Homework 6
36-705
Due: Wednesday October 31st by Midnight.
Last Update: October 28th at 6.30pm. Modified Question 2.

Note: We will post the solutions at Midnight on Oct 31st to allow students to prepare for the exam. For this HW you will receive no credit for a late assignment.

1. [20 points] TV distance: In studying the NP classifier we introduced the TV distance. For the entire question to simplify things we will assume everything is defined on a discrete domain \(\Omega\). Suppose we have two distributions \(P\) and \(Q\) with pmfs \(p\) and \(q\). We defined the total variation distance between these distributions as:

\[
\text{TV}(P,Q) = \sum_{\{x \in \Omega : p(x) \geq q(x)\}} [p(x) - q(x)].
\]

This distance is closely related to the \(\ell_1\) distance between the two distributions. Show that,

\[
\text{TV}(P,Q) = \frac{1}{2} \sum_{x \in \Omega} |p(x) - q(x)|.
\]

The TV distance is a very strong notion of distance. In particular, if the TV is small then the probability of any event under the two distributions must be close. Show that,

\[
\text{TV}(P,Q) = \sup_{A \subseteq \Omega} |P(A) - Q(A)|.
\]

The Total Variation distance also belongs to a popular class of distances between probability distributions. It is an integral probability metric (other popular examples of IPMs include the Wasserstein distance). Show that, the TV distance can also be written as:

\[
\text{TV}(P,Q) = \frac{1}{2} \sup_{f : \|f\|_\infty \leq 1} |E_{X \sim P} f(X) - E_{Y \sim Q} f(Y)|
\]

2. [10 points] Exponential LRT: Suppose that, \(X_1, \ldots, X_n\) is drawn from an Exponential distribution with density:

\[
p(x|\theta) = \begin{cases} 
\exp(-(x-\theta)) & x \geq \theta, \\
0 & \text{otherwise}.
\end{cases}
\]

Construct the (generalized) LRT statistic for distinguishing the hypotheses:

\[
H_0 : \theta \leq \theta_0, \\
H_1 : \theta > \theta_0,
\]

for some fixed \(\theta_0\).
3. [20 points]

(a) Let \( X_1, \ldots, X_n \sim N(\mu, \sigma^2) \). Construct the likelihood ratio test for

\[
H_0 : \sigma = \sigma_0, \mu \text{ unknown} \\
H_1 : \sigma \neq \sigma_0, \mu \text{ unknown}.
\]

Compare to the Wald test.

(b) Let \( X \sim \text{Bin}(n, p) \). Construct the likelihood ratio test for

\[
H_0 : p = p_0 \\
H_1 : p \neq p_0.
\]

Compare to the Wald test.

4. [20 points] p-values : Throughout this question we fix an arbitrary test statistic. Most natural hypothesis tests have (strictly) nested rejection regions. What this means is that if we denote the rejection region as \( R_\alpha \) then:

\[
R_\alpha \subset R_{\alpha'} \quad \text{if} \quad \alpha < \alpha'.
\]

It is in this setting that p-values make the most sense. We will assume that our test statistic has nested rejection regions. Now, consider the general hypothesis testing problem:

\[
H_0 : \theta \in \Theta_0 \\
H_1 : \theta \notin \Theta_0.
\]

(a) Show that if the test is valid, i.e.,

\[
\sup_{\theta \in \Theta_0} \mathbb{P}_\theta(\{X_1, \ldots, X_n\} \in R_\alpha) \leq \alpha \quad \text{for all} \quad 0 \leq \alpha \leq 1,
\]

then for any \( \theta \in \Theta_0 \) we have that the distribution of the p-value satisfies:

\[
\mathbb{P}_\theta(p \leq u) \leq u \quad \text{for all} \quad 0 \leq u \leq 1.
\]

This means that the p-value is “stochastically dominated” by the uniform distribution.

(b) Show that if there is a \( \theta \in \Theta_0 \) such that,

\[
\mathbb{P}_\theta(\{X_1, \ldots, X_n\} \in R_\alpha) = \alpha \quad \text{for all} \quad 0 \leq \alpha \leq 1,
\]

then

\[
\mathbb{P}_\theta(p \leq u) = u \quad \text{for all} \quad 0 \leq u \leq 1.
\]

This shows that in cases when the null is simple (or if there is a “worst-case” null distribution) then the p-value has a uniform distribution.
5. **[10 points] More LRT:** Suppose that, $X_1, \ldots, X_n$ is drawn from a Pareto distribution with density:

$$p(x|\theta, \nu) = \frac{\theta \nu^\theta}{x^{\theta+1}} I_{[\nu, \infty)}(x), \quad \theta > 0, \nu > 0.$$ 

Construct the (generalized) LRT statistic for distinguishing the hypotheses:

$$H_0 : \theta = 1, \nu \text{ unknown},$$

$$H_1 : \theta \neq 1, \nu \text{ unknown}.$$ 

Suppose Wilk's approximation was accurate. Derive the critical region for the GLRT.

6. **[20 points] Minimax Rate Lower Bounds:** We will derive something that is popularly known as Le Cam’s (two-point) lower bound. Minimax rate lower bounds for point estimation are usually proved via a reduction to testing. Suppose we have a collection of distributions $P_\theta = \{p_\theta : \theta \in \Theta\}$, and we are interested in proving a lower bound on the minimax rate for estimating $\theta$. Concretely, for some loss function $\ell$, which we will assume is symmetric and satisfies the triangle inequality, we want to show something of the form:

$$\inf_{\hat{\theta}(X_1, \ldots, X_n)} \sup_{\theta \in \Theta} \mathbb{E}[\ell(\hat{\theta}, \theta)] \geq \ldots.$$ 

(a) Fix an estimator $\hat{\theta}(X_1, \ldots, X_n)$. First show that for any $\delta > 0$,

$$\mathbb{E}[\ell(\hat{\theta}, \theta)] \geq \delta \mathbb{P}(\ell(\hat{\theta}, \theta) \geq \delta).$$ 

(b) Now, select $\theta_1, \theta_2 \in \Theta$, such that $\ell(\theta_1, \theta_2) > 2\delta$. From our estimator $\hat{\theta}$ we are going to create a test $\phi(\hat{\theta})$ for distinguishing:

$$H_0 : \theta = \theta_1$$

$$H_1 : \theta = \theta_2.$$ 

In particular,

$$\phi(\hat{\theta}) = \begin{cases} 
\theta_1 & \text{if } \ell(\hat{\theta}, \theta_1) < \ell(\hat{\theta}, \theta_2), \\
\theta_2 & \text{otherwise}.
\end{cases}$$

Show that,

$$\sup_{\theta \in \Theta} \delta \mathbb{P}(\ell(\hat{\theta}, \theta) \geq \delta) \geq \frac{\delta}{2} \left[ \mathbb{P}_{\theta_1}(\phi(\hat{\theta}) \neq \theta_1) + \mathbb{P}_{\theta_2}(\phi(\hat{\theta}) \neq \theta_2) \right].$$

(c) Now, show that:

$$\inf_{\hat{\theta}(X_1, \ldots, X_n)} \sup_{\theta \in \Theta} \mathbb{E}[\ell(\hat{\theta}, \theta)] \geq \frac{\delta}{2} \inf_{\phi} \left[ \mathbb{P}_{\theta_1}(\phi(X_1, \ldots, X_n) \neq \theta_1) + \mathbb{P}_{\theta_2}(\phi(X_1, \ldots, X_n) \neq \theta_2) \right],$$

where $\phi$ is a test of the form above for distinguishing the hypotheses $H_0$ and $H_1$. 

3
(d) Consider the (optimal) NP classifier for distinguishing these hypotheses. Show that this classifier has error \(1 - TV(p^n_{\theta_1}, p^n_{\theta_2})\), where \(p^n_{\theta}\) denotes the \(n\)-fold product distribution (i.e. the distribution of \(n\) i.i.d. samples). Using this show the following minimax lower bound, for any \(\theta_1, \theta_2\) such that \(\ell(\theta_1, \theta_2) > 2\delta\),

\[
\inf_{\hat{\theta}(X_1, \ldots, X_n)} \sup_{\theta \in \Theta} \mathbb{E}[\ell(\hat{\theta}, \theta)] \geq \frac{\delta}{2} \left[1 - TV(p^n_{\theta_1}, p^n_{\theta_2})\right].
\]

(e) Explain the bound intuitively, i.e. how should we try to select \(\theta_1, \theta_2\) to obtain a strong lower bound.

7. [20 points] **Extra Credit: The \(\chi^2\) test:** In lecture we claimed that the Pearson goodness-of-fit statistic for testing a multinomial \(p_0\),

\[
T = \sum_{i=1}^{d} \frac{(X_i - np_{0i})^2}{np_{0i}},
\]

has a \(\chi^2\) distribution with \(d - 1\) degrees of freedom (we actually subtracted the mean of this statistic to center the test statistic to have mean 0 under the null but this is not important). We will assume throughout that \(d\) is fixed, and that \(p_{0i} > 0\) for each category. Here \(X_1, \ldots, X_d\) are the counts for the \(d\) categories, so we have that:

\[
\sum_{i=1}^{d} X_i = n \quad \text{and} \quad \sum_{i=1}^{d} p_{0i} = 1.
\]

(a) As a warmup: Suppose that there are only two categories, i.e. the multinomial is \((p_{01}, p_{02})\). Show that the test statistic can be equivalently written as:

\[
T = \frac{(X_1 - np_{01})^2}{np_{01}(1 - p_{01})}.
\]

Show that \(\frac{(X_1 - np_{01})}{\sqrt{np_{01}(1 - p_{01})}}\) under the null has an asymptotically standard normal distribution, and conclude that in this case the Pearson statistic has a \(\chi^2_1\) distribution.

(b) The general case roughly follows the same ideas but is more involved. I will outline the steps to help you but this is only one way to get to the answer (and ideally you should try to solve it without looking through these steps first).

- The first thing to note is that our representation is redundant in the sense that the sum of the counts and the sum of the probabilities is constrained, so we will eliminate \(p_{0d}\), and denote the vector \(\tilde{p} = (p_{01}, \ldots, p_{0(d-1)})\). Now, using
the usual MLE asymptotics (you can look this up in the Wasserman book), you can conclude that under the null,

\[
\sqrt{n} \left( \begin{bmatrix} X_1/n \\ X_2/n \\ \vdots \\ X_{d-1}/n \end{bmatrix} - \tilde{p} \right) \xrightarrow{d} N \left( 0, \begin{bmatrix} p_{01}(1 - p_{01}) & -p_{01}p_{02} & \cdots & -p_{01}p_{0(d-1)} \\ -p_{01}p_{02} & p_{02}(1 - p_{02}) & \cdots & -p_{02}p_{0(d-1)} \\ \vdots & \vdots & \ddots & \vdots \\ -p_{01}p_{0(d-1)} & -p_{02}p_{0(d-1)} & \cdots & p_{0(d-1)}(1 - p_{0(d-1)}) \end{bmatrix} \right) \]

Convince yourself that the covariance matrix is non-degenerate (this is why we eliminated one of the categories).

• Show that the inverse of the covariance matrix is given by:

\[
I(p) = \begin{bmatrix} \frac{1}{p_{01}} + \frac{1}{p_{0d}} & \frac{1}{p_{02}} & \cdots & \frac{1}{p_{0d}} \\ \frac{1}{p_{02}} & \frac{1}{p_{0d}} + \frac{1}{p_{0d}} & \cdots & \frac{1}{p_{0d}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{1}{p_{0(d-1)}} & \frac{1}{p_{0d}} & \cdots & \frac{1}{p_{0(d-1)}} + \frac{1}{p_{0d}} \end{bmatrix} 
\]

Conclude that,

\[
n \left( \begin{bmatrix} X_1/n \\ X_2/n \\ \vdots \\ X_{d-1}/n \end{bmatrix} - \tilde{p} \right)^T I(p) \left( \begin{bmatrix} X_1/n \\ X_2/n \\ \vdots \\ X_{d-1}/n \end{bmatrix} - \tilde{p} \right) \xrightarrow{d} \chi^2_{d-1}.
\]

• Show that the above statistic is equal to the Pearson \( \chi^2 \) statistic.