**8th problem set,**

Since you're not all math majors I include this "reminder": If $A$ is a set of real numbers then "$\sup A$" does not quite mean the same thing as "$\max A$". For example, the set $A = [2,3] = \{\, x : 2 \le x \le 3 \,\}$ has a "max", which is 3, but the set $B = [2,3) = \{\, x : 2 \le x < 3 \,\}$, although it has no "max", still has a "sup", which is also 3. That is because 3 is the smallest number that no member of $B$ exceeds (even though $B$ has no largest member). And similarly 3 is the smallest number that no member of $A$ exceeds (because 3 is the largest member of $A$).

In class we have seen likelihood-ratio statistics of the form $\dfrac{\sup\{\, f(x \mid \theta) : \theta \in \Omega_0 \,\}}{\sup\{\, f(x \mid \theta) : \theta \in \Omega_0 \cup \Omega_1 \,\}}$. Here we will work with a simpler kind of likelihood ratio:

$$\Lambda(x) = \frac{f(x \mid \theta = \theta_1)}{f(x \mid \theta = \theta_0)} = \frac{L(\theta_1)}{L(\theta_0)}.$$

With a lower-case "$x$", this defines a function. With a capital "$X$", this is a random variable. We will consider how this random variable is distributed when $\theta_0$ rather than $\theta_1$ is the "true" value of $\theta$. Assume $X$ is a continuous random variable whose density, given that $\theta = \theta_0$, is $f(x \mid \theta = \theta_0)$. Use "$\mathbf{E}_{\theta_0}(\quad)$", "$\mathbf{var}_{\theta_0}(\quad)$", "$\mathbf{cov}_{\theta_0}(\quad)$", etc., to mean these operators are to be evaluated assuming $\theta = \theta_0$.

1. Show that $\mathbf{E}_{\theta_0}(\Lambda(X)) = 1$.

2. (a) The following may or may not be useful in doing part (b), perhaps depending on your tastes: Suppose $\mathbf{E}(U) = \mu$ and $\mathbf{E}(V) = \nu$. Show that $\mathbf{cov}(U,V)$ can be written in any of these forms (the second one is the definition, so there is nothing to show):

$$\mathbf{E}(U(V-\nu)) = \mathbf{E}((U-\mu)(V-\nu)) = \mathbf{E}((U-\mu)V).$$

   (b) Suppose $T(X)$ is an unbiased estimator of $\theta$. Show that $\mathbf{cov}_{\theta_0}(T(X), \Lambda(X)) = \theta_1 - \theta_0$.

3. Use the Cauchy-Schwartz inequality in the form $|\mathbf{cov}(U,V)| \le \mathbf{SD}(U)\,\mathbf{SD}(V)$ and the result of #2(b) to show that when $\theta = \theta_0$ then the mean squared error of $T(X)$ as an estimator of $\theta$ cannot be less than

$$\frac{(\theta_1 - \theta_0)^2}{\mathbf{var}_{\theta_0}(\Lambda(X))}.$$

CONTINUED⟶

4. Briefly comment on the justification for, and validity of, each step below.

$$\begin{aligned}
\frac{\mathbf{var}_{\theta_0}(\Lambda(X))}{(\theta_1-\theta_0)^2} &= \mathbf{var}_{\theta_0}\left(\frac{\Lambda(X)-1}{\theta_1-\theta_0}\right) \\
&= \mathbf{var}_{\theta_0}\left(\frac{1}{f(X\mid\theta=\theta_0)}\cdot\frac{f(X\mid\theta=\theta_1)-f(X\mid\theta=\theta_0)}{\theta_1-\theta_0}\right) \\
&\longrightarrow \mathbf{var}_{\theta_0}\left(\left.\frac{\partial}{\partial\theta}\log f(X\mid\theta)\right|_{\theta=\theta_0}\right) \quad \text{as } \theta_1\longrightarrow\theta_0.
\end{aligned}$$

5. Let $I(\theta)=\mathbf{var}_\theta\left(\frac{\partial}{\partial\theta}\log f(X\mid\theta)\right)$. This quantity $I(\theta)$ is called the "Fisher information" in the sample $X$. It is thought of as a measure of how much information about $\theta$ is conveyed by the sample $X$. If $X$ happens to be a sample of size $n$, so that $X=(X_1,\ldots,X_n)$, then let us denote this by $I_n(\theta)$, so that, in particular, $I_1(\theta)$ would be the amount of information in a sample of size 1. Show that $n$ times as much information is in a sample of size $n$ than in a sample of size 1, i.e., show that $I_n(\theta)=nI_1(\theta)$.

6. Use conclusions from #3 and #4 to show that the mean squared error of an unbiased estimator $T(X)$ of $\theta$ cannot be less than $1/I(\theta)$. (This is the "Cramer-Rao inequality" or the "information inequality." The quantity $1/I(\theta)$ is the "Cramer-Rao lower bound.")

7. The Fisher information $I(\theta)$ can sometimes be more readily computed by the result you will derive in this problem than by using the definition of it given in #5. In this problem you may assume that

$$\frac{\partial}{\partial\theta}\int_{-\infty}^{\infty}\cdots\cdots\,dx=\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}\cdots\cdots\,dx.$$

(This is valid if the function being integrated is sufficiently well-behaved.)

(a) Show that $\int_{-\infty}^{\infty}\frac{\partial}{\partial\theta}f(x\mid\theta)\,dx=0$.

(b) Show that $\int_{-\infty}^{\infty}\frac{\partial^2}{\partial\theta^2}f(x\mid\theta)\,dx=0$.

(c) Show that

$$\frac{\partial^2}{\partial\theta^2}\log f(x\mid\theta)=\frac{\frac{\partial^2}{\partial\theta^2}f(x\mid\theta)}{f(x\mid\theta)}-\left(\frac{\partial}{\partial\theta}\log f(x\mid\theta)\right)^2.$$

Then take expected values of all three terms, use the result of part (b) to show that one of the expected values is 0, and finally draw the conclusion that

$$I(\theta)=-\,\mathbf{E}_\theta\left(\frac{\partial^2}{\partial\theta^2}\log f(X\mid\theta)\right).$$

CONTINUED⟶

8. Suppose $X_1, \ldots, X_n \sim \text{i. i. d.}\, N(\theta, 1^2)$. The result of #6 then says no function of $X_1, \ldots, X_n$ that is an unbiased estimator of $\theta$ can have smaller variance that something. Find and simplify what goes in the role of "something" in the previous sentence. You can use the result of #7(c), but you are not required to do so.

9. Suppose $X_1, \ldots, X_n \sim \text{i. i. d.}\, \text{Bin}(10, \theta)$. Answer the same question as in #8. (The results you use work just as well if the probability distributions involved are discrete.)

10. Suppose $X_1, \ldots, X_n \sim \text{i. i. d.}\, \text{Poisson}(\theta)$. Answer the same question as in #8.

11. Suppose $X_1, \ldots, X_n \sim$ i. i. d. have a "memoryless" (continuous) exponential distribution with expected value $\theta$, i.e., each $X_i$ is distributed as the (continuous) waiting time in a Poisson process with intensity $1/\theta$ occurrences per unit time. Answer the same question as in #8.

Now we move on from estimation to hypothesis testing, but $\Lambda(X)$ is still defined as above.

12. Consider testing the null hypothesis $H_0 : \theta = \theta_0$ against the alternative hypothesis $H_1 : \theta = \theta_1$.

    The likelihood-ratio test rejects the null hypothesis if and only if $\Lambda(X) > c$ (where $c =$ the "critical value"). Suppose this test's probability of Type I error is 3%, i.e., $\mathbf{Pr}(\text{this test rejects } H_0 \mid H_0) = 0.03$. The power of the test is $\mathbf{Pr}(\text{this test rejects } H_0 \mid H_1)$.

    Some other test, based on some statistic $K(X)$ different from $\Lambda(X)$, rejects the null hypothesis if and only if $K(X) > d$. Suppose this test's probability of Type I error is also 3%, i.e., $\mathbf{Pr}(\text{this test rejects } H_0 \mid H_0) = 0.03$. The power of the test is $\mathbf{Pr}(\text{this test rejects } H_0 \mid H_1)$.

    These are different tests. That means one test could reject $H_0$ while the other does not, with the same data, and the probability that the two tests disagree on whether to reject $H_0$ is more than 0.

$$
\begin{aligned}
\text{Let } A &= [\text{the likelihood-ratio test rejects } H_0 \text{ and the other test does not}], \\
B &= [\text{the other test rejects } H_0 \text{ and the likelihood-ratio test does not}], \\
\text{and } C &= [\text{both tests reject } H_0].
\end{aligned}
$$

(a) Explain and justify each of the seven relations marked by "?" below.

$$
\begin{aligned}
\mathbf{Pr}(A \mid H_1) &\overset{\boxed{?}\downarrow}{=} \int_A f(x \mid \theta = \theta_1)\, dx \overset{\boxed{?}\downarrow}{>} \int_A c\, f(x \mid \theta = \theta_0)\, dx \overset{\boxed{?}\downarrow}{=} c\, \mathbf{Pr}(A \mid H_0) \\
&\underset{\uparrow\boxed{?}}{=} c\, \mathbf{Pr}(B \mid H_0) \underset{\uparrow\boxed{?}}{=} \int_B c\, f(x \mid \theta = \theta_0)\, dx \underset{\uparrow\boxed{?}}{\geq} \int_B f(x \mid \theta = \theta_1)\, dx \underset{\uparrow\boxed{?}}{=} \mathbf{Pr}(B \mid H_1).
\end{aligned}
$$

(b) By thinking about $\mathsf{Pr}(A \text{ or } C \mid H_1)$ and $\mathsf{Pr}(B \text{ or C} \mid H_1)$, explain why the conclusion of part (a) can be summarized by saying that the likelihood-ratio test is better than the other test. (This is the **"Neyman-Pearson lemma."**)

HINT: The verbal statement

**"**$\big[$The likelihood-ratio test rejects $H_0$ and the other test does not$\big]$ **or** $\big[$both tests reject $H_0\big]$**"**

can be enormously simplified!