Modern Statistics 52311, 2020-21
Homework 1 - Hypothesis Testing

April 6, 2021

The grade for entire exercise is the sum of the 3 highest scores for individual questions - a complete and correct solution for each question is worth 33.3% 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 22/4/2021.

Problems

1. (a) Consider the two-sample problem with $Y_i \sim Bernoulli(0.5)$ and $X_i|Y_i = k \sim N(\mu_k, \sigma^2)$. Suppose that we sample $n = 100$ independent pairs (X_i, Y_i) from this model, with $\mu_k = k$, $k = 0, 1$ and $\sigma^2 = 4$.
   Our goal is to test the hypothesis $H_0 : \mu_0 = \mu_1$ with the two-sided alternative $H_1 : \mu_0 \neq \mu_1$ using the standard t-statistic (i.e. we perform a two-sided t-test).
   Simulate $R = 200$ samples of size $n = 100$ from the above model and estimate the power of the test at significance level $\alpha = 0.05$.

   (b) Repeat the above but with the Rank-sum test - that is, the test statistic is $R = \sum_{i:Y_i=0} r(X_i)$ where $r(X_i) \in \{1, .., n\}$ is the rank of the i-th observation among $X_1, .., X_n$.
   For each sample use $N = 500$ random permutations of the labels Y_i to compute the empirical p-value for this test for each random sample.

   (c) Suppose now that $X_i|Y_i = k \sim N(0, \sigma_k^2)$, i.e. the two samples have the same mean but different variances. We are interested in testing $H_0 : \sigma_0 = \sigma_1$ vs. $H_1 : \sigma_0 \neq \sigma_1$.
   Is the rank sum statistic appropriate for this test? Simulate $R = 200$ samples of size $n = 100$ from this model with $\sigma_0 = 1, \sigma_1 = 10$ and compute the power of the rank sum test at significance level $\alpha = 0.05$. What is your conclusion?

(d) Repeat the above but this time use the test statistic: $S = \sum_{i:Y_i=0} |r(X_i) - n/2|$. Is the power increased? why?

2. Consider the independence testing problem $H_0 : X \perp\!\!\!\perp Y$ vs. the alternative: $H_1 : X \not\perp\!\!\!\perp Y$, and suppose that the true joint distribution F_{XY} of X, Y (under H_1) is uniform in the triangle with vertices: $(0,0), (0,1)$ and $(1,0)$.

   (a) Simulate a sample of $n = 500$ points from the above model. Similarly, derive the marginals F_X, F_Y of F_{XY}, and simulate and plot a sample of $n = 1000$ form the product of the marginals $F_X F_Y$ representing the null hypothesis. Do the samples look visually different?

   (b) Derive the likelihood-ratio test (LRT) statistic where under H_1, (X, Y) follow the above distribution, and under H_o they are independent with the marginals of the above distribution F_{XY}. Write a formula for the statistic as a function of a sample $\{(x_1, y_1), ..., (x_n, y_n)\}$ drawn from the above distribution. Under H_0 this means that the data from the product $(X_i, Y_i) \sim F_X F_Y$ i.i.d. where $F_X F_Y$ are the marginals of the distribution F_{XY}.

   (c) For $n = 5, 10, 15, ..., 100$ simulate data from alternative distribution H_1 (i.e. from F_{XY}) with sample size n, using $R = 200$ repetitions for each n.
   For each sample apply two tests: (i) using the Pearson correlation test statistic, and (ii) using the LRT test statistic.
   For each statistic compute the empirical p-value using a permutation test with $N = 500$ permutations, and compute the average power at significance level $\alpha = 0.05$ for each n.
   Plot the power as a function of n for both tests. Which test is more powerful? is this surprising? which test would you use? explain.

3. 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.

Remark 1. *Recall that for a distribution-free statistic, the distribution of T under H_0 (i.e. $F(X^{(1)}, X^{(2)}) = F_1(X^{(1)})F_2(X^{(2)})$) does not depend on F_1, F_2.*

4. 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 the proof for independence for $i, k = 1, .., m$:

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

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

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

Use the representation of the FDR as

$$FDR = \sum_{i=0}^{m_0} \sum_{k=1}^{m} \frac{1}{k} Pr\Big(\Big\{ P_i \le \frac{k\alpha}{m} \Big\} \cap C_k^{(i)} \Big) \tag{2}$$

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:

$$A_{kj}^{(i)} \equiv \Big\{ P_i \in (\frac{(j-1)\alpha}{m}, \frac{j\alpha}{m}] \Big\} \cap C_k^{(i)} \tag{3}$$

(b) Find an explicit set of test statistics $X_1, ..., X_m$ and rejection regions $\mathcal{R}_1, .., \mathcal{R}_m$ such that the for the BH procedure we have $FDR > \frac{\alpha m_0}{m}$.
Hint: We know that under PRDS this cannot hold, hence we're seeking negative dependencies

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