CSE 544, Spring 2020: Probability and Statistics for Data Science

Assignment 5: Hypothesis Testing Due: 4/20, 2:30pm via google forms

(5 questions, 70 points total)

I/We understand and agree to the following:

(a) Academic dishonesty will result in an 'F' grade and referral to the Academic Judiciary.

(b) Late submission, beyond the 'due' date/time, will result in a score of 0 on this assignment.

(write down the name of all collaborating students on the line below)

---

1. Hypothesis Testing for a single population (Total 10 points)

Consider the 10 samples: {1.68, 1.34, 1.98, 0.97, 1.09, 2.65, 1.23, 1.78, 0.6, 0.85}. Use the K-S test to check whether these samples are from the Uniform(0, 3) distribution. First, set up the hypotheses. Then, create a 10 X 6 table with entries: $[x, F_Y(x), \hat{F}_X^-(x), \hat{F}_X^+(x), \left|\hat{F}_X^-(x) - F_Y(x)\right|, |\hat{F}_X^+(x) - F_Y(x)|]$, where $\hat{F}_X^-(x)$ and $\hat{F}_X^+(x)$ are the values of the eCDF to the left and right of x, and $F_Y(x)$ is the CDF of Uniform(0, 3) at x; this is the same notation as in class. Finally, compare the max difference with the threshold of 0.37 to Reject/Accept. Show all rows and columns.

2. Toy Example for Permutation Test (Total 10 points)

Let X = {2, 9} and Y = {4}. The null hypothesis is that X and Y are from the same distribution. Use the permutation test to decide this using a p-value threshold of 0.05. Please show all steps for each permutation clearly.

3. Independence Tests to Save Your Casino (Total 15 points)

Being the owner of Casino 544, you are concerned that you are losing a lot of money because of the dealers at the blackjack tables. The Null hypothesis is that the outcome of the tables should be independent of the dealer, but you aren't sure.

(a) Validate your claim based on the dealer observations for a day, using the χ^2 test. Use α=0.05. You can use tools/online resources to find the CDF of χ^2; one such tool is https://www.danielsoper.com/statcalc/calculator.aspx?id=62. (10 points)

	Dealer A	Dealer B	Dealer C
Win	48	54	19
Draw	7	5	4
Loose	55	50	25

(b) You want to be more certain about the loyalty of your dealers, so you collect more data: number of wins from each dealer for 10 days. Find the Pearson correlation coefficient for each pair of dealers. What can you conclude? (5 points)

	Day-1	Day-2	Day-3	Day-4	Day-5	Day-6	Day-7	Day-8	Day-9	Day-10
Dealer A	48	40	58	53	65	25	52	34	30	45
Dealer B	54	48	51	47	62	35	70	20	25	40
Dealer C	19	40	35	41	38	32	32	37	37	15

4. Real-World Example Based on NBA (Total 20 points)

NBA veterans often complain about how easy it is to score points in today's games as opposed to prior decades. This question attempts to verify this claim. We will compare the distribution of points per game for all teams between 1999 and 2009, and between 2009 and 2019. The *Null hypothesis is that the scoring distribution remains the same*. Refer to the "**Team Per Game Stats**" table found on "https://www.basketball-reference.com/leagues/NBA_2019.html#all_team-stats-per_game" (this is 2019 data); you can obtain data for different years by changing the year in the URL. We only care about the last column i.e. Points (PTS); this column is the points per game. Use the "Share & more" option to format data. You can use any programming tool for this problem. Submit code for this question as q4.py.

(a) Using Permutation test, check if the Null hypothesis can be accepted or rejected for the two cases (1999 vs. 2009, and 2009 vs. 2019). Choose n=200 random permutations and use a p-value threshold of 0.05. Clearly show the p-value you obtain. (7 points)

(b) Repeat part (a) with n=2000. (7 points)

(c) Repeat part (a) but using two-sample K-S test with a max difference threshold of 0.05. You do not need to show the full table but do create the two eCDF graphs on the same figure and identify the max difference in the figure along with its value. Print and attach this figure. (6 points)

5. Type-1 and Type-2 error for one-sided unpaired T-test (Total 15 points)

Let $\{X_1, X_2, \dots, X_n\}$ be i.i.d. from Normal(μ_1, σ_1^2) and $\{Y_1, Y_2, \dots, Y_m\}$ be i.i.d. from Normal(μ_2, σ_2^2). Also suppose *X*'s and *Y's* are independent, and μ_1, σ_1^2, μ_2, σ_2^2 are unknown. Let S_x and S_Y be the sample standard deviations of the two populations. Assume that *n* and *m* are large. Let H_0: $\mu_1 > \mu_2$ be the null hypothesis and H_1: $\mu_1 <= \mu_2$ be the alternate hypothesis. Consider the T statistic for the unpaired T test, as in class, with $\delta > 0$ being the critical value.

(a) For the above test, show that the probability of Type-1 and Type-2 errors are given by $\Phi(-\delta - \frac{\mu_1-\mu_2}{\sqrt{\frac{Sx^2}{n}+\frac{Sy^2}{m}}})$ and $1 - \Phi(-\delta - \frac{\mu_1-\mu_2}{\sqrt{\frac{Sx^2}{n}+\frac{Sy^2}{m}}})$, respectively. (10 points)

(b) Show that the p-value is given by $\Phi(\frac{\bar{X}-\bar{Y}}{\sqrt{\frac{Sx^2}{n}+\frac{Sy^2}{m}}})$. (5 points)