Modern Statistics 52311, 2018-19
Homework 2 - Multiple Hypothesis Testing

March 26, 2019

The grade for entire exercise is the sum of the 4 highest scores for individual questions - a complete and correct solution for each question is worth 25% of this exercise grade. To get full credit for a question, you need to solve it correctly and completely, and explain your answer clearly. For the computerized questions, you need to supply plots, explanations and your code.

You may submit in groups of size two or less. Submit your solutions by 30/4/2019.

Problems

1. Suppose that we have m different random variables $X^{(1)}, .., X^{(m)}$ with a joint distribution $(X^{(1)}, .., X^{(m)}) \sim F$, and possibly different marginal distributions $X^{(i)} \sim F_i$. We observe n samples drawn from from the joint distribution F and want to test dependency between all pairs $(X^{(i)}, X^{(j)})$ using the same test statistic T, applied to the data $(x_1^{(i)}, x_1^{(j)}), .., (x_n^{(i)}, x_n^{(j)})$. For each pair, we perform a permutation test by computing a test statistic T on the data on N permuted versions of the data. We keep the $FWER$ over all pairs at level α using Bonferonni correction, and set the number of permutations N to be the minimal number which still allows us to rejection the null hypothesis with positive probability under $FWER \leq \alpha$. Suppose that each computation of the test statistic T on a sample of length n (for two r.v.s. $X^{(i)}, X^{(j)}$) requires $g(n)$ basic operations for some function g (we neglect other computations required, such as comparing a test statistic to a threshold).

   (a) Describe an algorithm for testing all dependencies of all pairs and express the total number of basic operations required to perform all tests as a function of m, n, α and g for a general test statistic T. Try to find the most efficient algorithm you can.

   (b) Describe an algorithm for testing all dependencies of all pairs and express the total number of basic operations required to perform all tests as a function of m, n, α and

g for a **distribution-free** test statistic T. Try to find the most efficient algorithm you can.

2. In this question we complete the proof showing that the Benjamini-Hochberg procedure with parameter α controls the FDR at level α for m independent test statistics. Let $\mathcal{R}_{BH}(P_1,..,P_m;\alpha) \subset \{1,..,m\}$ be the set of rejected hypotheses when using the BH procedure with parameter α on the p-values $P_1,..,P_m$. Let $C_k^{(i)}$ be the events defined as in the lectures for $i,k = 1,..,m$:

$$C_k^{(i)} = \tag{1}$$

$$\bigcap_{q\in[0,1]} \left\{ \left\{ i \notin \mathcal{R}_{BH}(P_1,..,P_{i-1},q,P_{i+1},..,P_m;\alpha) \right\} \cup \left\{ |\mathcal{R}_{BH}(P_1,..,P_{i-1},q,P_{i+1},..,P_m;\alpha)| = k \right\} \right\}$$

(here q is treated as a constant random variable taking the value q).

   (a) Write an explicit characterization of each event $C_k^{(i)}$ in terms of the order statistics of all $m-1$ p-values *except* P_i.

   (b) Prove that the events $C_k^{(i)} \cap \{P_i \leq \frac{k\alpha}{m}\}$ and $\{R = k\} \cap \{P_i \leq \frac{k\alpha}{m}\}$ are identical.

   (c) Prove that the events $C_k^{(i)}$ and $\{P_i \leq \frac{k\alpha}{m}\}$ are independent.

   (d) Prove that the events $C_k^{(i)}$ form a disjoint union of the sample space $\Omega = [0,1]^m$,

$$C_k^{(i)} \cap C_j^{(i)} = \emptyset, \forall k \neq j \quad ; \quad \bigcup_{k=1}^{m} C_k^{(i)} = \Omega \tag{2}$$

3. In this question we study the behavior of the FDR of the BH procedure for general dependency.

   (a) Prove that for general test statistics $X_1,..,X_m$ and rejection regions $\mathcal{R}_1,..,\mathcal{R}_m$, yielding p-values $P_1,..,P_m$, performing the BH procedure with parameter α controls the FDR at level $\frac{\alpha m_0(\log m+1)}{m}$. (Recall that the number of true null hypotheses m_0 can be $0,1,..,m$, and for the m_0 statistics X_i's corresponding to the null hypothesis H_0, we must have $P_i \sim U[0,1]$ and $Pr(X_i \in \mathcal{R}_i) = \alpha$.)
   Guidance: Let $C_k^{(i)}$ be the events defined in 2. Use the representation of the FDR as

$$FDR = \sum_{i=0}^{m_0} \sum_{k=1}^{m} \frac{1}{k} Pr\left(\left\{ P_i \leq \frac{k\alpha}{m} \right\} \cap C_k^{(i)} \right) \tag{3}$$

where w.l.o.g. the first m_0 null hypotheses are true, and decompose the events in the above sum further into the events:

$$A_{kj}^{(i)} \equiv \left\{ P_i \in \left(\frac{(j-1)\alpha}{m}, \frac{j\alpha}{m} \right] \right\} \cap C_k^{(i)} \tag{4}$$

(b) Find an explicit set of test statistics $X_1, ..., X_m$ and rejection regions $\mathcal{R}_1, .., \mathcal{R}_m$ such that the FDR of the BH procedure is strictly bigger than $\frac{\alpha m_0}{m}$.

(c) (* Bonus) Find a set of test statistics $X_1, ..., X_m$ and rejection regions $\mathcal{R}_1, .., \mathcal{R}_m$ such that the FDR is strictly bigger than α.

4. In this question we study empirically the effect of dependency on the FDR. Simulate test statistics from the same model as in question 5, but with parameters $m = 1000, m_0 = 500, \mu = 2$ and different values of ρ (to be specified in the sub-questions). For each set of parameters simulate $N = 5,000$ independent realizations of all m test statistics, and for each realization, apply the BH procedure with parameter $\alpha = 0.1$ and record the number of rejections R and false rejections V.

   (a) Draw a histogram of the total number of rejections, R for $\rho = 0, 0.95$. Explain your results.

   (b) Draw a histogram of the total number of false rejections, V for $\rho = 0, 0.95$. Explain your results.

   (c) Draw a histogram of the false discovery proportion, $Q = \frac{V}{R^+}$ for $\rho = 0, 0.95$. Explain your results.

   (d) For each $\rho = 0, 0.05, 0.1, .., 0.95, 1$ simulate N realizations and estimate the mean (FDR) and standard deviation of $Q = \frac{V}{R^+}$.
   Plot the estimators you got as a function of ρ . Compare the resulting estimated FDR to the theoretical guarantees according on the BH procedure we learned in class. Explain your results for both the mean and variance.

   Instructions: one way to simulate positively correlated Gaussians is as follows:

   First, simulate a set of independent Gaussians, $Y_0, Y_1, ...Y_m \sim N(0, 1)$.
   Then, for a parameter $0 \leq \rho \leq 1$ and for all $i = 1, .., m$ set $X_i = \rho Y_0 + (1 - \rho)Y_i + \mu_i$.

5. This question is meant to explain the empirical Bayes approach to FDR in a simplified manner.

   Assume that we test m hypotheses, each with a z-score test statistic X_i, such that the statistics have the following joint Gaussian distribution:

$$X \sim (\vec{\mu}, \Sigma) \tag{5}$$

with $\mu_i = 0$ for $i = 1, .., m_0$ (corresponding to H_0) and $\mu_i = \mu$ for $i = m_0 + 1, .., m$ (corresponding to H_1), and with $\Sigma_{ii} = 1, \Sigma_{ij} = \rho$ for $i \neq j$.

Assume that we perform a one-sided test, and for each test statistic we reject the null hypothesis if $X_i \geq C$, where C is determined by the procedure.

Simulate tests statistics $X_1, .., X_m$ from the above joint distribution with the following parameters: $m = 10000, m_0 = 5000, \mu = 2, \rho = 0$.

Assume a prior of $\pi_0 = 0.5$ for each hypotheses to be true null, and imagine that we estimated from the data the distribution of the z-scores under the null and alternative hypotheses to be $F_0 = N(0, 1)$ and $F_1 = N(2, 1)$ respectively (in reality we'll have to estimate π_0 and F_1 from the z-values themselves).

(a) Draw a histogram of the X_i with 100 equally spaced bins. Draw on the same plot the densities for $\pi_0 F_0, (1 - \pi_0)F_1$ and the mixture $\pi_0 F_0 + (1 - \pi_0)F_1$, scaled to match the number of observations (i.e. the areas under the histogram, F_0, F_1 and $\pi_0 F_0 + (1 - \pi_0)F_1$ should all be the same). Explain your results.

(b) Give an expression for the $FDR(z)$ and the local $fdr(z)$ as a function of z. Plot for $z \in [-4, 4]$ the expressions for both the $FDR(z)$ and local $fdr(z)$ a function and compare them. Explain your results.

(c) Convert each X_i to a right-tail p-value P_i (assuming X_i is a z-score), and compute the q-value for each P_i - this is minimal α for which the BH procedure will reject the i-th hypothesis. Plot on the same plot the values $\pi_0 \times q_i$ as a function of the X_i's (again, treating them as z-values) and compare to the empirical Bayes calculation of $FDR(z)$. Explain your results.