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

Assignment 3: Statistical Inference Due: 10/29, in class

(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 and Properties of Estimator **(Total 10 points)**

(a) The Beta distribution with parameters α and β has mean $\frac{\alpha}{(\alpha+\beta)}$ and variance $\frac{\alpha\beta}{(\alpha+\beta+1)(\alpha+\beta)^2}$. Find the MME estimate for α and β. (5 points)

(b) Recall the three properties of estimators from class: (i) bias($\hat{\theta}$) = E[$\hat{\theta}$] - θ, (ii) se($\hat{\theta}$) = $\sqrt{Var[\hat{\theta}]}$, and (iii) MSE($\hat{\theta}$) = E[($\hat{\theta}$ - θ)^2]. Find these for $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$, where $X_i \sim$ Poisson(θ). (5 points)

2. Plug-in estimates (Total 10 points)

(a) Show that the plug-in estimator of the variance of X is $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$, where $\bar{X}_n$ is the sample mean, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. (3 points)

(b) Show that the bias of $\hat{\sigma}^2$ is $-\sigma^2/n$, where σ^2 is the true variance. (4 points)

(c) The skew for a RV X with mean μ and variance σ^2 is defined a$s\ Skew[X] = E[(X-\mu)^3]\ /\ \sigma^3$. Derive the plug-in estimate of the skew in terms of the sample data. (3 points)

3. Method of Moments Estimator (MME) with Data Samples (Total 5 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. (2 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. (3 points)

4. Normal-based CI for $\widehat{\boldsymbol{F}}$ **(Total 10 points)**

$X_1, X_2, \dots, X_n$ be i.i.d. RVs with true CDF *F*. Let $\widehat{F}$ be their empirical distribution function.

(a) Derive Var[$\widehat{F}$]. (4 points)

(b) Derive the Normal-based (1-α) CI for $\widehat{F}$. (6 points)

5. MME in Practice (Total 10 points)

Elon Husk wants to make space tourism more accessible and for that he wants to know the amount of fuel required to travel to the space station. He calls his chief statistician (Adam) and asks him to do this.

Adam: *Umm, do you have any past reference?*

Husk: *I have sent payload through similar technology 10 times and those can be used as past data.*

Adam: *That's great. And I can assume fuel consumption is Normally distributed?*

Husk: *Sure.*

The past 10 fuel consumption samples in million gallons are: {5.3, 7.2, 6.2, 5.5, 6.9, 6.0, 4.3, 7, 5.2, 5.5}.

(a) Use MME to estimate the fuel consumption based on prior samples. (2 points)

(b) Husk also asks for confidence intervals of the estimate with α=0.05. (5 points)

(c) How many additional trips have to be made if Husk wants the confidence interval size to be less than or equal to 0.1? CI size is the difference between upper and lower limits of the CI. (3 points)

6. Histogram, meet the empirical distribution function (Total 5 points)

Let $X_1, X_2, \ldots, X_n$ be i.i.d. RVs with true CDF *F* with sample space [0, 1]. Let $\widehat{F}$ be the associated empirical distribution function. In class, we derived the histogram estimator for a bin B, as, say $\widehat{H}_B$. Let $\hat{f} = \widehat{H}_B/h$ be the empirical p.d.f. for the values in bin B with some bin size h. For some x in (0, 1), show that: $\hat{f}(x) \approx d\widehat{F}(\mathrm{x})$. Use the fact that the derivative of a function, g(), at x, is $\lim_{\Delta x \to 0} \frac{g(x+\Delta x)-g(x)}{\Delta x}$.

7. Programming fun with $\widehat{F}$ (Total 12 points)

For this question and the next, we require some programs/scripts. Feel free to use Python and the scripts provided on the class website as templates to build upon. Do not use any libraries or functions to bypass the programming effort. Please email all your code to the TA (Parth) with sufficient documentation so the code can be evaluated. Attach each plot as a separate sheet to your submission. All plots must be neat, legible (large fonts), with appropriate legends, axis labels, titles, etc.

(a) Write a program to plot $\widehat{F}$ given a list of samples as input. Your plot must have y-limits from 0 to 1, and x-limits from 0 to the largest sample. Also show the input points as Xs on the x-axis. (2 points)

(b) Use a binomial random number generator with range [0, 199] and p=1/2 to draw n=10, 100, 1000, 10000 samples. Feed these as input to (a) to generate four plots. What do you observe? (4 points)

(c) Modify (a) above so that it takes as input a collection of list of samples; that is, a 2-D array of sorts where each row is a list of samples (row=student). The program should now compute the average $\widehat{F}$ across the rows and plot it. That is, compute the $\widehat{F}$ for each row, average them all out, and plot the corresponding average of $\widehat{F}$. Show all input points as Xs on the x-axis. (2 points)

(d) Use the same binomial random number generator from (b) to draw n=10 samples for m=10, 100, 1000 rows. Feed these as input to (c) to generate three plots. What do you observe? (4 points)

8. Comparing CIs for $\widehat{F}$ **(Total 8 points)**

(a) Modify the program from 7 (a) to now also add 95% Normal-based CI lines for $\widehat{F}$, using the result from 4 (b), given a list of samples as input. Draw a plot showing $\widehat{F}$ and the CI lines for the q8.dat data file (799 samples) on the class website. Use x-limits of 0 to 2, and y-limits of 0 to 1. (4 points)

(b) Modify the program from 8 (a) to add 95% DKW-based CI lines for $\widehat{F}$. Draw a single plot showing $\widehat{F}$ and both sets of CI lines (Normal and DKW) for the q8.dat data. Which CI is tighter? (4 points)