# 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 $\text{Var}_\eta(a'T(X)) > 0$ 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. $\text{Unif}[0, u]$, 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.