Homework 9: Hypothesis Testing
Due Thursday, November 14, 2019

This homework will cement your understanding of the material in Rice Chapters 9 and 11. The assignment is due before the end of class on the date listed above. You may discuss problems with your classmates, but you must turn in your own writeup, and list all students with whom you discussed the assignment. You must show all your work in order to receive full credit. Stop by office hours if you need additional help or hints.

Problem 1 (Testing and p-values). This problem includes a few questions to solidify your conceptual understanding of hypothesis testing, test statistics and p-values.

(a) Suppose that we perform a hypothesis test at level $\alpha$. Explain (a sentence or two is fine) why the $p$-value associated with this test must be less than or equal to $\alpha$ in order to reject.

(b) Suppose that Alice and Bob use the same test statistic $T$, evaluated on the same data, but Alice chooses to reject whenever $T < t_A$ and Bob rejects whenever $T < t_B$, where $t_A < t_B$. $T$ has the property that smaller $T$ constitutes evidence against the null.

(i) Which test has larger size, Alice’s or Bob’s?

(ii) Which test has larger power, Alice’s or Bob’s?

(c) In the likelihood ratio test, explain (a sentence or two is fine) why lowering the threshold for rejection ($c$ in lecture) corresponds to a smaller significance level.

(d) A common misconception about testing is that the $p$-value is the probability that the null hypothesis is true. Explain (a sentence or two is fine) why this is not the case.

Problem 2 (Proving the Neyman-Pearson Lemma). In lecture, we stated the Neyman-Pearson lemma without proof. This problem will walk you through the proof. In what follows, let $X$ be a random variable denoting the data. Thus, we can think of $X = (X_1, X_2, \ldots, X_n)$, where $X_1, X_2, \ldots, X_n$ denote n i.i.d. random observations, but we stress that this object $X$ could be something much more complicated and it need not consist of independent observations. For example, in my own research, I spend a lot of time working with brain imaging data, so we might think of $X$ as denoting the 3-dimensional reconstructed image of a patient’s brain.

(a) A hypothesis test is a procedure whereby, given data, we decide whether or not to reject the null hypothesis. Thus, we can think of a hypothesis test as a function $d(x)$ with $d(x) = 1$ if our test rejects the null hypothesis when we observe data $x$, and $d(x) = 0$ otherwise. Thus, $d(X)$ is a Bernoulli random variable. Let $E_0$ and $E_A$ denote expectation when the null hypothesis is true and the alternative is true, respectively. Show that the level of the test is given by $E_0 d(X)$, and the power of the test is given by $E_A d(X)$. 
(b) Now, let \( d_{LR}(x) \) denote the decision of the likelihood ratio test, so that \( d_{LR}(x) = 1 \) if the LRT rejects \( H_0 \) when presented with data \( x \). Let \( f_0 \) and \( f_A \) denote the probability (mass or density) under the null and alternative, respectively. Explain why \( d_{LR}(x) = 1 \) if and only if \( cf_A(x) > f_0(x) \), where \( c \) is the threshold of rejection for the LRT.

(c) Let \( d(x) \) denote any other test with significance level at most that of the LRT. The main claim of the Neyman-Pearson Lemma is that the power of \( d(x) \) (or any other test with significance level at most that of our LRT) must be at most that of the LRT. Thus, we need to show that \( \mathbb{E}_A d(X) \leq \mathbb{E}_A d_{LR}(X) \). Our main tool for showing this is that for all \( x \),

\[
d(x) [cf_A(x) - f_0(x)] \leq d_{LR}(x) [cf_A(x) - f_0(x)].
\]  

(1)

Use the previous part to show why Equation (1) is true for all \( x \). Hint: consider separately the cases where \( d_{LR}(x) = 1 \) and \( d_{LR}(x) = 0 \).

(d) Summing or integrating over all \( x \) (depending on whether it is discrete or continuous data) and rearranging slightly, Equation (1) yields

\[
\mathbb{E}_0 d_{LR}(X) - \mathbb{E}_0 d(X) \leq c [\mathbb{E}_A d_{LR}(X) - \mathbb{E}_A d(X)].
\]

The right-hand side of this equation is the power of the LRT minus the power of the other test (and then multiplied by \( c \)). We have \( \mathbb{E}_A d_{LR}(X) \geq \mathbb{E}_A d(X) \) because the left-hand side of this equation is nonnegative. Why is this the case? That is, explain why \( \mathbb{E}_0 d_{LR}(X) - \mathbb{E}_0 d(X) \geq 0 \). Hint: go back over our assumptions.

(e) One issue we might be concerned with is the case where \( c = 0 \). Suppose that \( f_0(x) > 0 \) and \( f_A(x) > 0 \) for all \( x \). What is the significance level of the LRT when \( c = 0 \)? What is the power of the LRT when \( c = 0 \)?

**Problem 3** (Likelihood ratio for the multinomial). This problem is designed to walk you through the ideas developed in Rice Section 9.5. I recommend at least skimming that section before proceeding with this problem. In many settings, the outcome of an experiment is one of a few possible categorical outcomes. For example, in Gregor Mendel’s famous plant breeding experiments discussed in Example C of Rice Section 9.5, each pea plant could be one of four possible phenotypes. In survey data, respondents are often made to choose among a fixed set of answers to a question. Repeating the experiment, say, \( n \) times, we obtain counts of how many experiments yielded each outcome. If there are \( m \) possible outcomes, then our data takes the form of counts \( x_1, x_2, \ldots, x_m \) such that \( \sum_{i=1}^{m} x_i = n \). A common model for categorical outcomes such as this is the multinomial distribution, which has probability mass function given by

\[
f(x_1, x_2, \ldots, x_m; p) = \begin{cases} 
\frac{n!}{x_1!x_2!\cdots x_m!} p_1^{x_1} p_2^{x_2} \cdots p_m^{x_m} & \text{if } \sum_{i} x_i = n \\
0 & \text{otherwise,}
\end{cases}
\]

where \( p \in \mathbb{R}^m \) is a probability vector, so that \( p_i \geq 0 \) for \( i = 1, 2, \ldots, m \) and \( \sum_{i} p_i = 1 \).

Suppose that we have a parameter \( \theta \in \Theta = \omega_0 \cup \omega_A \), where \( \omega_0 \cap \omega_A = \emptyset \), and each \( \theta \in \Theta \) has a corresponding probability vector \( p(\theta) \), satisfying \( \sum_{i} p_i(\theta) = 1 \) and \( p_i(\theta) \geq 0 \) for all \( i = 1, 2, \ldots, m \).

\[1\text{https://en.wikipedia.org/wiki/Gregor_Mendel#Experiments_on_plant_hybridization}
(a) Let $H_0 : \theta \in \omega_0$ and $H_A : \theta \in \omega_A$. Write down an expression for the likelihood ratio $\Lambda$ (note that this is the LR in which the denominator is a maximum over $\omega_0 \cup \omega_A$).

(b) The MLE for $p$ if no restriction is placed on $p$ (i.e., if $p$ is allowed to be any probability vector) is $\hat{p}_i = x_i/n$ for $i = 1, 2, \ldots, m$. This can be proved using Lagrange multipliers, but you may take this fact as given, here. Suppose that for every probability vector $q$, there exists a $\theta \in \Theta$ such that $q = p(\theta)$. Use this assumption to show why the denominator in the likelihood ratio $\Lambda$ is given by

$$\frac{n!}{x_1!x_2! \cdots x_m! \hat{p}_1^{x_1} \hat{p}_2^{x_2} \cdots \hat{p}_m^{x_m}}.$$

(c) Let $\hat{\theta} \in \omega_0$ be such that

$$f(x_1, x_2, \ldots, x_m; p(\hat{\theta})) = \max_{\theta \in \omega_0} f(x_1, x_2, \ldots, x_m; p(\theta)).$$

Show that

$$\Lambda = \prod_{i=1}^{m} \left( \frac{p_i(\hat{\theta})}{\hat{p}_i} \right)^{x_i}.$$

**Hint:** the factorial terms should cancel out in the likelihood ratio.

(d) Use the preceding part and the definition of $\hat{p}_i$ to show that

$$\log \Lambda = n \sum_{i=1}^{m} \hat{p}_i \log \frac{p_i(\hat{\theta})}{\hat{p}_i}.$$

(e) Wilk’s Theorem (Rice Section 9.4, Theorem A), states that for suitably smooth probability densities or frequency functions, $-2 \log \Lambda$ converges in distribution to a $\chi^2$ distribution with degrees of freedom equal to the dimension of $\omega_0 \cup \omega_A$ minus the dimension of $\omega_0$. Argue why $\omega_0 \cup \omega_A$ has dimension $m - 1$. **Hint:** the vector $p$ has $m$ dimensions, but it is constrained in a certain way.

(f) Let $x_0 > 0$ be fixed and consider the function

$$f(x) = x \log \frac{x}{x_0}.$$

The Taylor series for $f(x)$ about $x = x_0$ is

$$f(x) = (x - x_0) + \frac{(x - x_0)^2}{2x_0} + \cdots.$$

Under $H_0$, if $n$ is large, we have $\hat{p}_i \approx p_i(\hat{\theta})$. Use this to show that when $H_0$ is true and $n$ is large,

$$-2 \log \Lambda \approx \sum_{i=1}^{m} \left( \frac{x_i - np_i(\hat{\theta})}{np_i(\hat{\theta})} \right)^2.$$

If you have used Pearson’s chi-square statistic before, this quantity should look familiar! We’ve just shown that the likelihood ratio test and Pearson’s chi-square test are (approximately) equivalent!
Problem 4 (Wilcoxon-Mann-Whitney). As mentioned in lecture, the Wilcoxon-Mann-Whitney (WMW) test is a two-sample test, but one that makes no assumptions about the distribution of the underlying data. This is in contrast to, e.g., the $t$-test, which assumes that the data are normally distributed. Since such distributional assumptions are rarely true, we often prefer nonparametric tests like the WMW test, i.e., tests that do not make such strong assumptions. This problem follows the presentation in Rice Section 11.2.3, so you may wish to read that section before proceeding.

The Wilcoxon-Mann-Whitney test assumes that we have $m + n$ total independent observations, $n$ from the control group and $m$ that are given treatment. For each unit in the control group, we obtain an observation, say $X_1, X_2, \ldots, X_n$, drawn i.i.d. from some distribution $F$. Similarly, let $Y_1, Y_2, \ldots, Y_m$ denote the observations from the treatment group, drawn i.i.d. from distribution $G$, possibly different from $F$. A key assumption of the WMW test is that in addition to the independence of the $X_i$ and the $Y_j$, the collections $\{X_1, X_2, \ldots, X_n\}$ and $\{Y_1, Y_2, \ldots, Y_m\}$ are independent of one another.

(a) The null hypothesis of the WMW test is that the treatment has no effect. Explain why (a sentence or two is fine), in that case, we can think of $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m$ as an i.i.d. sample of size $m + n$ from $F$.

(b) The WMW test begins by taking all $m+n$ observations and ranking them by increasing order (for the sake of simplicity, we will assume throughout this problem that there are no ties, i.e., $X_1, X_2, \ldots, X_n, Y_1, Y_2, \ldots, Y_m$ are distinct). Show why, under the null hypothesis $H_0 : F = G$, 

$$
\Pr \left[ X_1 = z_1, X_2 = z_2, \ldots, X_n = z_n, Y_1 = z_{n+1}, Y_2 = z_{n+2}, \ldots, Y_m = z_{n+m} \right]
$$

$$
= \Pr \left[ X_1 = z_{\pi(1)}, X_2 = z_{\pi(2)}, \ldots, X_n = z_{\pi(n)}, Y_1 = z_{\pi(n+1)}, Y_2 = z_{\pi(n+2)}, \ldots, Y_m = z_{\pi(n+m)} \right]
$$

for all permutations $\pi : \{1, 2, \ldots, m+n\} \rightarrow \{1, 2, \ldots, m+n\}$, and all $(z_1, z_2, \ldots, z_{n+m})$. Explain (in math, if you can, using words, if not), why this implies that under the null hypothesis, all assignments of rankings to units are equally likely.

(c) The test statistic in the WMW test is the sum of the ranks of the treatment units. Using the preceding results, explain why, under the null, the ranks of the treatment units (i.e., the ranks of $Y_1, Y_2, \ldots, Y_m$) are a simple random sample without replacement of size $m$ from the population $1, 2, \ldots, m + n$.

(d) Consider the finite population $1, 2, \ldots, m + n$. Use the identities 

$$
\sum_{k=1}^{N} k = \frac{N(N+1)}{2}
$$

and 

$$
\sum_{k=1}^{N} k^2 = \frac{N(N+1)(2N+1)}{6}
$$

to show that the population mean and variance are

$$
\mu = \frac{m+n+1}{2} \quad \text{and} \quad \sigma^2 = \frac{m^2 + n^2 + 2mn - 1}{12}.
$$

(e) Let $R_Y$ denote the sum of the ranks of the treatment group, i.e., the WMW test statistic. Using the previous two parts, show that 

$$
\mathbb{E} R_Y = \frac{m(m+n+1)}{2} \quad \text{and} \quad \text{Var} R_Y = \frac{mn(m+n+1)}{12}.
$$

(f) The WMW test rejects the null if $R_Y$ is much smaller or much larger than $\mathbb{E} R_Y$. Explain why (a sentence or two is fine) when $H_0$ is not true (i.e., $F$ and $G$ are different), we might expect $R_Y$ to be much bigger or much smaller than $\mathbb{E} R_Y$. 
