CSE 544, Fall 2018, Probability and Statistics for Data Science

Assignment 4: Statistical Inference Due: 11/12, in class

(7 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. More on Wald's test **(Total 5 points)**

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.)

2. Posterior for Normal **(Total 10 points)**

Let $X_1, X_2, \dots, X_n$ be distributed as Normal(θ, σ^2), where σ is assumed to be known. You are also given that the prior for θ is Normal(a, b^2).

(a) Show that the posterior of θ is Normal(x, y^2), such that: (7 points)

$x = \frac{b^2\bar{X}+se^2a}{b^2+se^2}$ $and\ y^2 = \frac{b^2se^2}{b^2+se^2}$; where $\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i$ and $se^2 = {\sigma^2}/{n}$.

(Hint: less messier if you ignore the constants, but please justify why you can ignore them)

(b) Compute the (1-α) posterior interval for θ. (3 points)

3. Consistency of MLE (Total 5 points)

Let $X_1, X_2, \ldots, X_n$ be distributed as Exponential $(\frac{1}{\beta})$, 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 ∞.

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

5. Bayesian Inference in action **(Total 15 points)**

You will need the q5_sigma3.dat and q5_sigma100.dat files for this question; these files are on the class website. Each file contains 5 rows of 100 samples each. Refer back to Q 2 (a); you can use its result even if you have not solved that question.

(a) Assume that $\sigma = 3$ (meaning $\sigma^2 = 9$). Let the prior be the standard Normal (mean 0, variance 1). Read in the 1st row of q5_sigma3.dat and compute the new posterior. Now, assuming this posterior is your new prior, read in the 2nd row of q5_sigma3.dat and compute the new posterior. Repeat till the 5th row. Please provide your steps here and draw a table with your estimates of the mean and variance of the posterior for all 5 steps (table should have 5 rows, 2 columns). Also plot each of the 5 posterior distributions on a single graph and attach this graph. What do you observe? (7 points)

(b) Now assume that $\sigma = 100$ and repeat part (a) above but with q5_sigma100.dat. Assume the same prior of a standard Normal. Provide the table and final graph. What do you observe? (7 points)

(c) Based on the comparison of answers of (a) and (b), what can you conclude? (1 point)

6. Hypothesis Testing for a single population **(Total 13 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}. Use the K-S test to check whether these samples are from the Uniform(0, 3) distribution or not. First, set up the hypotheses. Then, create a 10 X 6 table with entries: $[x, \hat{F}_{-}(x), \hat{F}_{+}(x), F_0(x), |\hat{F}_{-}(x) - F_0(x)|, |\hat{F}_{+}(x) - F_0(x)|]$, where $\hat{F}_{-}(x)$ and $\hat{F}_{+}(x)$ are the values of the empirical distribution function to the left and right of x, and $F_0(x)$ is the CDF of Uniform(0, 3) at x. Finally, compare the max of the last two columns with the α = 0.05 threshold of 0.41 to Reject/Accept. (5 points)

(b) Assuming that the 10 samples are normally distributed, use the t-test to decide the null hypothesis that the populatiøn mean is 1.5. Use the α = 0.05 threshold of 2.228 to Reject/Accept. (3 points)

(c) 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 α = 0.05 to Reject/Accept the null. What is the p-value? What if the null hypothesis is that the coin has p=0.7? What is the p-value? (5 points)

7. Hypothesis Testing for two populations (Total 12 points)

You will need q7_X.dat and q7_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). Show your steps, but feel free to use any software to compute mean and variance, etc.

(a) Use Wald's 2-population test with $\alpha = 0.05$. Is this test applicable here? (4 points)

(b) Use Welch's test with $\alpha = 0.05$ threshold of 1.962. Is this test applicable here? (4 points)

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