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)

   (b) TPR (true positive rate)

   (c) FNR (false negative rate)

   (d) FDP (false discovery proportion)

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

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.

   (b) Complete the statement: Under the null hypothesis, the distribution of p-values is ______.

   (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?

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?

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

   (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?

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?

   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 $P_{(1)} \le P_{(2)} \le \cdots \le P_{(n)}$

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

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?

   (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.

   (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.

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.

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

   (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.)