CSE 544, Spring 2021: Probability and Statistics for Data Science

Assignment 5: Hypothesis Testing Due: 4/20, 1:15pm, 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: {2.78, 0.84, 1.88, 2.23, 1.99, 0.04, 2.65, 0.74, 1.19, 2.57}. 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 0.25 to Reject/Accept. Show all rows and columns.

2. Toy Example for Permutation Test (Total 5 points)

Let X = {5} and Y = {2, 7}. 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 $\alpha=0.05$. 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 Healthcare (Total 20 points)

According to the World Health Organization (WHO) stroke is the 2nd leading cause of death globally, responsible for approximately 11% of total deaths. Whether a patient is likely to get stroke depends on the gender, average glucose level, etc. This question is to verify the relationship of stroke and these factors. Factors like age and glucose level were collected when the stroke was recorded. The original data comes from https://www.kaggle.com/fedesoriano/stroke-prediction-dataset. Datasets in this question are created from the original one. You can use any programming tool for this problem. Submit code for this question as q4.py.

(a) Use the two columns (stroke and avg_glucose_level) in the first dataset "data_q4_1.csv", where stroke=1 means the patient gets stroke. The Null hypothesis is that people getting stroke tend to have the same glucose level as people who do not get stroke. Use Permutation test to check if the Null hypothesis should be accepted or rejected. Choose n=200 and n=1000 random permutations, respectively. Use a p-value threshold of 0.05. Clearly show the p-value you obtain. (7 points)

(b) Dataset "data_q4_2.csv" has two columns, gender and age, when the patients get stroke. The Null hypothesis this time is that female patients get a stroke at the same age as male patients. Using Permutation test, check if the Null hypothesis should be accepted. Choose n = 1000 permutations with a p-value threshold of 0.05. (6 points)

(c) Repeat part (b) 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. (7 points)

5. Type-1 and Type-2 error for one-sided unpaired T-test (Total 10 points)

Let $\{X_1, X_2, \ldots, X_n\}$ be i.i.d. from Normal(μ_1, σ_1^2) and $\{Y_1, Y_2, \ldots, Y_m\}$ 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. (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. Z- and T-testing (Total 13 points)

We have learned from the lecture that the difference between T-test and Z-test is that T-test uses sample deviation while Z-test uses the true deviation of the distribution. Intuitively, Z-test will perform better in a sample of a small size if given the true deviation. Is that the case in practice? We will check this in this question. You will need the q6_X1.csv, q6_Y1.csv for (a) and q6_X2.csv, q6_Y2.csv for (b) (available at the class website). For the below questions, you do not need to provide any code but please provide your steps and intermediate values (mean, deviation, p-value) as needed, so the TA can evaluate your solution.

(a) For q6_X1.csv and q6_Y1.csv, each has 20 samples independently sampled from $X \sim N(1.5, 1)$ and $Y \sim N(1,1)$ respectively. Now assume that you don't know the mean of X and Y but their true deviations are given. The Null hypothesis is that X and Y have the same mean value. Using Z-test and T-test to check the hypothesis. Under $\alpha = 0.05$, use a threshold (or comparison value on right-hand-side of test) of 1.962 for Z-test and 2.086 for T-test. (6 points)

(a) Do the same thing as in (a) but use q6_X2.csv and q6_Y2.csv, each having 1000 samples sampled from $X \sim N(1.5, 1)$ and $Y \sim N(1,1)$ respectively. Note that the threshold for T-test in this case is the same as Z-test, which is 1.962. Compare the result with (a) and compare p-values from Z-test and T-test in each question. Is there a significant advantage of using Z-test in a small sample? (7 points)