Modern Statistics 52311, 2019-20
Homework 1 - Hypothesis Testing, James-Stein Estimator

March 19, 2020

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 19/4/2020.

Problems

1. Let T be a continuous test statistic computed for a two-sample test on $(x_1, y_1), .., (x_n, y_n)$ for testing independence, $H_0 : X \perp\!\!\!\perp Y$. Suppose that we perform a permutation test by applying N independently drawn random permutations to the dataset, computing T_i for each permuted dataset $(x_1, y_{\pi_i(1)}), .., (x_n, y_{\pi_i(n)})$, and rejecting H_0 if $\#\{T_i > T\} \leq \lfloor \alpha N \rfloor - 1$. We assume that ties between the T_i's are broken randomly.

   (a) Let $P_I(N; \alpha | H_0)$ be the probability for type-1-error for the test. Prove that for any $N \in \mathbb{N}$, $P_I(N; \alpha | H_0) \leq \alpha$.

   (b) Prove that $\lim_{N \to \infty} P_I(N; \alpha | H_0) = \alpha$.

   (c) Suppose that the distribution of T under the null and the alternative are $T|H_0 \sim G_0$ and $T|H_1 \sim G_1$, respectively. (here G_0, G_1 are cumulative distributions which you can assume are strictly increasing). Let $P_{II}(N; \alpha | H_1)$ be the probability for type-2-error. Prove that $\lim_{N \to \infty} P_{II}(N; \alpha | H_1) = C$ for some constant $0 \leq C \leq 1$. Give an explicit *approximate* expression for C in terms of α, G_0 and G_1. You may assume that only permutations without fixed points are allowed (i.e. we reject a permutation P if $P(i) = i$ for some i). What does the expression for C give if the variables X, Y are independent also under H_1? (i.e. $G_0 = G_1$).

2. Download the file *data_EX1.txt* from the course moodle. The file contains a sample of $n = 1000$ observations for two random variables X and Y. Your goal is to test whether these two r.v.s. are independent, that is $H_0 : X \perp\!\!\!\perp Y$ vs. the alternative: $H_1 : X \not\perp\!\!\!\perp Y$.

   (a) Compute the Pearson correlation coefficient test statistic and do a permutation test with $\alpha = 0.01$ significance level using at least $M = 1000$ permutations. Plot a histogram of your permuted test statistics with the original test statistic clearly marked and compute a p-value for the permutation test. Did you reject the null hypothesis saying that the r.v.s. are independent? Explain your results.

   (b) Repeat the above but replace Pearson's correlation by Hoeffding's test statistic. Compare your results to the previous test. Are they similar or different? explain.

   (c) By sampling from the original sample $\{(x_1, y_1), ..., (x_n, y_n)\}$ with replacement we can obtain a bootstrap null sample: $\{(x_{I_1}, y_{J_1}), ..., (x_{I_n}, y_{J_n})\}$ where $I_1, .., I_n, J_1, ..., J_n \overset{i.i.d.}{\sim} U\{1, n\}$. Compute the null distribution of the two test statistics from above using M bootstrap samples (for same M you used in 2a) and compute a p-value for the bootstrap-test in both cases. Are the null distributions similar to the permutation-test null distributions?

3. Consider the same independence testing problem as in 2, and suppose that the true joint distribution of X, Y is as follows, for Z independent of X, Y':

$$\begin{pmatrix} X \\ Y' \end{pmatrix} \sim N(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & 0.6 \\ 0.6 & 1 \end{pmatrix}),\ Z \sim Ber(0.5),\ Y = (2Z - 1)Y' \tag{1}$$

   (a) Derive the likelihood-ratio test statistic where (X, Y) follow the above distribution under the alternative H_1 and are independent with the marginals of the above distribution under the null H_0. 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.

   (b) For $n = 100, 200, ..., 1000$ simulate data from the product $F_X F_Y$ of the distribution F_{XY} in eq. (1), representing the null H_0. Perform a permutation test and a bootstrap test using both Hoeffding's test statistic and the likelihood ratio test statistic (four tests in total) with significant level $\alpha = 0.05$. Plot the actual observed type-1-error as a function of n for the four tests. Perform at least $R = 100$ repetitions and use at least $M = 100$ permutation or bootstrap samples in each repetition.

   (c) For $n = 100, 200, ..., 1000$ simulate data from alternative distribution in eq. (1), representing the alternative H_1 and repeat the four tests from 3b, plotting for each test the obtained power as a function of n. Which test is more powerful? is this surprising? which test would you use? explain.

4. In the next two questions we derive the risk of the Bayes and James-Stein estimators for the means of independent Gaussian random variables.

   Let $x_i \sim N(\mu_i, 1)$, $i = 1, .., n$ be independent Gaussian random variables (in matrix form: $x \sim N(\mu, I_n)$.) We use the squared loss. Recall that the risk under this loss of the MLE estimator $\hat{\mu}^{(MLE)} = x$ is n. For a given prior $P(\beta)$ define the Bayes estimator as the posterior mean, $\hat{\mu}^{(Bayes)} = E[\beta|x]$.

   (a) Suppose that the parameters μ_i have i.i.d. prior $\mu_i \sim N(0, \sigma^2)$. Prove that the Bayes estimator is $\hat{\mu}^{(Bayes)} = \frac{\sigma^2}{\sigma^2+1}x$.

   (b) Prove that if the true parameters vector is μ, then the risk of $\hat{\mu}^{(Bayes)}$ is:

$$R_\mu(\hat{\mu}^{(Bayes)}) = \Big[1 - \frac{\sigma^2}{\sigma^2+1}\Big]^2 ||\mu||^2 + n\Big[\frac{\sigma^2}{\sigma^2+1}\Big]^2 \tag{2}$$

   (c) Suppose that the parameters μ_i have i.i.d. prior $\mu_i \sim N(0, \sigma^2)$. Prove that the overall Bayes risk of the Bayes estimator is:

$$R^{(Bayes)}(\hat{\mu}^{(Bayes)}) = E_\mu\Big[R^{(Bayes)}(\mu)\Big] = \frac{\sigma^2}{\sigma^2+1}n \tag{3}$$

   (d) Show that for *any* estimator (with finite moments) $\hat{\mu}$ we have:

$$E[||\hat{\mu} - \mu||^2] = E[||\hat{\mu} - x||^2] - n + 2\sum_{i=1}^{n} COV(x_i, \hat{\mu}_i) \tag{4}$$

5. Under the conditions of question 4, and assuming $n \geq 3$, define the James-Stein estimator $\hat{\mu}^{(JS)} = (1 - \frac{n-2}{||x||^2})x$.

   (a) Use integration by parts and the multivariate Gaussian density to show that for any continuously differentiable estimator (with finite moments):

$$COV(x_i, \hat{\mu}_i) = E[\frac{\partial \hat{\mu}_i}{\partial x_i}] \tag{5}$$

   (b) Use the previous result and 4d to show that the risk of the James-Stein estimator for a given parameter μ is given by:

$$R_\mu(\hat{\mu}^{(JS)}) = n - E\Big[\frac{(n-2)^2}{||x||^2}\Big]\Big(< n = R_\mu(\hat{\mu}^{(MLE)})\Big) \tag{6}$$

   (c) Suppose that the parameters μ_i have i.i.d. prior $\mu_i \sim N(0, \sigma^2)$. Prove that the overall Bayes risk of the James-Stein estimator is:

$$R^{(Bayes)}(\hat{\mu}^{(JS)}) = E_\mu\Big[R_\mu(\hat{\mu}^{(JS)})\Big] = \frac{n\sigma^2 + 2}{\sigma^2 + 1} \tag{7}$$