Data 102 (Fall 2022) **Names:**
Discussion Worksheet 2
09/02/2022

---

1. In the following table, rows represent reality and columns represent the decision:

	0	1
0	4	7
1	3	2

Compute the:

(a) FPR (false positive rate) **Solution:** $FPR = \frac{FP}{FP+TN}$, or $\frac{7}{4+7} = \frac{7}{11}$.

(b) TPR (true positive rate) **Solution:** $TPR = \frac{TP}{TP+FN}$, or $\frac{2}{2+3} = \frac{2}{5}$.

(c) FNR (false negative rate) **Solution:** $FNR = \frac{FN}{TP+FN}$, or $\frac{3}{2+3} = \frac{3}{5}$.

(d) FDP (false discovery proportion) **Solution:** Unlike the previous three, this is a column-wise rather than row-wise rate. We have $FDP = \frac{FP}{FP+TP}$, or $\frac{7}{7+2} = \frac{7}{9}$.

Is the classifier represented by this table a good or a bad classifier? Why?

Solution: It is a bad classifier, because flipping the decision (all 0s become 1s and vice versa) would improve every single metric. In other words, it is worse than chance.

2. Conceptual questions about p-values:

(a) Circle the correct bold-faced word: Small p-values give evidence that the **null**/**alternative** hypothesis is true. **Solution:** Alternative.

(b) Complete the statement: Under the null hypothesis, the distribution of p-values is ______ . **Solution:** Uniform [0,1].

(c) If the null hypothesis is true, what is the probability of obtaining a p-value that is less than or equal to 0.01? **Solution:** 0.01.

3. Consider a test for a disease, where the test has a false positive rate of 1% and a true positive rate of 80%.

(a) Suppose the proportion of people that have the disease is 5%. Conditional on a positive test, how likely is a person to actually have the disease?

Solution: Algebraically, we can apply Bayes' rule. But the easiest way to visualize this is to imagine 100 people, 5 of whom have the disease and 95 of whom don't. Now if all of them take the test, 4 of the people with the disease will be positive and 0.95 of those who don't will test positive. So the probability of having the disease is $\frac{4}{4+0.95} = \frac{80}{99}$.

(b) What is the name of the conditional probability in the previous question?

Solution: True discovery rate.

(c) Suppose instead that the proportion of people with the disease is 0.1%. Now how likely is a person to have the disease, conditional on a positive test?

Solution: We can apply the same visualization as before. Imagine 1000 people, 1 with the disease and 999 without. Then in expectation 0.8 of the ones with the disease test positive, and 9.99 without the disease test positive. So the probability is $\frac{0.8}{0.8+9.99} = \frac{80}{1079}$.

4. For a real-valued random variable X, consider the following two hypothesis:

$$H_0 : X \sim \text{Uniform}([0, 1]), \quad H_1 : X \sim \text{Uniform}([0.5, 1.5]). \tag{1}$$

We would like a test with a false positive rate of at most 0.05. What is the best possible true positive rate?

Solution: The best we can do is reject the null when $X > 0.95$, which has a false positive rate of exactly 0.05. Then under H_1, we correctly output 1 when $X \in [0.95, 1.5]$ and incorrectly output 0 when $X \in [0.5, 0.95]$. So the true positive rate is 0.55.

Note interestingly that Neyman-Pearson does not work here, because the likelihood ratio is not a continuous function.

5. **Multiple Hypothesis Testing with the Benjamini-Hochberg Procedure**

In this question we analyze the properties of the Benjamini-Hochberg (BH) procedure. Recall the steps of the procedure:

Algorithm 1 The Benjamini-Hochberg Procedure

input: FDR level α, set of n p-values $P_1, \ldots, P_n$

Sort the p-values $P_1, \ldots, P_n$ in non-decreasing order $P_{(1)} \leq P_{(2)} \leq \cdots \leq P_{(n)}$

Find $K = \max\{i \in \{1, \ldots, n\} : P_{(i)} \leq \frac{\alpha}{n} i\}$

Reject the null hypotheses (declare discoveries) corresponding to $P_{(1)}, \ldots, P_{(K)}$

(a) We have 10 p-values for multiple hypothesis testing: 0.001, 0.003, 0.012, 0.015, 0.08,0.09, 0.1, 0.14, 0.16, 0.28. Suppose we would like to control the FDR at the level 0.05. How many tests does the BH procedure reject?

Solution: By first sorting the p-values and comparing the k-th p-value with $k \cdot \alpha/n = 0.005k$, we will see that the largest k such that its p-value is smaller than $0.005k$ is $k = 4$. So 4 tests will be rejected.

(b) Suppose $P_1 = P_2 = \cdots = P_n = \alpha$, and we run BH under level α on these p-values. How many discoveries does BH make? Explain.

Solution: It makes n discoveries, because the highest p-value (equal to α) is less than or equal to $\frac{\alpha}{n} n = \alpha$.

(c) Suppose $P_1 = P_2 = \cdots = P_{n-1} = \alpha, P_n = \alpha + 0.001\alpha$, and we run BH under level α on these p-values. How many discoveries does BH make? Explain.

Solution: It makes 0 discoveries, because no p-value is under the corresponding threshold $\frac{\alpha}{n} k$.

6. **Challenge question.** Consider the probability density function $f_\theta(x) = \theta x^{\theta-1}$ where $0 < x < 1$. We wish to design a test to discern between the null hypothesis that $\theta = 3$, and the alternative hypothesis that $\theta = 4$.

   (a) Derive the most powerful test that has significance level less than α.

   Solution: Leveraging the Neyman-Pearson Lemma, we design a likelihood-ratio test. The likelihood ratio has the form:

$$\frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} = \frac{4x^3}{3x^2} = \frac{4x}{3}.$$

   Now we need to solve for η such that the significance level is α, or

$$\mathbb{P}\Big(\frac{4x}{3} > \eta \,|\, H_0\Big) = \alpha \implies \mathbb{P}\Big(x > \frac{3\eta}{4} \,|\, H_0\Big) = \alpha.$$

   That is, we need

$$\int_{\frac{3\eta}{4}}^{1} f_{\theta_0}(x)dx = \int_{\frac{3\eta}{4}}^{1} 3x^2 dx = \alpha.$$

   Solving for this gives $1 - \left(\frac{3\eta}{4}\right)^3 = \alpha$, which yields $\eta = \frac{4}{3}(1-\alpha)^{1/3}$. Therefore, the most powerful likelihood ratio test is given by:

$$\delta(x) = \begin{cases} \text{Reject Null} & : \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} > \frac{4}{3}(1-\alpha)^{1/3} \\ \text{Accept Null} & : \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} \leq \frac{4}{3}(1-\alpha)^{1/3} \end{cases}$$

   (b) What is the power of the test, i.e. its true positive rate?

   Solution: Using the solution to (a), we need to calculate $\mathbb{P}(x > \frac{4}{3}(1-\alpha)^{1/3} \,|\, H_1)$. More explicitly,

$$\begin{aligned} \mathbb{P}\left(\frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} > \eta \mid H_1\right) &= \mathbb{P}\left(\frac{4x}{3} > \frac{4}{3}(1-\alpha)^{1/3} \,|\, H_1\right) \\ &= \mathbb{P}(x > (1-\alpha)^{1/3} \,|\, H_1) \\ &= \int_{(1-\alpha)^{1/3}}^{1} 4x^3 dx = 1 - (1-\alpha)^{4/3} \end{aligned}$$

   (c) Now suppose that there were two possible alternatives: either $\theta = 2$ or $\theta = 4$. You would like to design a test with significance level α, which has good power under both alternatives. How would you go about doing this? (*Note:* This question is somewhat open-ended: there may be multiple good strategies.)