$$

(a) Derive the correct generalization of (1) for general $d \geq 1$, where the MSE is defined as

$$\text{MSE}(\theta; \delta) = \mathbb{E}_\theta \|\delta(X) - \theta\|_2^2.$$

It might help to start with $d = 1$.

(b) Suppose that we are estimating the false positive rate of a new diagnostic test for some disease, using a sample of $n$ specimens taken from a population known not to have the disease we are testing for. If $X$ is the number of false positives and $\theta \in (0, 1)$ is the false positive rate, assume $X \sim \text{Binom}(n, \theta)$. The "obvious" estimator is $\delta_0(X) = X/n$.

However, biological samples are expensive to obtain and the new test is a slightly modified version of an old test whose false positive rate is known to be $\theta_0 \in (0, 1)$, so we might want to "shrink" the estimator toward $\theta_0$ as follows:

$$\delta_\gamma(X) = \gamma\theta_0 + (1 - \gamma)\frac{X}{n}, \quad \text{for } \gamma \in [0, 1],$$

where taking $\gamma = 0$ reduces to the "obvious" estimator $\delta_0(X) = X/n$.

Find the MSE of $\delta_\gamma(X)$ as an explicit expression in $\theta_0, \theta, n$, and $\gamma$.

(c) Find the parameter $\gamma^*$ for which the MSE is minimized, as an expression in $n, \theta$, and $\theta_0$. What happens to $\gamma^*$ if we send $\theta \to \theta_0$ holding $\theta_0$ and $n$ fixed? What if we send $n \to \infty$ holding $\theta$ and $\theta_0$ fixed instead? Explain why these limits make sense.

(d) In our calculation above, $\gamma^*$ is never exactly zero. That is, a smidgeon of shrinkage always beats no shrinkage. Does this prove that $\delta_0$ is inadmissible? Prove or disprove whether $\delta_0$ is dominated by any $\delta_\gamma$.

**Moral:** Shading our estimate toward some "hunch" value can be an effective technique to improve an estimator's performance. This is a central idea in statistics and machine learning that goes by many names: regularization, shrinkage, and inductive bias, to name a few. The optimal amount of bias in an estimator depends on the sample size, and the accuracy of our hunch, but is rarely zero. This may give us pause about insisting that estimators should be unbiased, a theme to which we will return later.

**2. Convexity of $A(\eta)$ and $\Xi_1$**

Let $\mathcal{P} = \{p_\eta : \eta \in \Xi_1\}$ denote an $s$-parameter exponential family in canonical form

$$p_\eta(x) = e^{\eta' T(x) - A(\eta)} h(x), \qquad A(\eta) = \log \int_{\mathcal{X}} e^{\eta' T(x)} h(x) \, \mathrm{d}\mu(x),$$

where $\Xi_1 = \{\eta : A(\eta) < \infty\}$ is the natural parameter space.

Recall Hölder's inequality: if $q_1, q_2 \geq 1$ with $q_1^{-1} + q_2^{-1} = 1$, and $f_1$ and $f_2$ are ($\mu$-measurable) functions from $\mathcal{X}$ to $\mathbb{R}$, then

$$\|f_1 f_2\|_{L^1(\mu)} \leq \|f_1\|_{L^{q_1}(\mu)} \|f_2\|_{L^{q_2}(\mu)}, \quad \text{where } \|f\|_{L^q(\mu)} = \left( \int_{\mathcal{X}} |f(x)|^q \, \mathrm{d}\mu(x) \right)^{1/q}.$$

(**Note** that $q_1 = q_2 = 2$ reduces to Cauchy-Schwarz).

(a) Show that $A(\eta) : \mathbb{R}^s \to [0, \infty]$ is a convex function: that is, for *any* $\eta_1, \eta_2 \in \mathbb{R}^s$ (not just in $\Xi_1$), and $c \in [0, 1]$ then

$$A(c\eta_1 + (1-c)\eta_2) \leq cA(\eta_1) + (1-c)A(\eta_2) \tag{2}$$

(**Hint**: try $q_1 = c^{-1}$, $f_1(x)^{1/c} = e^{\eta_1' T(x)} h(x)$.)

(b) Conclude that $\Xi_1 \subseteq \mathbb{R}^s$ is convex.

**Moral:** The natural parameter space for any exponential family (meaning the set of all parameters $\eta$ that give normalizable densities) is a convex subset of $\mathbb{R}^s$.

**3. Expectation of an increasing function**

(a) Assume $X \sim P$ is a real-valued random variable. Show that if $f(x)$ and $g(x)$ are non-decreasing functions of $x$, then

$$\mathrm{Cov}(f(X), g(X)) \geq 0$$

(**Hint**: derive the identity $\mathbb{E}\left[(f(X_1) - f(X_2))(g(X_1) - g(X_2))\right] = 2\mathrm{Cov}(f(X_1), g(X_1))$, where $X_1, X_2 \overset{\text{i.i.d.}}{\sim} P$).

(b) Let $p_\eta(x)$ be a one-parameter canonical exponential family with non-decreasing sufficient statistic $T(x)$, where $x \in \mathcal{X} \subseteq \mathbb{R}$:

$$p_\eta(x) = e^{\eta T(x) - A(\eta)} h(x).$$

Let $\psi(x)$ be any non-decreasing bounded function. Show that, for $\eta \in \Xi_1^\circ$, $\frac{d}{d\eta} \mathbb{E}_\eta[\psi(X)] \geq 0$.

(**Hint**: find an expression for $\frac{d}{d\eta} \mathbb{E}_\eta[\psi(X)]$ by using methods akin to the ones we used in class to derive the differential identities. You may appeal to Keener Theorem 2.4 to justify differentiating under the integral sign.)

(c) Conclude that $X$ is stochastically increasing in $\eta$; that is, show $\mathbb{P}_\eta(X \leq c)$ is non-increasing in $\eta$, for every $c \in \mathbb{R}$.

**Moral:** This exercise confirms something that we should intuitively expect to be true: that increasing the natural parameter $\eta$, which "tilts" the distribution toward larger values of $T(X)$, will also shift the distribution of $X$ to the right if $T$ is an increasing function. It also illustrates the usefulness of differential identities for understanding exponential families' structure.

**4. Exponential families maximize entropy**

The entropy (with respect to $\mu$) of a random variable $X$ with density $p$, is defined by

$$h(p) = \mathbb{E}_p(-\log p(X)) = -\int_{\{x:\, p(x) > 0\}} \log(p(x)) p(x) \, \mathrm{d}\mu(x).$$

Here, as always in this course, $\log$ denotes the natural logarithm, but $h$ is also commonly defined in terms of the log with base 2. Entropy arises naturally in information theory as a minimal expected code length (for the base-2 log), or in statistical mechanics as a measure of the disorder in a physical system.

Let $T : \mathcal{X} \to \mathbb{R}^s$ denote a generic function, and let $\alpha$ be some vector in the interior of the convex hull of $T(\mathcal{X}) = \{T(x) : x \in \mathcal{X}\}$. Consider the problem of maximizing $h(p)$ over all probability densities subject to the constraint that $\mathbb{E}_p[T(X)] = \alpha$. That is, we want to solve

$$\begin{aligned} \text{maximize} \quad & -\int_{\{x:\, p(x)>0\}} \log(p(x))p(x)\,\mathrm{d}\mu(x) \\ \text{s.t.} \quad & p(x) \geq 0, \ \int_{\mathcal{X}} p(x)\,\mathrm{d}\mu(x) = 1, \text{ and } \int_{\mathcal{X}} p(x)T(x)\,\mathrm{d}\mu(x) = \alpha \in \mathbb{R}^s. \end{aligned}$$

(a) If $\mathcal{X}$ is a finite set with $\mu(\{x\}) > 0$ for all $x \in \mathcal{X}$, show that the optimal $p^*$ is a member the $s$-parameter exponential family

$$p_\eta(x) = e^{\eta' T(x) - A(\eta)},$$

with parameter $\eta^* \in \mathbb{R}^s$ chosen so that $p_{\eta^*}$ satisfies the constraints.
(**Hint**: use Lagrange multipliers).

(b) Blithely[1] applying the result of (a) to $\mathcal{X} = \mathbb{R}$, find the distribution that maximizes entropy with respect to the Lebesgue measure, subject to the constraint that $\mathbb{E}(X) = \mu, \mathrm{Var}(X) = \sigma^2$.

(c) Assume that we need to place $n$ balls into $d$ bins. The number of ways to place the balls resulting in $k_i$ total balls in bin $i$, for $i = 1, \ldots, d$, is given by the combinatorial expression $\frac{n!}{k_1!k_2!\cdots k_d!}$.
Now consider the empirical distribution of the balls. Its probability mass function is $p(i) = k_i/n$ with respect to the counting measure on $\{1, \ldots, d\}$. Let $N_p$ denote the number of configurations with empirical distribution $p$, and show that

$$\log(N_p) = nh(p) + O(\log n),$$

where $h(p)$ is the entropy with respect to the counting measure on $\{1, \ldots, d\}$.
In other words, there are many more high-entropy configurations than low-entropy configurations. This suggests the intuition that, if we consider a physical system at a "macro level" (such as the distribution of gas particles in a container) then we should expect it to drift toward high-entropy configurations.
**Hint:** It may be helpful to recall Stirling's approximation:

$$\log(n!) = n\log n - n + O(\log n)$$

**Moral:** This exercise illustrates additional reasons why exponential family distributions are natural objects of study in statistics.

## 5. 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

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

[1] Meaning naively, without any concern that anything new might go wrong in a continuous space

The gamma distribution generalizes the exponential distribution

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

and the chi-squared distribution

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

(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.

# Stats 210A, Fall 2023
# Homework 2

**Due date**: Wednesday, Sep. 13

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).

If you need to write code to answer a question, show your code. If you need to include a plot, make sure the plot is readable, with appropriate axis labels and a legend if necessary. Points will be deducted for very hard-to-read code or plots.

**1. Minimal sufficiency of the likelihood ratio**

Suppose that $\mathcal{P} = \{p_\theta : \theta \in \Theta\}$ is a family of densities defined with respect to a common measure $\mu$ on $\mathcal{X}$. Assume for simplicity that $p_\theta(x) > 0$ for all $\theta \in \Theta$ and $x \in \mathcal{X}$.

For $\theta_1, \theta_2 \in \Theta$, define the likelihood ratio as

$$\text{LR}(\theta_1, \theta_2; X) = \frac{p_{\theta_1}(X)}{p_{\theta_2}(X)} \in (0, \infty).$$

(a) Use the factorization theorem directly to prove that the *likelihood ratio process*

$$R(X) = (\text{LR}(\theta_1, \theta_2; X) : \theta_1, \theta_2 \in \Theta)$$

is minimal sufficient.

The statistic $R(X)$ should be understood as a stochastic process, i.e. a collection of real random variables $R_{\theta_1,\theta_2}(X) = \text{LR}(\theta_1, \theta_2; X)$, indexed by $(\theta_1, \theta_2) \in \Theta^2$.

**Hint:** Don't forget to prove that $R(X)$ is sufficient.

**Hint:** If you find the concept of a stochastic process over a generic index set perplexing and unintuitive, I suggest you warm up by working through the problem assuming that $\Theta = \{1, 2, \ldots, d\}$ for some finite integer $d$. Then $R$ is simply a $d \times d$ random matrix with $R_{i,j} = \text{LR}(i, j; X)$.

**Note:** You could trivialize this problem by starting from the essentially equivalent result from class about the "likelihood shape." I *don't* want you to use the likelihood shape because the point of this exercise is for you to work out a more concrete version of what is essentially the same result.

(b) Show by counterexample that the *likelihood function*, defined as

$$\text{Lik}(\theta; X) = (p_\theta(X))_{\theta \in \Theta}$$

is *not*, in general, minimal sufficient.

**Note:** If you try to construct a counterexample by playing dirty tricks with measure-zero sets, it probably won't be a real counterexample for the rigorous measure-theoretic definition of minimal sufficient, the rigorous statement of the factorization theorem, and so on. These kind of shenanigans should not be necessary; once you have understood the essence of the problem it will not be hard to come up with a counterexample for discrete $\mathcal{X}$.

(c) **Optional** (not graded, no extra points). If we want to be more concrete we can define the "likelihood shape" concretely as the equivalency class of all functions on $\Theta$ that are proportional to Lik:

$$S(X) = (0, \infty) \cdot \text{Lik}(\cdot; X) = \{c \cdot \text{Lik}(\cdot; x) : c \in (0, \infty)\}$$

Show that the likelihood shape $S(X)$ is minimal sufficient by appealing to your result from part (a).

**Moral:** The collection of likelihood ratios is minimal sufficient, as is the likelihood shape. However, the likelihood function is not minimal sufficient because the scaling constant might be irrelevant for estimating $\theta$.

**2. Bayesian interpretation of sufficiency**

Assume we have a family $\mathcal{P}$ of densities $p_\theta(x)$ with respect to a common measure $\mu$ on $\mathcal{X}$, for $\theta \in \Theta \subseteq \mathbb{R}^n$. Additionally, assume the parameter $\theta$ is itself random, following *prior density* $q(\theta)$ with respect to the Lebesgue measure on $\Theta$.

Then, we can write the *posterior density* (distribution of $\theta$ given $X = x$) as

$$q_{\text{post}}(\theta \mid x) = \frac{p_\theta(x)q(\theta)}{\int_\Theta p_\zeta(x)q(\zeta)\,\mathrm{d}\zeta}.$$

(**Note:** this manipulation of the densities generally works even though we might worry about conditioning on a measure zero set. Feel free to make similar manipulations yourself in the problem).

In this setting, prove the following claims:

(a) Suppose a statistic $T(X)$ has the property that, for any prior distribution $q(\theta)$, the posterior distribution $q_{\text{post}}(\theta \mid x)$ depends on $x$ only through $T(x)$. Show that $T(X)$ is sufficient for $\mathcal{P}$.

(b) Conversely, assume that if $T(X)$ is sufficient for $\mathcal{P}$ and show that, for any prior $q$, the posterior depends on $x$ only through $T(x)$.

**Moral:** If we have a prior opinion about $\theta$ in the form of a distribution, and then we rationally update our opinion after observing $X$, then we will naturally adhere to the sufficiency principle. This gives an alternative epistemological motivation for the principle.

**3. Mean parameterization of an exponential family**

Consider the $s$-parameter exponential family $\mathcal{P} = \{P_\eta : \eta \in \Xi\}$ on $\mathcal{X}$ with densities $p_\eta(x) = e^{\eta' T(x) - A(\eta)}h(x)$ with respect to a common dominating measure $\nu$. Assume $\Xi = \Xi_1^\circ$, the interior of the full natural parameter space, and that $\mathrm{Var}_\eta(a'T(X)) > 0$ for all $a \neq 0$ and $\eta \in \Xi$.

Define the *mean parameter*

$$\mu(\eta) = \mathbb{E}_\eta[T(X)].$$

We will show that this is a one-to-one mapping, so $\mathcal{P}$ can be alternatively be parameterized by $\mu(\eta)$ instead of $\eta$. The Bernoulli, Poisson, and exponential distributions are exponential families that are most often parameterized by their means, and parameterizations of other distributions like the normal and binomial are closely related to the mean parameterization.

Throughout this problem, you may use without proof that if the variance of any statistic $S(X)$ is positive under one $P_\eta \in \mathcal{P}$ then it is positive under all $P_\eta \in \mathcal{P}$ (as an optional exercise, try to prove this).

(a) For $s = 1$, show that $\eta \mapsto \mathbb{E}_\eta[T(X)]$ is a one-to-one mapping; that is, show that if $\eta_1 \neq \eta_2$ then $\mathbb{E}_{\eta_1}[T(X)] \neq \mathbb{E}_{\eta_2}[T(X)]$.
**Hint:** You can use the differential identities.

(b) For $s > 1$ and $\eta_1, \eta_2 \in \Xi$, consider the subfamily whose parameter space is the line segment between $\eta_1$ and $\eta_2$. For $\theta \in [0, 1]$, let

$$\eta(\theta) = (1 - \theta)\eta_1 + \theta\eta_2.$$

Show that this subfamily is a one-parameter exponential family on $\mathcal{X}$ with natural parameter $\theta$, and write it in standard exponential family form.

(c) Combine (a) and (b) to show that $\eta \mapsto \mathbb{E}_\eta[T(X)]$ is a one-to-one mapping for $s \geq 1$.

**4. Multinomial family**

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 $\pi$ 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 *not* i.i.d. samples; each one corresponds to a different bin and $X_1$ is not independent of $X_2$.

(a) Rewrite the densities as a $(d-1)$-parameter exponential family, giving an explicit form for $T(x)$, $h(x)$, $\eta$, and $A(\eta)$. Is $X = (X_1, \ldots, X_d)$ minimal sufficient?

(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 $\theta^2$, $2\theta(1-\theta)$, and $(1-\theta)^2$, respectively.

We can estimate $\theta$ 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 $\theta$:

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

Can $\mathcal{P}$ be written as a one-parameter exponential family? Find a minimal sufficient statistic for $\mathcal{P}$.

**5. Uniform location-scale family**

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Unif}[\mu - \sigma, \mu + \sigma]$, with $\mu \in \mathbb{R}$ and $\sigma > 0$ unknown.

(a) Show that $T(X) = (X_{(1)}, X_{(n)})$ is minimal sufficient.

(b) If $B \sim \text{Beta}(\alpha, \beta)$ then its density is proportional to $x^{\alpha-1}(1-x)^{\beta-1}$ on $x \in [0, 1]$.

If $U_1, \ldots, U_n \overset{\text{i.i.d.}}{\sim} U[0, 1]$, show that

$$U_{(n)} \sim \text{Beta}(n, 1), \quad \text{which has density } p(x) = nx^{n-1},$$

and

$$U_{(1)}/U_{(n)} \sim \text{Beta}(1, n-1) \quad \text{which has density } p(x) = (n-1)(1-x)^{n-2},$$

independently of $U_{(n)}$.

**Hint:** For the first part, start by writing down the CDF of $U_{(n)}$.

**Hint:** For the second part, you may use without proof the fact that, conditional on $U_{(n)} = u$, the remaining $n-1$ values are i.i.d. Unif$[0, u]$, then proceed similarly to what you did for the first part.

(c) Suppose that we wish to estimate $\mu$ under the squared error loss. The sample mean $\overline{X}$ may appear to be a reasonable estimator of $\mu$, but we might worry about the fact that it is not a function of $T(X)$.

Guided by the sufficiency principle, we could instead consider the estimator

$$\delta(X) = \frac{X_{(1)} + X_{(n)}}{2}.$$

Compute the MSE of each estimator as a function of $n, \mu$, and $\sigma$, and show that $\delta$ strictly dominates $\overline{X}$ for $n > 2$ (the estimators coincide for $n = 2$). What happens to the ratio of their MSE's as $n \to \infty$?

**Hint:** The results from part (b) should be useful. You may use without proof that Beta$(\alpha, \beta)$ has mean $\frac{\alpha}{\alpha+\beta}$ and variance $\frac{\alpha\beta}{(\alpha+\beta)^2(\alpha+\beta+1)}$.

(d) Simulate the distribution for $\mu = 0, \sigma = 1, n = 1000$. For each estimator, plot a histogram of simulated estimates.

**Moral:** Understanding and respecting the statistical structure of a model sometimes helps us to come up with estimators that perform dramatically better than the estimator we would have naïvely thought of. Here is a case where applying the sufficiency principle helped us get a much better estimator than the sample mean.

# 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 $P_\theta$. 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 $\theta$ (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 $\Theta$ has an accumulation point at 0, i.e. $\Theta$ 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 $P_\theta^T$ "spans" a certain vector space, so that only the zero function is orthogonal to all $P_\theta^T$.

**Moral 2:** If $\mathcal{P} = \{P_\eta : \ \eta \in \Xi\}$ is a full-rank exponential family with natural parameter $\eta$, meaning $\Xi$ 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 $\Xi$ 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* $\zeta$ if its distribution is independent of $\zeta$; that is, if $T(X)$ is ancillary in the subfamily where $\lambda$ is known, for each possible value of $\lambda$.

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 $\mu$ 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 $\mu$ (we can call this the *naive* estimator since it ignores the selection bias).

(b) Show that $X_2$ is unbiased for $\mu$, 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 $\mu$ 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 $\mu$. 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

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

**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.

# Stats 210A, Fall 2023
# Homework 4

**Due date**: Wednesday, Sep. 27

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. Bayesian law of large numbers**

(a) Let $p(x)$ and $q(x)$ denote two strictly positive probability densities with respect to a common dominating measure $\mu$. The *Kullback–Leibler divergence* between $p$ and $q$ is defined as

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

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 $\Theta$ is a finite set, and the densities are strictly positive on $\mathcal{X}$. Let $\lambda$ denote a prior density w.r.t. the counting measure on $\Theta$, 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 $\theta_0$ (that is, we are doing a *frequentist* analysis 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 $\lambda$ puts positive mass on all values in $\Theta$, show that as $n \to \infty$, the posterior density eventually concentrates nearly all its mass on the true value $\theta_0$. 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:** use the law of large numbers).

**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."

**2. Fisher information for location and scale families**

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.

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

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

Try to explain in your own words why it makes sense that the Fisher information should be proportional to $\theta^{-2}$ (the verbal explanation will be graded leniently).

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

**3. Ridge regression**

Consider the *Gaussian linear model* where

$$y_i = x_i' \beta + \varepsilon_i, \quad \text{with } \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2) \text{ for } i = 1, \dots n,$$

where $\beta \in \mathbb{R}^d$ is unknown, and the covariate vectors $x_i \in \mathbb{R}^d$ are fixed and known. Assume the error variance $\sigma^2 > 0$ is also known. We observe the response vector $y \in \mathbb{R}^n$.

(a) Assume that $d \leq n$, and the design matrix $\mathbf{X}$ (the $n \times d$ matrix whose $i$th row is $x_i'$) has full column rank. Show that the OLS estimator $\hat{\beta} = (\mathbf{X}'\mathbf{X})^{-1}\mathbf{X}'y$ is the UMVU estimator of $\beta$.

**Note:** Remember that the design matrix $\mathbf{X}$ is not data in the same sense $y$ is; it is more like a known parameter.

(b) Now consider Bayesian estimation with the prior $\beta \sim N(\mu, \tau^2 I_d)$. Under the same prior as in part (b), find the posterior distribution of $\beta$. Does it matter whether $d > n$, or whether $\mathbf{X}$ has full column rank?

(c) Suppose that $\mathbf{X}\gamma = 0$ for some nonzero $\gamma \in \mathbb{R}^d$. Show that no unbiased estimator exists for $g(\beta) = \beta'\gamma$. What is the posterior distribution for $g(\beta)$?

**4. Other loss functions**

Assume for each problem below that there exists an estimator with finite Bayes risk.

(a) Consider a Bayesian model with a discrete parameter $\theta$. What is the Bayes estimator for the loss $L(\theta, d) = 1\{\theta \neq d\}$?

(b) Next consider a Bayesian model with a single real parameter $\theta$, and assume that the posterior distribution of $\theta$ given $X = x$ is absolutely continuous (with respect to the Lebesgue measure) for all $x$. What is the Bayes estimator for the *absolute error loss* $L(\theta, d) = |\theta - d|$?

(c) Under the same assumptions as part (b), what loss function $L_\gamma(\theta, d)$ would give the posterior $\gamma$ quantile as its Bayes estimator; that is, the estimator $\delta_\gamma(X)$ has $\mathbb{P}(\theta < \delta_\gamma(X) \mid X) = \gamma$.

**5. Exponential-exponential model**

Consider a Bayesian model with prior distribution $\lambda(\theta) = e^{-\theta} 1\{\theta > 0\}$ for $\theta$ (the standard exponential distribution) and whose likelihood is the exponential location family:

$$p_\theta(x) = e^{\theta - x} 1\{x > \theta\},$$

where we observe a sample $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} p_\theta(x)$ given $\theta$.

(a) Calculate the posterior distribution for $\theta$ for $n > 1$.

(b) For $n = 1$, calculate the posterior distribution and the Bayes estimator under squared error loss.

(c) Still for $n = 1$, calculate the MSE for the Bayes estimator and the UMVU estimator as a function of $\theta$. Plot the risk function for $\theta \in [0, 5]$. For what values of $\theta$ does the Bayes estimator perform better?

(d) Still for $n = 1$, calculate the Bayes risk for the Bayes estimator, and for the UMVU estimator $X_1 - 1$, using squared error loss.

**Moral:** The Bayes estimator tends to have better risk in places where the prior is large, sometimes at the cost of performing very poorly where the prior puts very little mass.

# Stats 210A, Fall 2023
# Homework 5

**Due date**: Wednesday, Oct. 4

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).

If you need to write code to answer a question, show your code. If you need to include a plot, make sure the plot is readable, with appropriate axis labels and a legend if necessary. Points will be deducted for very hard-to-read code or plots.

**1. Admissibility and Bayes estimators**

One of the frequentist motivations for Bayes estimators is their connection to admissibility.

(a) Suppose that the Bayes estimator $\delta_\Lambda$ for the prior $\Lambda$ is unique up to $\mathcal{P}$-almost-sure equality. That is, for any other Bayes estimator $\tilde{\delta}_\Lambda$, we have $\delta_\Lambda(X) = \tilde{\delta}_\Lambda(X)$ almost surely, for every $P_\theta \in \mathcal{P}$. Show that $\delta_\Lambda$ is admissible.

(b) Now suppose that $\Theta$ is discrete (possibly countably infinite) and $\Lambda$ is a probability distribution putting positive mass on every value $\theta \in \Theta$. Show that any Bayes estimator with finite Bayes risk is admissible.

(c) A *randomized estimator* is an estimator that is a random function of the data. We can formalize it generically as $\delta(X, W)$ where $X \sim P_\theta$ as usual and $W$ is some auxiliary random variable generated by the analyst. For this part, "admissible" and "Bayes" are defined with respect to all estimators including randomized ones.

Now consider a model with a finite parameter space, $|\Theta| = n < \infty$ and assume we are estimating some real-valued $g(\theta)$ using a bounded non-negative loss $L : \Theta \times \mathbb{R} \to [0, \infty)$. Show that every admissible estimator is a (possibly randomized) Bayes estimator for some prior.

**Hint:** consider the set $\mathcal{A}$ of all achievable risk functions, and the set $\mathcal{D}_\delta$ of all (possibly unachievable) risk functions that would dominate a given estimator $\delta$. Recall the *hyperplane separation theorem*: for any two disjoint non-empty convex subsets $A, B \subseteq \mathbb{R}^n$ there exist $c \in \mathbb{R}$ and nonzero $\lambda \in \mathbb{R}^n$ such that $\lambda' a \geq c \geq \lambda' b$ for all $a \in A, b \in B$. It might help to draw pictures for $n = 2$.

**Moral:** Minimizing average-case risk is closely related to admissibility.

**2. MCMC algorithms**

This problem considers MCMC sampling from a generic posterior density $\lambda(\theta \mid x)$ where $\theta \in \mathbb{R}^d$.

(a) The Metropolis–Hastings algorithm is a Markov chain using the following update rule: First, sample $\zeta \sim f(\cdot \mid \theta^{(t)})$ according to some "proposal distribution" $f(\zeta \mid \theta) : \Theta \times \Theta \to (0, \infty)$,

where $f(\cdot \mid \theta)$ is a probability density for each $\theta$ (assume $\lambda$ and $f(\cdot \mid \theta)$ are densities w.r.t. the same dominating measure). Next, compute the "accept probability" as

$$a(\zeta \mid \theta) = \min\left\{1, \ \frac{\lambda(\zeta \mid X)}{\lambda(\theta \mid X)} \frac{f(\theta \mid \zeta)}{f(\zeta \mid \theta)}\right\}.$$

Finally, let $\theta^{(t+1)} = \zeta$ with probability $a(\zeta \mid \theta^{(t)})$ and $\theta^{(t+1)} = \theta^{(t)}$ with probability $1 - a(\zeta \mid \theta^{(t)})$. Show that $\lambda(\theta \mid X)$ is stationary for the Metropolis–Hastings algorithm.

(b) Consider the version of the Gibbs sampler that updates a *single* random index $J^{(t+1)} \sim \text{Unif}\{1, \ldots, d\}$ at each step, so

$$\theta_j^{(t+1)} = \begin{cases} \zeta_j^{(t+1)} & \text{if } j = J^{(t+1)} \\ \theta_j^{(t)} & \text{if } j \neq J^{(t+1)} \end{cases},$$

with

$$\zeta_j^{(t+1)} \mid \theta^{(t)} \sim \lambda(\theta_j \mid \theta_{\setminus j} = \theta^{(t)}, X),$$

where $\lambda$ above is the conditional density for the $j$th coordinate of $\theta$ given the others, and the data, in the full Bayes model. Show that this algorithm is a special case of the Metropolis–Hastings algorithm.

**Note:** The Metropolis-Hastings algorithm is computationally attractive because we can can always implement it using only the unnormalized posterior $p_\theta(X)\lambda(\theta)$ (or any function $g(\theta)$ that is proportional to it), which is often much easier to compute than the normalized posterior.

**3. Empirical Bayes for exponential families**

Consider an $s$-parameter exponential family model in canonical form:

$$p_\theta(x) = e^{\theta' T(x) - A(\theta)} h(x)$$

where $x = (x_1, \ldots, x_n)$. We will consider a Bayes prior for the random vector $\theta$ with density $\lambda_\gamma(\theta)$, which is itself parameterized by a hyperparameter $\gamma \in \Gamma$. We will consider an empirical Bayes model where $\gamma$ is fixed and $\theta$ and $X$ are both random. Let $\lambda_\gamma(\theta \mid x)$ and $q_\gamma(x)$ denote the posterior and marginal, respectively, for a given choice of $\gamma$, and $\mathbb{E}_\gamma$ represent expectations (or conditional expectations) with respect to the joint distribution over $\theta$ and $X$.

Assume both $\Gamma$ and the natural parameter space $\Xi_1$ are open subsets of $\mathbb{R}$ and $\mathbb{R}^s$, respectively. Assume also that all relevant quantities are suitably differentiable and/or integrable, and that derivatives can always be taken inside the integral sign.

(a) Show that for $i = 1, \ldots, n$, we have

$$\mathbb{E}_\gamma\left[\sum_{j=1}^{s} \theta_j \frac{\partial T_j(x)}{\partial x_i} \mid X = x\right] = \frac{\partial}{\partial x_i} \log q_\gamma(x) - \frac{\partial}{\partial x_i} \log h(x),$$

(b) Now assume we have $n = s$ with $T(x) = x$:

$$p_\theta(x) = e^{\theta' x - A(\theta)} h(x).$$

Let $\hat{\gamma}(X)$ denote the maximum likelihood estimator (MLE) of $\gamma$ based on the observed data:

$$\hat{\gamma}(X) = \arg\max_{\gamma \in \Gamma} q_\gamma(X),$$

which we assume always exists and is unique.

Show that the empirical posterior mean of $\theta$, using $\hat{\gamma}$ to estimate $\gamma$, is

$$\mathbb{E}_{\hat{\gamma}}\left[\theta \mid X = x\right] = \nabla_x \left(\log q_{\hat{\gamma}(x)}(x) - \log h(x)\right)$$

**Note:** You should interpret $\mathbb{E}_{\hat{\gamma}}[\cdot]$ as $\mathbb{E}_\gamma[\cdot]\big|_{\gamma=\hat{\gamma}}$, and $q_{\hat{\gamma}(x)}(x)$ as $q_\gamma(x)\big|_{\gamma=\hat{\gamma}(x)}$. Note that the second expression depends on $x$ in two places.

**Moral:** This gives easy-to-evaluate rules for calculating empirical Bayes estimators in simple exponential family models.

**4. Gamma-Poisson empirical Bayes**

Consider the Bayes model with

$$\theta_i \overset{\text{i.i.d.}}{\sim} \text{Gamma}(k, \sigma), \quad i = 1, \ldots, n$$
$$X_{ij} \mid \theta_i \overset{\text{ind.}}{\sim} \text{Pois}(\theta_i), \quad i = 1, \ldots, n, \quad j = 1, \ldots, m$$

Assume $k > 0$ (shape parameter) is known and $\sigma > 0$ (scale parameter) is unknown and estimated via the MLE. In addition, assume $\sum_{ij} X_{ij} > 0$ (though the formulae below would be basically correct in a limiting sense if the sum were zero, too).

(a) If $m = 1$, show that the empirical Bayes posterior mean for $\theta_i$ is

$$\frac{\overline{X}}{\overline{X} + k}(k + X_{i1}), \quad \text{where } \overline{X} = n^{-1} \sum_i X_{i1}.$$

You may use without proof the fact that the marginal distribution of $X_i$ is negative binomial.

(b) For general $m$, show that the empirical Bayes posterior mean for $\theta_i$ is

$$\frac{\overline{X}}{\overline{X} + k/m}(k/m + \overline{X}_i), \quad \text{where } \overline{X}_i = m^{-1} \sum_j X_{ij} \quad \text{and } \overline{X} = (nm)^{-1} \sum_{ij} X_{ij}.$$

**Hint:** Make a sufficiency reduction and remember that $\sigma$ is a scale parameter.

**5. Gibbs Sampler for Gamma-Poisson model**

Consider a hierarchical Bayes model instead, where

$$\sigma^{-1} \sim \text{Exp}(1)$$
$$\theta_i \mid \sigma \overset{\text{i.i.d.}}{\sim} \text{Gamma}(k, \sigma), \quad i = 1, \ldots, n$$
$$X_{ij} \mid \sigma, \theta \overset{\text{ind.}}{\sim} \text{Pois}(\theta_i), \quad i = 1, \ldots, n, \quad j = 1, \ldots, m$$

where $\sigma$ is a scale parameter, and $k, n, m$, are fixed and known.

**Note:** For parts (b) and (c) below, be sure to read the instructions on coding problems at the top of this problem set.

(a) Give an explicit algorithm for one full iteration of the Gibbs sampler. It may be helpful to look up the inverse gamma distribution on Wikipedia.

(b) Implement the Gibbs sampler in a programming language of your choice (R is recommended since it is easy to draw random draws from standard distributions; Python or Matlab will probably also work fine). For $k = m = 3$ and $n = 100$, download the matrix $X \in \mathbb{R}^{n \times m}$, in `hw5.csv` from the course website and implement the Gibbs sampler (the standard version where you update all variables in every round; use 100 rounds of burn-in and take the next 10,000 rounds of sampling, without thinning). Make a trace plot of your draws from $\sigma$ and $\theta_1$ and include them in your homework submission. Report the following three estimators of $\theta_1$, to three significant digits:

  (i) the hierarchical Bayes estimator (for squared error loss),

  (ii) the empirical Bayes estimator from Problem 4, and

  (iii) the UMVU estimator (in the model where $\theta$ is fixed and unknown).

(c) Next, carry out a Monte Carlo simulation to estimate the Bayes risk conditional on $\sigma$, for four estimators: (i–iii) from part (b), plus the "oracle Bayes" estimator where the value of $\sigma$ is known. That is, for each estimator $\delta_1^{(\ell)}(X)$ of $\theta_1$, approximately evaluate:

$$R^{(\ell)}(\sigma) = \mathbb{E}[(\delta_1^{(\ell)}(X) - \theta_1)^2 \mid \sigma] = \mathbb{E}\left[ n^{-1} \sum_i (\delta_i^{(\ell)}(X) - \theta_i)^2 \mid \sigma \right],$$

where the expectation is taken over $\theta$ and $X$ (but *not* $\sigma$, since we are conditioning on that). The second equality follows from the exchangeability over different values of $i$ (you do not need to prove it yourself, but you should use it to save yourself computation). **Note:** for the hierarchical Bayes estimator, this does *not* mean you should hold $\sigma$ fixed in your MCMC chain: you should compute it just as you did in part (b). Use the values $\sigma = 0.1, 0.2, 0.5, 1, 2, 5, 10$ and include a $4 \times 7$ table of risk values, each reported to at least 3 significant figures, in your answer.

For each of the three non-oracle estimators, plot the relative excess risk

$$\frac{R^{(\ell)}(\sigma)}{R^{(\text{oracle})}(\sigma)} - 1$$

against $\sigma$ for the values above. Make an analogous plot for $m = 30, n = 100$ and another for $m = 3, n = 10$. I recommend using a log scale for the horizontal and vertical axis but it is not required.

**Note:** This exercise should not take you an absurd amount of computer time; using 100 MC runs per value of $\sigma$ and the 7 values of $\sigma$ above, takes my three-year-old laptop computer less than three minutes to produce each of the three plots requested above. If it is taking your computer much much longer you are probably doing something very inefficiently.

(d) **Optional:** Why do your three plots look the way they do? What's the moral of the story?

# Stats 210A, Fall 2023
# Homework 6

**Due date**: Wednesday, Oct. 11

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).

If you need to write code to answer a question, show your code. If you need to include a plot, make sure the plot is readable, with appropriate axis labels and a legend if necessary. Points will be deducted for very hard-to-read code or plots.

**1. Effective degrees of freedom**

We can write a standard Gaussian sequence model in the form

$$Y_i = \mu_i + \varepsilon_i, \quad \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2), \quad i = 1, \ldots, n$$

with $\mu \in \mathbb{R}^n$ and $\sigma^2 > 0$ possibly unknown. If we estimate $\mu$ by some estimator $\hat{\mu}(Y)$, we can compute the residual sum of squares (RSS):

$$\mathrm{RSS}(\hat{\mu}, Y) = \|\hat{\mu}(Y) - Y\|^2 = \sum_{i=1}^{n} (\hat{\mu}_i(Y) - Y_i)^2.$$

If we were to observe the same signal with independent noise $Y^* = \mu + \varepsilon^*$, the expected prediction error (EPE) is defined as

$$\mathrm{EPE}(\mu, \hat{\mu}) = \mathbb{E}_\mu \left[\|\hat{\mu}(Y) - Y^*\|^2\right] = \mathbb{E}_\mu \left[\|\hat{\mu}(Y) - \mu\|^2\right] + n\sigma^2.$$

Because $\hat{\mu}$ is typically chosen to make RSS small for the observed data $Y$ (i.e., to fit $Y$ well), the RSS is usually an optimistic estimator of the EPE, especially if $\hat{\mu}$ tends to overfit. To quantify how much $\hat{\mu}$ overfits, we can define the *effective degrees of freedom* (or simply the *degrees of freedom*) of $\hat{\mu}$ as

$$\mathrm{DF}(\mu, \hat{\mu}) = \frac{1}{2\sigma^2} \mathbb{E}\left[\mathrm{EPE} - \mathrm{RSS}\right],$$

which uses optimism as a proxy for overfitting.

For the following questions assume we also have a predictor matrix $X \in \mathbb{R}^{n \times d}$, which is simply a matrix of fixed real numbers. Suppose that $d \leq n$ and $X$ has full column rank.

(a) Show that if $\hat{\mu}$ is differentiable with $\mathbb{E}_\mu \|D\hat{\mu}(Y)\|_F < \infty$ then

$$\sum_{i=1}^{n} \frac{\partial \hat{\mu}_i(Y)}{\partial Y_i}$$

is an unbiased estimator of the DF. (Recall $D\hat{\mu}(Y)$ is the Jacobian matrix from class).

(b) Suppose $\hat{\mu} = X\hat{\beta}$, where $\hat{\beta}$ is the ordinary least squares estimator (i.e., chosen to minimize the RSS). Show that the DF is $d$. (This confirms that DF generalizes the intuitive notion of degrees of freedom as "the number of free variables").

(c) Suppose $\hat{\mu} = X\hat{\beta}$, where $\hat{\beta}$ minimizes the penalized least squares criterion:

$$\hat{\beta} = \arg\min_{\beta} \|Y - X\beta\|_2^2 + \rho\|\beta\|_2^2,$$

for some $\rho \geq 0$. Show that the DF is $\sum_{j=1}^{d} \frac{\lambda_j}{\rho+\lambda_j}$, where $\lambda_1 \geq \cdots \geq \lambda_d > 0$ are the eigenvalues of $X'X$ (counted with multiplicity) (**Hint:** use the singular value decomposition of $X$).

## 2. Soft thresholding

Consider the *soft thresholding operator* with parameter $\lambda \geq 0$, defined as

$$\eta_\lambda(x) = \begin{cases} x - \lambda & x > \lambda \\ 0 & |x| \leq \lambda \\ x + \lambda & x < -\lambda \end{cases}$$

Note that, although we didn't prove it in class, Stein's lemma applies for continuous functions $h(x)$ which are differentiable except on a measure zero set; you can apply it here without worrying.

Assume $X \sim N_d(\theta, I_d)$ for $\theta \in \mathbb{R}^d$, which we will estimate via $\delta_\lambda(X) = (\eta_\lambda(X_1), \ldots, \eta_\lambda(X_d))$. Soft thresholding is sometimes used when we expect *sparsity*: a small number of relatively large $\theta_i$ values. $\lambda$ here is called a *tuning parameter* since it determines what version of the estimator we use, but doesn't have an obvious statistical interpretation.

(a) Show that $|\{i : |X_i| > \lambda\}|$ is an unbiased estimator of the degrees of freedom of $\delta_\lambda$ (so, in a sense, the DF is the expected number of "free variables").

(b) Show that

$$d + \sum_i \min(X_i^2, \lambda^2) - 2\,|\{i : |X_i| \leq \lambda\}|$$

is an unbiased estimator for the MSE of $\delta_\lambda$.

(c) Show that the risk-minimizing value $\lambda^*$ solves

$$\lambda \sum_i \mathbb{P}_{\theta_i}(|X_i| > \lambda) = \sum_i \phi(\lambda - \theta_i) + \phi(\lambda + \theta_i),$$

where $\phi(z) = \frac{e^{-z^2/2}}{\sqrt{2\pi}}$ is the standard normal density.

(d) Consider a problem with $\theta_1 = \cdots = \theta_{20} = 10$ and $\theta_{21} = \cdots = \theta_{500} = 0$. Compute $\lambda^*$ numerically. Then simulate a vector $X$ from the model and use it to automatically tune the value of $\lambda$ by minimizing SURE. Call the automatically tuned value $\hat{\lambda}(X)$ and report both $\lambda^*$ and $\hat{\lambda}(X)$. Finally plot the true MSE of $\delta_\lambda$ along with its SURE estimate against $\lambda$ for a reasonable range of $\lambda$ values. Add a horizontal line for the risk of the UMVU estimator.

(e) Compute and report the squared error loss $\|\delta(X) - \theta\|^2$ for the following four estimators:

(i) the UMVU estimator $\delta_0(X) = X$,

(ii) the optimally tuned soft-thresholding estimator $\delta_{\lambda^*}(X)$,

(iii) the automatically tuned soft-thresholding estimator $\delta_{\hat{\lambda}(X)}(X)$, and

(iv) the James-Stein estimator.

You do not need to compute the MSE. Intuitively, what do you think accounts for the good performance of soft-thresholding in this example?

## 3. Mean estimation

(a) Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N_d(\theta, I_d)$ and consider estimating $\theta \in \mathbb{R}^d$. Show that $\overline{X} = \frac{1}{n}\sum_i X_i$ is the minimax estimator of $\theta$ under squared error loss.

**Hint:** Find a least favorable sequence of priors.

(b) Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$ where $P$ is any distribution over the real numbers such that $\text{Var}_P(X) \leq 1$. Show that $\overline{X} = \frac{1}{n}\sum_i X_i$ is minimax for estimating $\theta(P) = \mathbb{E}_P X$ under the squared error loss.

**Hint:** Try to relate this problem to the Gaussian problem with $d = 1$.

(c) Assume $X \sim N(\theta, 1)$ with the constraint that $|\theta| \leq 1$. Show that the minimax estimator for squared error loss is

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$

Plot its risk function.

**Hint:** Plot the risk function first. For this problem if you need to show that a function is maximized or minimized somewhere, you may do it numerically or by inspecting a graph if it is obvious enough.

## 4. James-Stein estimator with regression-based shrinkage

Consider estimating $\theta \in \mathbb{R}^d$ in the model $Y \sim N_d(\theta, I_d)$. In the standard James-Stein estimator, we shrink all the estimates toward zero, but it might make more sense to shrink them towards the average value $\overline{Y}$, or towards some other value based on observed side information.

(a) Consider the estimator

$$\delta_i^{(1)}(Y) = \overline{Y} + \left(1 - \frac{d-3}{\|Y - \overline{Y}1_d\|^2}\right)\left(Y_i - \overline{Y}\right)$$

Show that $\delta^{(1)}(Y)$ strictly dominates the estimator $\delta^{(0)}(Y) = Y$, for $d \geq 4$.

$$\text{MSE}(\theta; \delta^{(1)}) < \text{MSE}(\theta; \delta^{(0)}), \quad \text{for all } \theta \in \mathbb{R}^d.$$

Calculate the MSE of $\delta^{(1)}$ if $\theta_1 = \theta_2 = \cdots = \theta_d$. How would it compare to the MSE for the usual James-Stein estimator?

**Hint:** Change the basis using an appropriate orthogonal rotation and think about how the estimator operates on different subspaces.

**Hint:** Recall that if $Z \sim N_d(\mu, \Sigma)$ and $A \in \mathbb{R}^{k \times d}$ is a fixed matrix then $AZ \sim N_k(A\mu, A\Sigma A')$.

(b) Now suppose instead that we have side information about each $\theta_i$, represented by fixed covariate vectors $x_1, \ldots, x_d \in \mathbb{R}^k$. Assume the design matrix $X \in \mathbb{R}^{d \times k}$ whose $i$th row is $x_i'$ has full column rank. Suppose that we expect $\theta \approx X\beta$ for some $\beta \in \mathbb{R}^k$, but unlike the usual linear regression setup, we will not assume $\theta = X\beta$ with perfect equality.

Find an estimator $\delta^{(2)}$, analogous to the one in part (a), that dominates $\delta^{(0)}$ whenever $d - k \geq 3$:

$$\text{MSE}(\theta; \delta^{(2)}) < \text{MSE}(\theta; \delta^{(0)}), \quad \text{for all } \theta \in \mathbb{R}^d,$$

and for which $\text{MSE}(X\beta; \delta^{(2)}) = k + 2$, for any $\beta \in \mathbb{R}^k$.

**Hint:** Think of this setting as a generalization of part (a), which can be considered a special case with $d = 1$ and all $x_i = 1$. What is the right orthogonal rotation?

**Note:** Don't assume there is an additional intercept term for the regression; this could always be incorporated into the $X$ matrix by taking $x_{i,1} = 1$ for all $i = 1, \ldots, d$.

## 5. Upper-bounding $\theta$

(a) Let $X \sim N(\theta, 1)$ for $\theta \in \mathbb{R}$, and consider the loss function

$$L(\theta, d) = 1\{d < \theta\};$$

that is, we observe $X$ and try to come up with an upper bound $\delta(x) \in \mathbb{R}$ for $\theta$. Show that the minimax risk is 0 (note you may not be able to find a minimax estimator).

(b) Now, consider a problem with the same loss function but without observing any data. Show the minimax risk (considering both randomized and non-randomized estimators) is 1, but the Bayes risk $r_\Lambda = 0$ for any prior $\Lambda$ (note there may be no estimator $\delta_\Lambda$ that attains the minimum Bayes risk).

(**Note:** This problem exhibits a "duality gap" where the lower bounds we can get by trying different priors will always fall short of the minimax risk.)

(c) **Optional** (not graded, no extra points): Now consider the same loss function, but now $X \sim N(\theta, \sigma^2)$ and $\sigma^2$ is unknown too. Find the minimax risk.

**Hint:** consider estimators of the form $\delta(X) = c|X|$.

# Stats 210A, Fall 2023
# Homework 7

**Due date**:Thursday, Oct. 19

**Instructions:** The same standing instructions are in effect as in previous weeks.

**1. MLR and location families**

(a) Assume $X \sim p_\theta(x) = p_0(x-\theta)$, a location family with $p_0$ continuous and strictly positive. Show that the family has MLR in $x$ if and only if $\log p_0$ is concave.

**Note:** For full credit, you should not assume that $p_0$ is differentiable.

**Hint 1:** It may help to recall that $f(x)$ is convex if and only if

$$R(x_1, x_2) = \frac{f(x_1) - f(x_2)}{x_1 - x_2}$$

is non-decreasing in $x_1$ and $x_2$.

**Hint 2:** It may also help to recall that a continuous function $f$ is convex if and only if it is *midpoint convex* meaning

$$f\left(\frac{x_1+x_2}{2}\right) \leq \frac{f(x_1)+f(x_2)}{2}, \quad \text{for all } x_1, x_2.$$

(b) Consider testing in the Cauchy location family:

$$p_\theta(x) = \frac{1}{\pi(1+(x-\theta)^2)}.$$

Let $\theta_0, \theta_1$ be any two real numbers with $\theta_1 > \theta_0$ and consider the LRT for testing $H_0: \ \theta = \theta_0$ vs $H_1: \ \theta = \theta_1$ at some level $\alpha \in (0,1)$. Show that for some $\alpha^*(\theta_0, \theta_1)$, the rejection region for any $\alpha < \alpha^*$ is a bounded interval, and the rejection region for any $\alpha > \alpha^*$ is a union of two half intervals. Find $\alpha^*$.

**Hint:** recall that $\frac{d}{dx}\arctan(x) = \frac{1}{1+x^2}$.

(c) In the Cauchy location family, prove that, for any $\alpha \in (0,1)$, there exists no UMP level-$\alpha$ test of $H_0: \ \theta = 0$ vs. $H_1: \ \theta > 0$.

(d) Consider testing $H_0: \theta = 0$ vs. $H_1: \ \theta = 6$ in the Cauchy location family at level $\alpha = 0.01$. Numerically calculate the rejection region and the power for the LRT, and also for the one-tailed test that rejects for large values of $X$.

(e) **Optional:** (not graded, no extra points) In words, can you explain why the optimal LRT rejection regions for the Cauchy distribution take this odd form? Think about how you would explain to a scientific collaborator why you are proposing such an odd test, beyond "it fell out of an optimization problem."

**Moral:** When we think carefully about how to design rejection regions, we can get surprising results. In particular, for location families with heavy tails, extreme values are not that informative for distinguishing between two smaller values of the location parameter. Concretely, $X = 10^6$ doesn't help us distinguish between $\theta_1 = 1$ vs. $\theta_0 = 0$. By contrast, if the tails are lighter ($\log p_0$ concave implies the density shrinks at least exponentially) then more extreme $X$ values always give stronger evidence for distinguishing between any two parameter values; this is what MLR means.

**2. Some UMP tests**

Numerically find the UMP test for the following hypothesis testing problems at level $\alpha = 0.05$. For each problem,

(i) derive the appropriate test on paper,

(ii) numerically compute the cutoff value $c$ (and $\gamma$ if necessary), and

(iii) plot the power function of the level-$\alpha$ test for an appropriate range of parameter values.

(a) $X_i \overset{\text{ind.}}{\sim} \text{Pois}(a_i\lambda)$ for $i = 1, \ldots, n$, where $a_1, \ldots, a_n$ are known positive constants and $\lambda > 0$ is unknown. Test $H_0 : \lambda = 1$ vs. $H_1 : \lambda > 1$, with $n = 5$ and $a_i = i$.

(b) $X_i \overset{\text{ind.}}{\sim} N(\theta, \sigma_i^2)$ for $i = 1, \ldots, n$, where $\sigma_i^2$ are known positive constants and $\theta \in \mathbb{R}$ is unknown. Test $H_0 : \theta = 0$ vs. $H_1 : \theta > 0$, with $n = 20$ and $\sigma_i^2 = i$. On your power plot, also plot the power function of the (sub-optimal) test that rejects for large $\sum_i X_i$.

(c) $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Pareto}(\theta) = \theta x^{-(1+\theta)}$, for $\theta > 0$ and $x > 1$ (also called a power law distribution). Test $H_0 : \theta = 1$ vs. $H_1 : \theta < 1$, for $n = 100$. On your power plot, also plot the power function of the (sub-optimal) test that rejects for large $\sum_i X_i$.

**Moral:** Once again, when we use the right test we often can deliver noticeably better power than if we chose an *ad hoc* test.

**3. Uniform UMP test**

We usually can't get a UMP two-sided test, but this problem gives an amusing counterexample where it is possible. Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Unif}[0, \theta]$ for $\theta > 0$.

(a) Consider the problem of testing $H_0 : \theta = \theta_0$ versus $H_1 : \theta > \theta_0$. Show that any test $\phi$ for which $\phi(x) = 1$ when $x_{(n)} = \max\{x_1, \ldots, x_n\} > \theta_0$ is UMP at level $\alpha = \mathbb{E}_{\theta_0}[\phi(X)]$.

(b) Now consider the problem of testing $H_0 : \theta = \theta_0$ against $H_1 : \theta \neq \theta_0$. Show that a unique UMP level-$\alpha$ test exists, and is given by

$$\phi(x) = 1\left\{x_{(n)} > \theta_0 \text{ or } x_{(n)} < \theta_0 \alpha^{1/n}\right\}$$

**4. Bayesian hypothesis testing**

Consider a univariate Gaussian problem with $X \mid \theta \sim N(\theta, 1)$, where $\theta = 0$ under the null hypothesis and $\theta \sim \Lambda_1$ under the alternative hypothesis (assume $\Lambda_1(\{0\}) = 0$). In addition let $\pi_0$ denote the *a priori* probability that the null hypothesis is true; therefore the full prior is a mixture between a point mass at 0 and $\Lambda_1$.

(a) Compute the posterior probability that the null hypothesis is true, i.e.

$$\pi_{\text{post}}(x; \Lambda_1, \pi_0) = \mathbb{P}(\theta = 0 \mid X = x).$$

(b) Assume $\pi_0 = 0.5$ (we are initially agnostic between the null and the alternative), and find

$$\pi^*_{\text{post}}(x) = \min_{\Lambda_1} \pi_{\text{post}}(x; \Lambda_1, 0.5),$$

as a function of $x$, for $x > 0$. Give the minimizing prior $\Lambda_1$, which also depends on $x$.

**Note:** This is not an optimization problem the analyst is going to solve: it is definitely not allowed to choose the prior after seeing the data. Instead, think of a large and diverse population of analysts who all have different priors before seeing the data, and therefore different posteriors after seeing the data (but with the constraint that none of them are initially "biased" against the null). Then we

as theoreticians are calculating a lower bound $\pi^*_{\text{post}}$ for any of these analysts' conditional belief in the null: all of their posterior credences in the null will be at least $\pi^*_{\text{post}}$. So everyone has their own prior but the only way someone could be really convinced that the null is false (more convinced than $1 - \pi^*_{\text{post}}$) is if they already thought it was probably false before seeing the data.

(c) Now restrict $\Lambda_1 = N(0, \tau^2)$ for $\tau > 0$, a class of more "realistic" priors. Compute $\pi_{\text{post}}$ as a function of $\tau^2$ and $x$. Find

$$\pi^*_{\text{post},N}(x) = \min_{\tau^2 > 0} \pi_{\text{post}}(x; N(0, \tau^2), 0.5),$$

and give the minimizing value of $\tau^2$, both as functions of $x$, for $x > 1$.

(d) Now assume we observe a value of $X$ such that the two-sided $p$-value $p(X)$ (i.e., $p(x) = \mathbb{P}_0(|X| > |x|)$) takes the values $0.05, 0.01, 0.005$, or $0.001$. Numerically compute $\pi^*_{\text{post}}$ and $\pi^*_{\text{post},N}$ for each value and make a small table. In words, interpret the results.

**Moral:** $p$-values are commonly misinterpreted as representing "the probability that the null hypothesis is true, given the data." This is an Bayesian statement and it depends on our prior beliefs. In fact, as this problem shows, even in a Bayesian setting, the $p$-value is generally not a good approximation for the posterior probability that the null is true.

**5. Mean estimation**

(a) Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N_d(\theta, I_d)$ and consider estimating $\theta \in \mathbb{R}^d$. Show that $\overline{X} = \frac{1}{n}\sum_i X_i$ is the minimax estimator of $\theta$ under squared error loss.

**Hint:** Find a least favorable sequence of priors.

(b) Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$ where $P$ is any distribution over the real numbers such that $\text{Var}_P(X) \leq 1$. Show that $\overline{X} = \frac{1}{n}\sum_i X_i$ is minimax for estimating $\theta(P) = \mathbb{E}_P X$ under the squared error loss.

**Hint:** Try to relate this problem to the Gaussian problem with $d = 1$.

(c) Assume $X \sim N(\theta, 1)$ with the constraint that $|\theta| \leq 1$. Show that the minimax estimator for squared error loss is

$$\tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}.$$

Plot its risk function.

**Hint:** Plot the risk function first. For this problem if you need to show that a function is maximized or minimized somewhere, you may do it numerically or by inspecting a graph if it is obvious enough.

# Stats 210A, Fall 2023
# Homework 8

**Due date**: Wednesday, Oct. 25

## 1. Directional error claims

Suppose $\mathcal{P} = \{P_\theta : \theta \in \mathbb{R}\}$, and $T(X)$ is a continuous test statistic that is stochastically increasing in $\theta$, meaning

$$\mathbb{P}_{\theta_1}(T(X) \leq t) \leq \mathbb{P}_{\theta_0}(T(X) \leq t), \quad \text{for all } t \in \mathbb{R} \text{ and } \theta_1 > \theta_0.$$

As we have discussed in class, this property guarantees that a one-tailed test of $H_0 : \theta \leq 0$ vs. $H_1 : \theta > 0$ that rejects for large values of $T(X)$, with cutoff $c$ chosen to solve $\mathbb{P}_0(T(X) > c) = \alpha$, will give a valid level-$\alpha$ test over the whole null distribution.

Now let $\phi(X)$ represent any level-$\alpha$ *two-tailed* test of $H_0 : \theta = 0$ vs. $H_1 : \theta \neq 0$ that rejects for extreme values of $T(X)$.

Assume that we also always make a *directional claim* about sign$(\theta)$ whenever we reject $H_0$. That is, we make one of *three* decisions: if $T(X) \in [c_1, c_2]$ then we do not reject $H_0$ (and we make no claim about the sign of $\theta$); if $T(X) > c_2$, then we reject $H_0$ and claim further that $\theta > 0$; and if $T(X) < c_1$, then we reject $H_0$ and claim further that $\theta < 0$. So now there are two kinds of Type I errors we could make: we could reject $H_0$ when it is really true, *or* we could reject $H_0$ when it is really false but call the sign wrong. Show that we have control of the *directional error rate*:

$$\sup_{\theta \in \mathbb{R}} \mathbb{P}_\theta(\text{False rejection or wrong sign call}) \leq \alpha$$

**Moral:** Having MLR or exponential families are nice to be able to talk about optimal tests, but stochastically increasing is a useful condition for getting valid tests in various contexts.

In particular, people often complain that we do not learn anything about $\theta$ by rejecting $H_0 : \theta = 0$, because we should have already known $\theta$ was not exactly zero. This line of argument ignores the fact that (in most testing settings) we can also draw a definite conclusion about the sign of $\theta$ whenever we reject $H_0 : \theta = 0$, without inflating the error rate.

## 2. Some two-tailed tests

Consider testing $H_0 : \theta = \theta_0$ vs $H_1 : \theta \neq \theta_0$ in a one-parameter exponential family of the form $p_\theta(x) = e^{\theta T(x) - A(\theta)} h(x)$. We stated in class that among all *unbiased, level-*$\alpha$ tests, the one that rejects for extreme (i.e., large or small) values of $T(X)$ is uniformly most powerful (simultaneously maximizes power for all alternatives).

The equal-tailed level-$\alpha$ test that rejects for extreme values of $T(X)$ does not satisfy as interesting an optimality property but it is also a competitive test. Depending on the distribution, the equal-tailed test and the UMPU test may or may not coincide.

Numerically find the equal-tailed and UMPU test for the following hypothesis testing problems at level $\alpha = 0.05$. For each problem,

(i) derive the appropriate tests (leaving the cutoff values abstract),

(ii) numerically compute the cutoff values $c$ (no $\gamma$ necessary since these are continuous problems), and

(iii) invert the equal-tailed test to give an interval for the data value specified (no need to invert the unbiased test).

(a) $X_i \overset{\text{ind.}}{\sim} N(\theta, \sigma_i^2)$ for $i = 1, \ldots, n$, where $\sigma_i^2$ are known positive constants and $\theta \in \mathbb{R}$ is unknown. Test $H_0 : \ \theta = 0$ vs. $H_1 : \ \theta \neq 0$, with $n = 20$ and $\sigma_i^2 = i$. On your power plot, also plot the power function of the (sub-optimal) test that rejects for extreme values of $\sum_i X_i$.

(b) $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} \text{Pareto}(\theta) = \theta x^{-(1+\theta)}$, for $\theta > 0$ and $x > 1$ (also called a power law distribution). Test $H_0 : \ \theta = 1$ vs. $H_1 : \ \theta \neq 1$, for $n = 100$. On your power plot, also plot the power function of the (sub-optimal) test that rejects for large $\sum_i X_i$.

**3. Maximizing average power**

In situations where there is not a UMP test, we cannot simultaneously maximize power for all alternatives. However, if the null is simple ($\Theta_0 = \{\theta_0\}$) and we have a prior $\Lambda_1$ over the alternative parameter space $\Theta_1$, we can maximize average power by rejecting for large values of:

$$T(x) = \frac{\int_{\Theta_1} p_\theta(x) \, \mathrm{d}\Lambda_1(\theta)}{p_{\theta_0}(x)}$$

Show that this test maximizes the average-case power $\int_{\Theta_1} \mathbb{E}_\theta \phi(X) \, \mathrm{d}\Lambda_1(\theta)$ among all tests with level $\alpha$.

**Hint:** Show that it can be viewed as a Neyman-Pearson test for a particular simple alternative.

**4. $p$-value densities**

Suppose $\mathcal{P}$ is a family with monotone likelihood ratio in $T(X)$, and the distribution of $T(X)$ is continuous with common support for all $\theta$. Let $\phi_\alpha$ denote the UMP level-$\alpha$ test of $H_0 : \theta \leq \theta_0$ vs. $H_0 : \theta > \theta_0$ that rejects when $T(X)$ is large. Let $p(X)$ denote the resulting $p$-value. Show that $p(X) \sim \text{Unif}[0, 1]$ if $\theta = \theta_0$, has non-increasing density on $[0, 1]$ if $\theta > \theta_0$, and has non-decreasing density on $[0, 1]$ if $\theta < \theta_0$.

**Note:** As always there is some ambiguity in how we could define the density; to resolve this ambiguity note it is equivalent to show that the CDF is linear, concave, or convex, or you can define the density unambiguously as the derivative of the CDF.

# Stats 210A, Fall 2023
# Homework 9

**Due date**: Thursday, Nov. 2

**Instructions:** As usual.

**1. Fisher's exact test**

Suppose $X_i \sim \text{Binom}(n_i, \pi_i)$ independently for $i = 0, 1$ and $\pi_0, \pi_1 \in (0, 1)$. Consider testing $H_0 : \pi_1 \leq \pi_0$ vs. $H_1 : \pi_1 > \pi_0$.

(a) A natural object of inference in this model is the *odds ratio*:

$$\rho = \frac{\pi_1/(1-\pi_1)}{\pi_0/(1-\pi_0)}.$$

Write the model in exponential family form with $\theta = \log \rho$ as one of the natural parameters, and reframe $H_0$ as an equivalent hypothesis about $\theta$.

(b) Find the UMPU level-$\alpha$ test of $H_0$, giving the cutoffs $c(u), \gamma(u)$ in terms of solutions to integral equalities for a hypergeometric distribution.

(c) Suppose $n_0 = n_1 = 40$, $X_0 = 18$ and $X_1 = 7$. Give a $95\%$ confidence interval for the odds ratio $\rho$ by numerically inverting the two-sided, equal-tailed, conditional test of $H_0 : \rho = \rho_0$ vs. $H_1 : \rho \neq \rho_0$. Don't randomize the interval, just return the conservative non-randomized interval. (Hint: it is equivalent to set up the problem in terms of $\theta$, and may be a little easier to think about that way.)

**Note:** Fisher's exact test is almost certainly the most important non-Gaussian example of a UMPU test with nuisance parameters, and has been used in countless clinical trials and observational studies. For example, we might give $n_1$ cardiac disease patients a new drug and give $n_0$ a placebo, then observe how many patients in each group suffer a heart attack within the next 5 years.

**2. Comparing variances**

Consider testing $H_0 : \sigma^2 \leq \tau^2$ vs. $H_1 : \sigma^2 > \tau^2$ in the two-sample Gaussian model with

$$X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2), \quad Y_1, \ldots, Y_m \overset{\text{i.i.d.}}{\sim} N(\nu, \tau^2),$$

where $X$ is independent of $Y$ and all parameters are unknown.

Define the sample mean and sample variance as

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

and define $\overline{Y}$ and $S_Y^2$ analogously.

(a) Show that $S_X^2/S_Y^2 \sim F_{n-1,m-1}$ if $\sigma^2 = \tau^2$ (i.e., on the boundary of the null).

(b) Show that the test that rejects for large values of $S_X^2/S_Y^2$ is UMPU (Hint: it may be helpful to recall that $\overline{X}, S_X^2, \overline{Y}$, and $S_Y^2$ are mutually independent by Basu's theorem, and that $(n-1)S_X^2 = \|X\|^2 - n\overline{X}^2$.)

**3. One-sample $t$-interval**

If $Z \sim N(0,1)$ and $V \sim \chi^2_d$ with $Z, V$ independent, we say that $T = Z/\sqrt{V/d}$ follows a *Student's $t$ distribution* with $d$ degrees of freedom, denoted by $T \sim t_d$. Note that $T^2 \sim F_{1,d}$ but $T$ preserves sign information in case we want to do one-sided tests.

Now suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\mu, \sigma^2)$ with $\sigma^2 > 0$ unknown and consider testing $H_0 : \ \mu = \mu_0$ vs. $H_1 : \ \mu \neq \mu_0$.

We showed in class that the one-sided UMPU test for $H_0 : \ \mu \leq 0$ vs. $H_1 : \ \mu > 0$ rejects for large values of $T_X = \frac{\overline{X}\sqrt{n}}{\sqrt{S_X^2}}$, where $S_X^2$ is defined as in Problem 2.

(a) Show that $T_X \sim t_{n-1}$ if $\mu = 0$ (see hint for previous problem).

(b) To test $H_0 : \ \mu = 0$ vs. $H_1 : \ \mu \neq 0$, show that the UMPU test rejects for large values of $|T_X|$ (Hint: the simplest way is to use symmetry).

(c) Find a UMPU test of $H_0 : \ \mu = \mu_0$ for a generic $\mu_0 \in \mathbb{R}$, and invert to find a confidence interval for $\mu$ in terms of $\overline{X}$, $S_X^2$, quantiles of the $t_{n-1}$ distribution, and the desired level $\alpha$ (Hint: consider the distribution of $X_i - \mu_0$).

**4. McNemar's test**

Suppose we have paired binary data: for $i = 1, \ldots, n$ we observe $(X_i, Y_i) \in \{0,1\}^2$. The pairs are i.i.d. with

$$\mathbb{P}\left[(X_i, Y_i) = (a, b)\right] = \pi_{a,b} \quad a, b \in \{0, 1\}.$$

Write $\pi_X = \mathbb{P}(X_i = 1) = \pi_{1,0} + \pi_{1,1}$ and $\pi_Y = \mathbb{P}(Y_i = 1) = \pi_{0,1} + \pi_{1,1}$, and let $N_{a,b} = \sum_{i=1}^n 1\{X_i = a, Y_i = b\}$.

(a) Find the UMPU test of $H_0 : \ \pi_X \leq \pi_Y$ vs. $H_1 : \ \pi_X > \pi_Y$, giving the cutoffs $c(u), \gamma(u)$ in terms of solutions to integral equalities for a binomial distribution. (Hint: it may help to first reframe the hypothesis in terms of the $\pi_{a,b}$ parameters.)

(b) Suppose $N_{0,0} = N_{1,1} = 1000$, $N_{0,1} = 5$ and $N_{1,0} = 25$. Compute 95% confidence intervals for $\pi_X$ and $\pi_Y$ (invert the two-sided equal-tailed test but without randomizing). Then compute a $p$-value for $H_0 : \ \pi_X \leq \pi_Y$ (do not randomize). Does anything about the respective answers surprise you?

(Note: This test is called McNemar's test; it is very useful for clinical trials with matched pairs of subjects, and also for comparing the performance of different classifiers on a held-out sample.)

**Moral:** When we have paired data, we can often make much more precise comparisons between two distributions; even more precise than our ability to infer things about either of the distributions individually. This is often worth taking into account if we are designing an experiment: for example, if we match patients into pairs on demographic characteristics and then randomize a treatment/placebo assignment within each pair, we may get a very good inference about whether the treatment is better than the placebo, much better than we would get if we randomly assigned all $2n$ subjects independently of each other.

**5. Nonparametric tests**

In this problem you will design tests for two nonparametric hypothesis testing problems. There is necessarily some wiggle room in how you choose the test statistic, and it will probably not be possible to determine the cutoff explicitly. Just choose a reasonable one, define the cutoff in terms of a quantile of a well-defined distribution, and show that your test has significance level $\alpha$.

(a) Suppose $X_1, \ldots, X_n \in \mathbb{R}$ are independent random variables with $X_i \sim P_i$. Consider testing the null hypothesis $H_0 : \ P_1 = P_2 = \cdots = P_n$ (i.e., the observations are i.i.d.) against the alternative that there is a systematic trend toward larger values of $X_i$ as $i$ increases (this is sometimes called a *test of trend*). Design a level-$\alpha$ test.

(b) Suppose $(X_1, Y_1), \ldots, (X_n, Y_n) \overset{\text{i.i.d.}}{\sim} P$ where $P$ is an unknown joint distribution on $\mathbb{R}^2$. Consider testing the null hypothesis that $X_i$ and $Y_i$ are independent within each pair (i.e., $P = P_X \times P_Y$, with $P_X$ and $P_Y$ unknown and not necessarily the same) versus the alternative that $(X_i, Y_i)$ are positively correlated within each pair. Design a level-$\alpha$ test.

Note that the alternative is defined a little vaguely in each part above. If that troubles you, we could formally take the alternative be "$P_i$ are arbitrary but not all equal" in part (a), or "$P \neq P_X \times P_Y$" in part (b). The alternative hypotheses as I've defined them informally are meant to suggest which alternatives to prioritize when you design your test.

**Moral:** We can often design our own nonparametric tests by conditioning on an appropriate sufficient statistic for the null distribution.

# Stats 210A, Fall 2023
# Homework 10

**Due date**: Wednesday, Nov. 8

**1. Multidimensional testing**

Suppose $X \sim N_d(\mu, I_d)$ for unknown $\mu \in \mathbb{R}^d$. Consider testing $H_0 : \ \mu = 0$ vs. $H_1 : \ \mu \neq 0$. You may take as given the fact that if $d = 1$ the UMPU test for the Gaussian location family is unique: i.e., it is the only UMPU test for that model up to almost sure equality.

(a) Show that for any $d > 1$ and $\alpha \in (0, 1)$, there exists no UMP or UMPU level-$\alpha$ test.
**Hint:** what would we do if we knew $\mu = (\theta, 0, 0, \ldots, 0)$ for an unknown $\theta \in \mathbb{R}$?

(b) Suppose we have a prior $\Lambda_1$ for the value that $\mu$ takes under the alternative; that is, $\mu \sim \Lambda_1$ if $H_1$ is true and $\mu = 0$ if $H_0$ is true. Define the average power as

$$\int_{\mathbb{R}^d} \mathbb{E}_\mu[\phi(X)] \, \mathrm{d}\Lambda_1(\mu).$$

If $\Lambda_1 = N(\nu, \Sigma)$, with positive definite covariance matrix $\Sigma$, find the level-$\alpha$ test that maximizes the average power. Show that the acceptance region is an ellipse centered at 0 if $\nu = 0$.
**Hint:** You can use the result from homework 8.

(c) **Optional:** Show that if $\Lambda_1$ is rotationally invariant, the $\chi^2$ test that rejects for large $\|X\|^2$ maximizes the average power.

**Moral:** Choosing a test in higher dimensions requires us to think harder about how to compromise across different alternative directions, and Bayesian thinking can give us some guidance.

**2. James-Stein estimator with regression-based shrinkage**

Consider estimating $\theta \in \mathbb{R}^n$ in the model $Y \sim N_n(\theta, I_n)$. In the standard James-Stein estimator, we shrink all the estimates toward zero, but it might make more sense to shrink them towards the average value $\overline{Y}$, or towards some other value based on observed side information.

(a) Consider the estimator

$$\delta_i^{(1)}(Y) = \overline{Y} + \left(1 - \frac{n-3}{\|Y - \overline{Y}1_n\|^2}\right)(Y_i - \overline{Y})$$

Show that $\delta^{(1)}(Y)$ strictly dominates the estimator $\delta^{(0)}(Y) = Y$, for $n \geq 4$.

$$\text{MSE}(\theta; \delta^{(1)}) < \text{MSE}(\theta; \delta^{(0)}), \quad \text{for all } \theta \in \mathbb{R}^n.$$

Calculate the MSE of $\delta^{(1)}$ if $\theta_1 = \theta_2 = \cdots = \theta_n$.
**Hint:** Change the basis and think about how the estimator operates on different subspaces.

(b) Now suppose instead that we have side information about each $\theta_i$, represented by covariate vectors $x_1, \ldots, x_n \in \mathbb{R}^d$. Assume the design matrix $X \in \mathbb{R}^{n \times d}$ whose $i$th row is $x_i'$ has full column rank. Suppose that we expect $\theta \approx X\beta$ for some $\beta \in \mathbb{R}^d$, but unlike the usual linear regression setup, we will not assume $\theta = X\beta$ with perfect equality.

Find an estimator $\delta^{(2)}$, analogous to the one in part (a), that dominates $\delta^{(0)}$ whenever $n - d \geq 3$:

$$\text{MSE}(\theta; \delta^{(2)}) < \text{MSE}(\theta; \delta^{(0)}), \quad \text{for all } \theta \in \mathbb{R}^n,$$

and for which $\text{MSE}(X\beta; \delta^{(2)}) = d + 2$, for any $\beta \in \mathbb{R}^d$.

**Hint:** Think of this setting as a generalization of part (a), which can be considered a special case with $d = 1$ and all $x_i = 1$.

**3. Confidence regions for regression**

Assume we observe $x_1, \ldots, x_n \in \mathbb{R}$, which are not all identical (for some $i$ and $j$, $x_i \neq x_j$). We also observe

$$Y_i = \beta_0 + \beta_1 x_i + \varepsilon_i, \text{ for } \varepsilon_i \overset{\text{i.i.d.}}{\sim} N(0, \sigma^2).$$

$\beta_0, \beta_1 \in \mathbb{R}$ and $\sigma^2 > 0$ are unknown. Let $\bar{x}$ represent the mean value $\frac{1}{n}\sum_i x_i$.

(a) Give an explicit expression for the $t$-based confidence interval for $\beta_1$, in terms of a quantile of a Student's $t$ distribution with an appropriate number of degrees of freedom (feel free to break up the expression, for example by first giving an expression for $\hat{\beta}_1$ and then using $\hat{\beta}_1$ in your final expression). You do not need to show the interval is UMAU.

**Hint:** It may be helpful to consider a translation of the model similar to what we did in Problem 3 of Homework 8.

(b) Invert an $F$-test to give a *confidence ellipse* for $(\beta_0, \beta_1)$. It may be convenient to represent the set as an affine transformation of the unit ball in $\mathbb{R}^2$:

$$b + A\mathbb{B}_1(0) = \{b + Az : \ z \in \mathbb{R}^2, \|z\| \leq 1\}, \quad \text{for } b \in \mathbb{R}^2, A \in \mathbb{R}^{2\times 2}.$$

Give explicit expressions for $b$ and $A$ in terms of a quantile of an appropriate $F$ distribution.

**Hint:** Consider the joint distribution of $(\hat{\beta}_0 - \beta_0, \hat{\beta}_1 - \beta_1)$.

**Hint:** Use the fact that $\binom{\hat{\beta}_0}{\hat{\beta}_1} \sim N_2\left(\binom{\beta_0}{\beta_1}, \ \sigma^2(X'X)^{-1}\right)$. You do not need to show that the confidence ellipse you come up with has any optimality properties.

**4. Confidence bands for regression**

The setup for this problem is the same as for Problem 4 only now we are interested in giving *confidence bands* for the regression line $f(x) = \beta_0 + \beta_1 x$. In this problem you do not need to give explicit expressions for everything, but you should be explicit enough that someone could calculate the bands based on your description.

(a) For a fixed value $x_0 \in \mathbb{R}$ (not necessarily one of the observed $x_i$ values) give a $1 - \alpha$ $t$-based confidence interval for $f(x_0) = \beta_0 + \beta_1 x_0$. That is, we want to find $C_1^P(x_0), C_2^P(x_0)$ such that

$$\mathbb{P}\left(C_1^P(x_0) \leq f(x_0) \leq C_2^P(x_0)\right) = 1 - \alpha.$$

The functions $C_1^P(x), C_2^P(x)$ that we get from performing this operation on all $x$ values give a *pointwise confidence band* for the function $f(x)$.

(b) Now give a *simultaneous confidence band* around $f(x) = \beta_0 + \beta_1 x$. That is, give $C_1^S(x), C_2^S(x)$ with

$$\mathbb{P}\left(C_1^S(x) \leq f(x) \leq C_2^S(x), \text{ for all } x \in \mathbb{R}\right) = 1 - \alpha,$$

and show that your confidence band has this property.

**Hint:** If all we know is that $(\beta_0, \beta_1)$ is in the confidence ellipse from Problem 4, what can we deduce about $f(x)$?

(c) Download the data set in `hw10-4.csv` from the course web site and make a scatter plot of the data. Plot the OLS regression line as well as the two confidence bands. Describe what you see. What do the bands do as $x$ goes away from the data set, and why does this make sense?

# Stats 210A, Fall 2023
# Homework 11

**Due date**: Wednesday, Nov. 15

### 1. Some Maximum Likelihood Estimators

Find the MLE for each model below, show that it is consistent, and find its asymptotic distribution. You may assume our Taylor expansions from class are valid without checking conditions.

(a) Binomial: $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \text{Binom}(m, \theta)$. Find the MLE for $\theta$ and for the natural parameter $\eta = \log \frac{\theta}{1-\theta}$.

(b) Gaussian: $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, \sigma^2)$. Find (i) the MLE for $\theta$ if $\sigma^2$ is known, (ii) the MLE for $\sigma^2$ if $\theta$ is known, and (iii) the MLE for $(\theta, \sigma^2)$ if neither is known.

(c) Laplace: $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} \frac{1}{2} e^{-|x-\theta|}$. Assume $n$ is odd.
For this problem, the log-likelihood is non-differentiable at one point, but we can still use our formula for the asymptotic distribution of the MLE from class, with the Fisher information defined by $J_1(\theta) = \text{Var}_\theta[\dot{\ell}_1(\theta; X_i)]$. You may use this fact without proof.

(d) **Optional** (not graded, no extra points) For the Laplace, plot a few realizations of the log-likelihood for $n = 5000$ with $\theta_0 = 0$, and plot over it the quadratic approximation given by

$$\ell_n(\theta) - \ell_n(\theta_0) \approx \dot{\ell}_n(\theta_0)(\theta - \theta_0) - \frac{1}{2} n J_1(\theta_0)(\theta - \theta_0)^2.$$

Is the quadratic approximation pretty good in the neighborhood $\theta_0 \pm 3\sigma$, where $\sigma^2$ is the approximate variance of $\hat{\theta}_n$? Intuitively, what do you think might account for this when the second derivative doesn't exist?

### 2. Estimating the inverse of a mean

Suppose that $X_1, \dots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$, and that we are interested in estimating the quantity $1/\theta$. In order to do so, we use the estimator $\delta(X) = 1/\overline{X}_n$ where $\overline{X}_n = \frac{1}{n}\sum_{i=1}^n X_i$ is the sample mean. Assume $\theta \neq 0$.

(a) Show that $\delta$ is asymptotically normal, and find its asymptotic distribution.

(b) Show that the expectation $\mathbb{E}|1/\overline{X}_n| = \infty$ for every $n$. Why does this not contradict the result of part (a)?

(c) Simulate to find the distribution of $1/\overline{X}_n$ for $n = 10, 100, 10^4$ and $\theta = 0.1, 1, 10$. For each setting of the parameters, plot a histogram of the estimator and overlay its Gaussian approximation. When the Gaussian approximation is not good, what is going wrong? Is the sample size a reliable indicator of whether we should trust an asymptotic approximation?

**Hint:** If you are using R, the functions `hist` (with argument `freq = FALSE` to get a density histogram), `curve`, and `dnorm` will come in handy. Also, I recommend manually setting the `breaks` and `xlim` arguments in `hist` to stop enormous values from making your histogram uninformative: $\mu \pm 4\sigma$ is a reasonable range of values to plot, where $\mu$ and $\sigma^2$ are the mean and variance of the Gaussian approximation.

**3. Limiting distribution of $U$-statistics**

Suppose $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} P$ in some sample space $\mathcal{X}$. $U_n = U_n(X_1, \ldots, X_n)$ is called a rank-2 $U$-statistic if

$$U_n = \frac{1}{n(n-1)} \sum_{i=1}^{n} \sum_{j \neq i} h(X_i, X_j)$$

where $h$ is a symmetric function, i.e. $h(x_1, x_2) = h(x_2, x_1)$ for any $x_1, x_2 \in \mathcal{X}$.

In this problem, we denote $\theta = \mathbb{E}h(X_1, X_2)$ and assume that $\mathbb{E}h(X_1, X_2)^2 < \infty$. Note that $U_n$ is the nonparametric UMVU estimator of $\theta$.

Perhaps surprisingly, we can derive the asymptotic distribution of $U_n$ in a relatively small number of steps using a technique called *Hájek projection* where we approximate it by an additive function of the independent $X_i$ variables. We walk through the proof below.

(a) Define $g(x) = \mathbb{E}h(x, X_2) - \theta = \int h(x, u)\, \mathrm{d}P(u) - \theta$. Show that, for all $i$,

$$\mathbb{E}g(X_i) = 0, \quad \text{and} \quad \text{Var}(g(X_i)) < \infty.$$

(**Note:** $g$ is a specific function from $\mathcal{X}$ to $\mathbb{R}$. It is not a rule for naively substituting symbols into expressions. In particular, note that $g(X_i)$, a random variable, is not the same as the deterministic expression $\mathbb{E}h(X_i, X_2) - \theta$.)

(b) Define $\widehat{U}_n = \theta + \frac{2}{n} \sum_{i=1}^{n} g(X_i)$. Show that $\mathbb{E}[(U_n - \widehat{U}_n) f(X_i)] = 0$ for any $i$ and any measurable function $f(X_i)$ with $\mathbb{E}[f(X_i)^2] < \infty$.
(**Hint:** Condition on $X_i$)

(c) Show that $\sqrt{n}(U_n - \widehat{U}_n) \xrightarrow{p} 0$ as $n \to \infty$. (Hint: show that $U_n$ and $\widehat{U}_n$ have the same asymptotic variance, and then apply part (b)).

(d) Conclude that $\sqrt{n}(U_n - \theta) \Rightarrow N(0, 4\zeta_1)$, where $\zeta_1 = \text{Var}(g(X_1))$.

(e) Assume that $\mathcal{X} = \mathbb{R}$ with $\mathbb{E}X_i^4 < \infty$. Express the sample variance $S_n^2 = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \overline{X})^2$ as a rank-2 U-statistic and use the above results to derive its asymptotic distribution.

(**Note:** a similar result holds in general for rank-$r$ $U$-statistics if we set $\widehat{U}_n = \theta + \frac{r}{n} \sum_i g(X_i)$ where $g(x) = \mathbb{E}[h(x, X_2, \ldots, X_r)] - \theta$. )

**Moral:** If $P^n$ is the distribution of $(X_1, \ldots, X_n)$ then it is easy to check that the set of all square-integrable random variables of the form $f(X_1, \ldots, X_n)$ (where $f : \mathcal{X}^n \to \mathbb{R}$ is measurable) forms a vector space over $\mathbb{R}$, which we call $L^2(P^n)$, where we can define an inner product as

$$\langle f(X), g(X) \rangle_{L^2} = \mathbb{E}[f(X)g(X)] \leq \sqrt{\mathbb{E}[f(X)^2]\mathbb{E}[g(X)^2]} < \infty.$$

Moreover, the subset of those random variables that can be written as $\sum_i f_i(X_i)$, where each $f_i$ is measurable, forms a subspace. Part (b) establishes that the simpler random variable $\widehat{U}_n$ is the *projection* of $U_n$ onto this subspace, and part (c) establishes that $U_n$ is asymptotically very close to its projection.

**4. Probabilistic big-O notation**

Let $X_1, X_2, \ldots$ denote a sequence of random vectors (with $\|X_n\| < \infty$ almost surely for each $n$). We say the sequence is *bounded in probability* (or sometimes *tight*) if for every $\varepsilon > 0$ there exists a constant $M_\varepsilon > 0$ for which

$$\mathbb{P}(\|X_n\| > M_\varepsilon) < \varepsilon, \quad \forall n.$$

Informally, there is "no mass escaping to infinity" as $n$ grows. Like regular big-O notation, these symbols can help to make rigorous asymptotic proofs look clean and intuitive.

For a fixed sequence $a_n$, we say $X_n = o_p(a_n)$ if $X_n/a_n \xrightarrow{p} 0$ as $n \to \infty$, and $X_n = O_p(a_n)$ if the sequence $(X_n/a_n)_{n\geq 1}$ is bounded in probability.

Prove the following facts for $X_n, Y_n \in \mathbb{R}^d$:

(a) If $X_n \Rightarrow X$ for any random vector $X$, then $X_n = O_p(1)$.

(b) If $X_n = o_p(a_n)$ then $X_n = O_p(a_n)$.

(c) If $a_n/b_n \to 0$ and $X_n = O_p(a_n)$, then $X_n = o_p(b_n)$.

(d) If $X_n = O_p(a_n)$ and $Y_n = O_p(b_n)$ then $X_n + Y_n = O_p(\max\{a_n, b_n\})$.

(e) If $X_n = O_p(a_n)$ and $Y_n = o_p(b_n)$, then $X_n' Y_n = o_p(a_n b_n)$. If $X_n = O_p(a_n)$ and $Y_n = O_p(b_n)$, then $X_n' Y_n = O_p(a_n b_n)$.

(f) If $X_n = O_p(1)$ and $g : \mathbb{R}^d \to \mathbb{R}^k$ is continuous then $g(X_n) = O_p(1)$.

(g) For $d = 1$, if $X_n = O_p(a_n)$ with $a_n \to 0$ and $g : \mathbb{R} \to \mathbb{R}$ is continuously differentiable with $g(0) = \dot{g}(0) = 0$, then $g(X_n) = o_p(a_n)$. Show further that if $g$ is twice continuously differentiable then $g(X_n) = O_p(a_n^2)$. (**Hint:** Use the mean value theorem and apply a previous part of this problem.)

(h) For $d = 1$, if $\text{Var}(X_n) = a_n^2 < \infty$ and $\mathbb{E}X_n = b_n$ then $X_n = O_p(a_n + b_n)$. (**Hint:** Use Chebyshev's inequality.)

(i) **Optional** (not graded, no extra points): If $\text{Var}(X_n) = a_n^2 < \infty$, is it impossible to have $X_n = o_p(a_n)$? Prove or give a counterexample.

# Stats 210A, Fall 2023
# Homework 12
## Optional

### 1. MLE consistency for concave log-likelihoods

Assume $X_1, X_2, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta_0}(x)$ for some identifiable, dominated family with $\Theta = \mathbb{R}^d$. Assume additionally that $\ell_1(\theta; X_i) = \log p_\theta(X_i)$ is almost surely concave and continuously differentiable in $\theta$, and that for all compact sets $K \subseteq \mathbb{R}^d$, we have

$$\mathbb{E}_{\theta_0}\left[\sup_{\theta \in K} \|\nabla \ell_1(\theta; X_1)\|_2\right] < \infty.$$

Prove that the MLE is consistent: if $\hat{\theta}_n \in \text{argmax } \ell_n(\theta)$ then $\hat{\theta}_n \xrightarrow{p} \theta_0$ (You may assume a maximizer always exists; note we could always define $\hat{\theta}_n$ arbitrarily when there is none).

(**Hint:** The technique here is not just a small modification of what we used in our theorem from class for consistency with non-compact $\Theta$; it's a different argument entirely. But similarly to what we did in class, you should start by showing uniform convergence of $\overline{W}_n(\theta)$ on compact $K$, and then deal with the rest of $\mathbb{R}^d$.)

**Moral:** There is more than one way to get consistency of the MLE.

### 2. Logistic regression with random $X$

Consider a univariate logistic regression model where we observe $n$ i.i.d. pairs $(X_i, Y_i) \in \mathbb{R} \times \{0, 1\}$. The covariate is random with a known distribution, $X_i \overset{\text{i.i.d.}}{\sim} U[-1, 1]$, and

$$\mathbb{P}_{\alpha,\beta}(Y_i = 1 \mid X_i = x) = \frac{e^{\alpha+\beta x}}{1 + e^{\alpha+\beta x}}.$$

(a) Show that the maximum likelihood estimator for $(\alpha, \beta)$ solves

$$\sum_i Y_i = \sum_i \pi_i(\hat{\alpha}_n, \hat{\beta}_n)$$
$$\sum_i Y_i X_i = \sum_i \pi_i(\hat{\alpha}_n, \hat{\beta}_n) X_i,$$

where $\pi_i(\alpha, \beta) = e^{\alpha+\beta X_i}/(1 + e^{\alpha+\beta X_i})$.

(b) Use the result of the previous problem to show that the MLE is consistent, asymptotically Gaussian, and asymptotically efficient (you may ignore the fact that the MLE may not always exist in finite samples).

(c) For $\alpha = 0$, $\beta = 4$, calculate the Fisher information for a single pair $(X_i, Y_i)$; give it as an integral and also calculate it numerically (you do not need to analytically evaluate the integral). Note your answer should not depend on $X_i$, which is a random variable in this problem. Give the asymptotic distribution of the MLE, with a numerical answer for the asymptotic variance.

(d) For $\alpha = 0$, $\beta = 4$, and for each of a few different $n$ values:

(i) Generate a large number (e.g. 1000) of data sets of size $n$, and for each one compute the MLE $(\hat{\alpha}, \hat{\beta})$ (you can use statistical software to compute the MLE, e.g. the `glm` function in R).

(ii) Plot histograms of $\hat{\alpha}$ and $\hat{\beta}$ (if you use R, I recommend setting `freq=FALSE` to get a density histogram instead of a frequency histogram).

(iii) Overlay the Gaussian curve based on the approximate distribution from part (c) (you can use the `dnorm` function in R). About how big does $n$ need to be for the normal approximation to be pretty good?

(e) Repeat parts (c) and (d) for $\alpha = -6$ and $\beta = 4$. How is it the same or different, and what do you think accounts for why?

**3. Score test with nuisance parameters**

Consider a testing problem with $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} p_{\theta,\zeta}(x)$ with parameter of interest $\theta \in \mathbb{R}$ and nuisance parameter $\zeta \in \mathbb{R}$. That is, we are testing $H_0 : \theta = \theta_0$ vs. $H_1 : \theta \neq \theta_0$, and $\zeta$ is unknown; let $\zeta_0$ denote its true value. Then there is a version of the score test where we plug in an estimator for $\zeta$, but we must use a corrected version of the variance.

Let $\hat{\zeta}_0$ denote the maximum likelihood estimator of $\zeta$ under the null:

$$\hat{\zeta}_0(\theta_0) = \arg\max_{\zeta \in \mathbb{R}} \ \ell(\theta_0, \zeta; X).$$

Assume $\hat{\zeta}_0$ is consistent under the null hypothesis.

Let $J(\theta, \zeta)$ denote the full-sample Fisher Information (omitting the usual $n$ subscript), and assume it is continuous and positive-definite everywhere.

(a) Use Taylor expansions informally to show that, for large $n$,

$$\frac{\partial}{\partial \theta}\ell(\theta_0, \hat{\zeta}_0) \approx \frac{\partial}{\partial \theta}\ell(\theta_0, \zeta_0) - \frac{\frac{\partial^2}{\partial\theta\partial\zeta}\ell(\theta_0, \zeta_0)}{\frac{\partial^2}{\partial\zeta^2}\ell(\theta_0, \zeta_0)} \ \frac{\partial}{\partial \zeta}\ell(\theta_0, \zeta_0).$$

(Note: the LHS should be read as $[\frac{\partial}{\partial\theta}\ell(\theta, \zeta)]\big|_{\theta_0, \hat{\zeta}_0}$, and **not** $\frac{d}{d\theta_0}[\ell(\theta_0, \hat{\zeta}_0(\theta_0))]$).

(b) Using part (a), conclude that

$$\left(J_{11} - \frac{J_{12}^2}{J_{22}}\right)^{-1/2} \ \frac{\partial}{\partial \theta}\ell(\theta_0, \hat{\zeta}_0) \Rightarrow N(0, 1) \quad \text{as } n \to \infty$$

where $J = J(\theta_0, \hat{\zeta}_0)$. Compare this to the score test statistic we would use if $\zeta_0$ were known rather than estimated. (Note: you may assume without proof that the approximation error in part (a) is negligible; i.e. you may take the "$\approx$" as an exact equality).

**Moral:** The score test can be carried out with nuisance parameters, but the fact that we estimate the nuisance parameter affects the distribution of the test statistic in a way that we need to take into account.

**4. Poisson score test**

Suppose that for $i = 1, \ldots, x_n$ we observe a real covariate $x_i \in \mathbb{R}$ (fixed and known) and a Poisson response $Y_i \sim \text{Pois}(\lambda_i)$. We assume that $\lambda_i = \alpha + \beta x_i$, with the restriction that $\lambda_i \geq 0$ for all $i$, but with $\alpha, \beta \in \mathbb{R}$ otherwise unrestricted. Assume that

$$\lim_{n\to\infty} \frac{\sum_{i=1}^n |x_i - \bar{x}_n|^3}{\left(\sum_{i=1}^n (x_i - \bar{x}_n)^2\right)^{3/2}} = 0,$$

where $\bar{x}_n = n^{-1}\sum_{i=1}^n x_i$. We observe the first $n$ pairs $(x_i, y_i)$ and our goal is to test the hypothesis $H_0 : \beta = 0$ vs. $H_1 : \beta > 0$. Assume that there are at least 3 distinct values represented among $x_1, \ldots, x_n$.

(a) Show that this model is a curved exponential family.

(b) Derive the score test statistic for $H_0$ vs $H_1$. Give the test statistic and asymptotic rejection cutoff. It is not necessary to normalize it for this part.

(c) Show that your test statistic is indeed asymptotically normally distributed, and find an asymptotically valid rejection cutoff.

**Hint**: It may help to use the *Lyapunov CLT*, which applies to sums of independent random variables that are not necessarily identically distributed: Suppose $Z_1, Z_2, \ldots$ is a sequence of random variables with $Z_i \sim (\mu_i, \sigma_i^2)$, for $\sigma_i^2 < \infty$. Define $s_n^2 = \sum_{i=1}^n \sigma_i^2$. If for some $\delta > 0$, we have

$$\lim_{n\to\infty} \frac{1}{s_n^{2+\delta}} \sum_{i=1}^n \mathbb{E}\left[|Z_i - \mu_i|^{2+\delta}\right] = 0,$$

then $s_n^{-1} \sum_{i=1}^n (Z_i - \mu_i) \Rightarrow N(0,1)$.

(d) Suppose $n$ is small, so we don't want to rely on the asymptotic normality. Explain how we could find a finite-sample exact conditional cutoff for the score test from part (b) (it is not necessary to prove any optimality property).

**5. Super-Efficient Estimator**

Let $X_1, \ldots, X_n \overset{\text{i.i.d.}}{\sim} N(\theta, 1)$ and consider estimating $\theta$ via:

$$\delta_n(X) = \overline{X}_n 1\{|\overline{X}_n| > a_n\},$$

where $a_n \to 0$ but $a_n\sqrt{n} \to \infty$ as $n \to \infty$ (for example, $a_n = n^{-1/4}$).

(a) Show that $\delta_n$ has the same asymptotic distribution as $\overline{X}_n$ when $\theta \neq 0$, but that $\sqrt{n}(\delta_n - 0) \xrightarrow{p} 0$ if $\theta = 0$.

(b) Show that, pointwise in $\theta$, as $n \to \infty$,

$$n\,\text{MSE}(\delta_n; \theta) \to 1\{\theta \neq 0\},$$

but that the convergence is not uniform in $\theta$; in fact,

$$\sup_{\theta\in\mathbb{R}} \; n\,\text{MSE}(\delta_n; \theta) \to \infty.$$

(**Note**: this is an example of a situation where it is incorrect to exchange a limit with a supremum.)

(c) **Optional**: Can you find a scaling of $\delta_n$ that converges to a non-degenerate distribution when $\theta = 0$? What is the limiting distribution?

**Moral:** The sense in which asymptotically efficient estimators are "optimal" is not easy to define, and it isn't obvious how we should compare the asymptotic behavior of different estimators. In this example it would appear initially that the super-efficient estimator renders the sample mean inadmissible. But this is only true if we look at the pointwise limit for fixed $\theta$; at any $n$ there are some values of $\theta$ for which the estimator is performing very badly, and this gets worse and worse as $n$ gets larger.