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

Assignment 4: Parametric Inference & Hypothesis Testing Due: 4/08, 1:15pm, via Blackboard
(8 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. Practice with MME **(Total 9 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, $\overline{S^2} = ((\sum X_i^2)/n) - \bar{X}^2$. (5 points)

2. Consistency of MLE **(Total 6 points)**

Let $X_1, X_2, \dots, X_n$ be distributed as Exponential(1/β), all i.i.d. Show that the MLE($\hat{\beta}$) will converge to the unknown parameter β. Prove this by showing that bias($\hat{\beta}$) and se($\hat{\beta}$) tends to 0 as n tends to ∞. You can use the fact that the mean and variance of Exponential(λ) are 1/λ and $1/\lambda^2$, respectively.

3. Practice with MLE **(Total 10 points)**

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

(b) Let $X_1, X_2, \ldots, X_n$ be distributed i.i.d. as Normal(μ, σ^2). Show that the MLE of μ and σ^2 is the same as the sample mean and (uncorrected) sample variance, respectively. (4 points)

(c) Let $X_1, X_2, \ldots, X_n \sim$ Normal(θ, 1). Let δ $= E[I_{X_1>0}]$. Use the Equivariance property to show that the MLE of δ is $\Phi(\frac{1}{n}\sum_{i=1}^{n} X_i)$, where $\Phi()$ is the CDF of the standard Normal. You can use the MLE of the Normal as provided in 3(b). (3 points)

4. Parametric Inference with Data Samples **(Total 10 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) Provide a numerical estimate of the 95%ile confidence intervals for $\hat{\theta}_{MME}$. Start by deriving $\widehat{se}(\hat{\theta}_{MME})$: first derive $se(\hat{\theta}_{MME})$ in terms of θ, and then estimate $\widehat{se}(\hat{\theta}_{MME})$, as in class. Show all your steps. Your final answer should be a numerical range. (4 points)

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

5. MME versus MLE using real data (Total 10 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 Uniform MLE, you can directly use the results from class. For the Normal MLE, use the result from Q3(b); for Uniform MME, use the result from Q1(b). For the Exponential, 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. (3 points)

(d) Same as part (c), but this time find the MLE estimates. (3 points)

6. Clinical Testing **(Total 4 points)**

Consider the sick patient example from class. In a clinical trial of a new disease detection test, there were 100 healthy patients and 100 sick patients. The test correctly identified 98 out of the 100 healthy patients as healthy. The test also correctly identified 99 of the 100 sick patients as sick. The remaining patients were incorrectly classified.

(a) What is the precision of the test? (1 point)

(b) What is the recall of the test? (1 point)

(c) What is the Type I error of the test? (1 point)

(d) What is the Type II error of the test? (1 point)

7. Wald's test **(Total 11 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 $\Phi(\frac{\theta_0-\theta_*}{\widehat{se}} + z_{\alpha/2}) - \Phi(\frac{\theta_0-\theta_*}{\widehat{se}} - z_{\alpha/2})$.

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

(b) You 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 or 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)

8. More on Wald's test **(Total 10 points)**

(a) Use q8_a.csv dataset and assume it is distributed as Normal(θ, σ^2). Apply the Wald's test with α = 0.02 to check whether the true mean is $\theta_0 = 0.5$. Use sample mean to obtain $\hat{\theta}$ and corrected sample variance estimator for obtaining $\widehat{\sigma^2}$. (4 points)

(b) Use q8_b_X.csv and q8_b_Y.csv available at the class website for this question. Each contains 750 samples for X and Y drawn from two independent Normal distributions. Without worrying about the applicability of the test, use Wald's 2-population test with α = 0.05 to test whether the population means of X and Y are same (null) or not (alternative). Is this test applicable here? (6 points)