CSE 544, Spring 2022: Probability and Statistics for Data Science

Assignment 5: Hypothesis Testing Due: 4/18, 8:15pm, via Blackboard

(6 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 7 points)

Consider the 10 samples: {0.592, 0.774, 0.245, 0.424, 0.685, 0.436, 0.648, 0.959, 0.842, 0.995}. Use the K-S test to check whether these samples are from the distribution with pdf p(x) = 2x , x>=0. First, set up the hypotheses. Then, create a 10 X 6 table with entries: $[x, F_Y(x), \widehat{F}_X^-(x), \widehat{F}_X^+(x), \left|\widehat{F}_X^-(x) - F_Y(x)\right|$, $|\widehat{F}_X^+(x) - F_Y(x)|]$, where $\widehat{F}_X^-(x)$ and $\widehat{F}_X^+(x)$ are the values of the eCDF to the left and right of x, and $F_Y(x)$ is the CDF of p(x) at x; this is the same notation as in class. Finally, compare the max difference with the threshold of 0.25 to Reject/Accept. Show all rows and columns.

2. Toy Example for Permutation Test (Total 6 points)

Let X = {3,2} and Y = {2,5}. 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 Boxing Ring (Total 15 points)

Being the owner of a local boxing ring, you want to be sure that your ring's judges are scoring boxing matches correctly. The judges score the matches of your ring's best player (Player 1) against several other players roughly of the same caliber. Hence, it can be assumed that the outcome of a match is independent of Player 2. The Null hypothesis is that the outcome of the match should be independent of the judge, but you aren't sure.

(a) Validate your claim based on the judge 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)

	Judge A	Judge B	Judge C
Player 1 Wins	72	50	24
Draw	8	5	3
Player 1 Loses	20	8	9

(b) You want to be more certain about the expertise of your judges, so you collect more data: number of player 1 wins from each judge for 10 days. Find the Pearson correlation coefficient for each pair of judges. What can you conclude? Is a particular judge not doing their job? (5 points)

	Day-1	Day-2	Day-3	Day-4	Day-5	Day-6	Day-7	Day-8	Day-9	Day-10
Dealer A	40	51	38	50	32	32	44	55	35	49
Dealer B	42	48	42	46	48	35	46	45	29	49
Dealer C	30	32	36	42	43	36	45	45	48	45

4. Potato Farming (Total 8 points)

We have the potato yield from 12 different farms, given by D below. We know that the standard potato yield for the given variety is μ=20. (Use 0.05 significance level unless mentioned otherwise). D = [21.5, 24.5, 18.5, 17.2, 14.5, 23.2, 22.1, 20.5, 19.4, 18.1, 24.1, 18.5].

(a) Test if the potato yield from these farms is significantly better than the standard yield using the T-test. (4 points)

(b) Due to the introduction of a new fertilizer, the farming company believes there is an improvement in the standard potato yield. To see if this is the case, a sample of size 12 is chosen and their mean yield is found to be 22. Historically, the standard deviation of potato yields is 3. Can we support the claim that there is an improvement in the standard potato yield due to the new fertilizer? Assume that the population standard deviation has not changed. Use the test you think is most suitable for this objective. (4 points)

5. Type-1 and Type-2 error for one-sided unpaired T-test (Total 10 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 *H₀:* $\mu_1 > \mu_2$ be the null hypothesis and *H₁:* $\mu_1 <= \mu_2$ be the alternate hypothesis. Consider the T statistic for the unpaired T test, as in class, with $\delta > 0$ being the critical value.

(Hint: The condition that H0 or H1 is True may be ignored when computing the Type-1 and Type-2 error)

(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. (5 points)

(b) Show that the p-value is given by $\Phi\left(\frac{\bar{X} - \bar{Y} - (\mu_1 - \mu_2)}{\sqrt{\frac{Sx^2}{n} + \frac{Sy^2}{m}}}\right)$. (5 points)

6. COVID-19 Perceptions (Total 24 points)

We are going to verify/debunk a few common perceptions about COVID-19 outbreak in this question using the dataset on the class website a5_q6.csv. Use p-value of threshold of 0.05 in all parts.

(a) Using Z-Test, verify if we can accept or reject the Null hypothesis that temperature doesn't affect COVID-19 Outbreak. A temperature greater than 12 is considered warm and less than equal to 12 is considered cold. Clearly show the z-statistic and p-value obtained. (6 points)

(b) Using Z-Test, verify if we can accept or reject the Null hypothesis that humidity doesn't affect COVID-19 Outbreak. A humidity level greater than equal to 65 is considered humid and less than 65 is considered not humid. Clearly show the z-statistic and p-value obtained. (6 points)

(c) Verify the same hypothesis as in part (a) by using the permutation test. Use n=500 and n=1000 random permutations. Clearly show the p-value obtained. (6 points)

(d) Verify the same hypothesis as in part (b) by using the permutation test. Use n=500 and n=1000 random permutations. Clearly show the p-value obtained (6 points)

Submit all parts in a single a5_q6.py file.