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

Assignment 4: Parametric Inference & Hypothesis Testing Due: 4/8 2:30pm, via google form
(7 questions, 75 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. Practice with MME **(Total 10 points)**

(a) The Gamma(x, y) distribution has mean $x \cdot y$ and variance $x \cdot y^2$. Find MME for $\hat{x}$ and $\hat{y}$. (4 points)

(b) Find MME $\hat{a}$ and $\hat{b}$ for the Uniform(a, b) distribution. Express your final answer in terms of the sample mean, $\bar{X} = (\sum X_i)/n$, and sample variance, $\overline{S^2} = ((\sum X_i^2)/n\,) - \,\bar{X}^2$. (6 points)

2. Consistency of MLE (Total 5 points)

Let $X_1, X_2, \ldots, X_n$ be distributed as Exponential(λ), all i.i.d. Show that the MLE($\hat{\lambda}$) will converge to the unknown parameter λ. Prove this by showing that bias($\hat{\lambda}$) and se($\hat{\lambda}$) tends to 0 as n tends to ∞.

3. Practice with MLE **(Total 11 points)**

(a) Let $X_1, X_2, \dots, X_n$ be distributed as Poisson(λ). Find the MLE of λ. (3 points)

(b) Let $X_1, X_2, \dots, X_n$ be a sample from the distribution whose density function is given by f(x)= $\frac{1}{2}e^{-|x-\theta|}$, -∞ <x<∞. Determine the MLE of θ and comment on it. (4 points)

(c) Let $X_1, X_2, \dots, X_n \sim$ Normal(θ, 1). Let δ =$E[I_{X_1>0}]$. Use the Equivariance property to show that the MLE of δ is $\varphi(\frac{1}{n}\sum_{i=1}^{n} X_i)$. You can assume the MLE of the Normal as derived in class. (4 points)

4. Parametric Inference with Data Samples **(Total 13 points)**

Let $X = \begin{cases} 2 & with\ prob\ \theta \\ 3 & otherwise \end{cases}$, where θ is unknown. Let D = {2, 3, 2} be drawn i.i.d. from X.

(a) Derive $\hat{\theta}_{MME}$ using D as the sample data. Clearly show all your steps. (3 points)

(b) Derive $\widehat{se}(\hat{\theta})$ using estimates from part (a). Specifically, first derive $se(\hat{\theta})$ in terms of θ, and then estimate $\widehat{se}(\hat{\theta})$, as in class. Show all your steps clearly. (5 points)

(c) Derive $\hat{\theta}_{MLE}$ using D as the sample data. Clearly show all your steps. (5 points)

5. MME versus MLE using real data (Total 14 points)

For this question, we will use the acceleration, model, and mpg data from the Auto-mpg dataset (https://www.kaggle.com/uciml/autompg-dataset). Please use the data files on the class website. We will assume that acceleration is Normal(μ, σ^2) distributed, model year is Uniform(a, b) distributed, and mpg is Exponential(λ) distributed. You are to find the MME and MLE estimates of the parameters of the distributions for all 3 datasets. For the Normal MME and Normal and Uniform MLE, you can directly use the results from class. For the MME of Uniform, you can use the result from Q1. For the remaining cases (Exponential MME and MLE), we will first derive the estimates.

(a) For the Exp(λ) distribution, find the $\hat{\lambda}_{MME}$. (2 points)

(b) For the Exp(λ) distribution, find the $\hat{\lambda}_{MLE}$. (2 points)

(c) For the 3 datasets, find the MME estimates. That is, find the MME for μ and σ^2 for the acceleration dataset, a and b for the model dataset, and λ for the mpg dataset. Provide your answer as a number with 3 significant digits. (4 points)

(d) Same as part (c), but this time find the MLE estimates. (4 points)

(e) Based on your answers for (c) and (d), can you comment on which is more accurate among MME and MLE? This is an open-ended question, so subjective arguments will suffice. (2 points)

Report the required results for (c) and (d) in the hardcopy but submit the Python code for these parts via the google form link for A4 (link on piazza).

6. More on Hypothesis testing (Total 14 points)

(a) Suppose the null hypothesis is H_0: $\theta = \theta_0$, but the true value of θ is θ_*. Show that, under Wald's test, the probability of a Type II error is $\varphi(\frac{\theta_0-\theta_*}{\widehat{se}} + z_{\alpha/2}) - \varphi(\frac{\theta_0-\theta_*}{\widehat{se}} - z_{\alpha/2})$.

(Hints: (i) might help to draw a figure; (ii) think about the distribution of the estimate.) (6 points)

(b) You will need q6_X.dat and q6_Y.dat available at the class website for this question. Each contains 1000 samples for X and Y drawn from two independent Normal distributions. In the following, test whether the population means of X and Y are same (null) or not (alternative). Use Wald's 2-population test with $\alpha = 0.05$. Is this test applicable here? (4 points)

(c) Assume X and Y are dependent. Repeat part (b) but use the paired t-test with $\alpha = 0.05$ threshold of 1.962. Is this test applicable here? (4 points)

Report the required results for (b) and (c) in the hardcopy but submit the Python code for these parts via the google form link for A4 (link on piazza).

7. Hypothesis Testing for a single population **(Total 8 points)**

(a) Consider the following 10 samples: {1.87, 1.29, 2.01, 0.93, 1.02, 2.78, 2.33, 1.65, 0.50, 0.99}. Assuming that the 10 samples are normally distributed, use the t-test to decide the null hypothesis that the population mean is 1.5. Use the $\alpha = 0.05$ threshold of 2.228 to Reject/Accept. (3 points)

(b) You observe 46 successes in 100 trials of a coin. If the null hypothesis is that the coin is unbiased, use the Wald's test with the MLE/MME with $\alpha = 0.05$ to Reject/Accept the null. What if the null hypothesis is that the coin has $p=0.7$? (5 points)