DS 102 Discussion 2

Wednesday, February 2, 2022

1. **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?

(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.

(d) Suppose we run BH on $\{P_1, \ldots, P_n\}$, and we make $R < n$ discoveries. Now suppose we add an extra p-value equal to 0 to this set. Now we run BH on $\{P_1, \ldots, P_n, 0\}$ and get a new number of rejections R'. Which of the following are possible: $R' > R$, $R' = R$, $R' < R$? If multiple are possible, list all that are possible. Explain why.

(e) (Optional) Suppose we run BH on $P_1, \ldots, P_n$, and we make $R < n$ discoveries. Now suppose we add an extra p-value equal to the maximum p-value in this set; denote that p-value $P_{\max}$. Now we run BH on $P_1, \ldots, P_n, P_{\max}$ and get a new number of rejections R'. Which of the following are possible: $R' > R$, $R' = R$, $R < R$?

2. **Hypothesis Testing**

One can imagine different metrics for quantifying how "good" a decision is. For example, we would like our decisions to have both high true positive rate and low false positive rate. Our goal as statisticians is to develop reasonable strategies for doing well on both metrics. In other words, how should we pick a point on the ROC curve? Once we pick a point, how do we achieve it?

The Neyman-Pearson Lemma offers one solution, in the case of hypothesis testing. We call the probability of a false positive under null hypothesis H_0 the *significance level* α of a test and we call the probability of a true positive under the alternative hypothesis H_1 the *power* of a test.

The Neyman-Pearson formulation makes the following recommendation: fix a false positive rate you are willing to tolerate, then pick the point on the ROC curve which maximizes power. The Neyman-Pearson Lemma prescribes how to achieve this point:

Lemma (Neyman & Pearson, 1933) *Suppose* $\theta_1 > \theta_0$*. For any significance level* $\alpha \in [0, 1]$*, the following* likelihood-ratio test *maximizes power among all tests with level at most* α*:*

$$\delta(x) = \begin{cases} \textit{Reject Null } (1) & : \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} > \eta \\ \textit{Accept Null } (0) & : \frac{f_{\theta_1}(x)}{f_{\theta_0}(x)} \leq \eta \end{cases}$$

where $f_{\theta_0}, f_{\theta_1}$ *are the likelihoods under the null and alternative distributions, respectively, and* η *is the real value such that* $\mathbb{P}(\delta(x) = 1 \mid H_0) = \alpha$*.*

Let's try an example where you will apply the Neyman-Pearson Lemma. Suppose that you have a sample from a distribution with probability density function $f_\theta(x) = \theta x^{\theta-1}$ where $0 < x < 1$. You would like to design a test to discern between the null hypothesis that $\theta = 3$, and the alternative hypothesis that $\theta = 4$.

(a) *Deriving a Likelihood Ratio Test*

Derive the most powerful test for this problem such that the significance level is less than α.

(b) *Power of the Likelihood Ratio Test*

What is the power of the test, $\mathbb{P}(\delta(x) = 1 \mid H_1)$?

3. **ROC Curves**

In lecture we discussed ROC curves, or "receiver operating characteristic" curves. ROC curves plot the true positive rate (TPR) and false positive rates (FPR) for a binary classifier at different decision thresholds. Recall that the TPR and FPR are defined as:

$$\text{TPR} = \frac{\text{TP}}{\text{FN} + \text{TP}}, \quad \text{FPR} = \frac{\text{FP}}{\text{TN} + \text{FP}},$$

where TP, TN, FP, and FN correspond to the entries of the 2×2 table for binary decisions.

In this exercise, we will examine the ROC curve for a toy dataset. Let Y be the label, X_1, X_2 be features, and consider the model function $f(X_1, X_2) = 3X_1 + 2X_2 + 1$.

Table 1: Example dataset

Y	$f(X_1, X_2)$	X_1	X_2
0	-1	-1	0.5
1	-0.5	-1	0.75
0	0	-1	1
1	1	0.2	-0.3
1	0.25	-0.25	0
0	0.25	-0.05	-0.3

(a) *Drawing a ROC Curve*

Plot the ROC curve for the model $f(X_1, X_2)$ with respect to the label Y.

(b) *Feasible Points on the ROC Curve*

Is it possible to choose two decision thresholds α_1, α_2 and probabilities of using each decision threshold such that the expected true positive rate is $\frac{1}{3}$, and the expected false positive rate is $\frac{2}{3}$?