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

Assignment 3: Non-Parametric Inference Due: 3/18, 1:15pm, via Blackboard

(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)

AMEYA SANKHE (1132 19562), SHUBHAM AGRAWAL (113166701), PRATIK NAGELIA (114122014), RANJAN KUMAR (1132 62 786)

1. MSE in terms of bias **(Total 5 points)**

For some estimator $\hat{\theta}$, show that MSE = bias$^2(\hat{\theta})$ + Var$(\hat{\theta})$. Show your steps clearly.

We Know, $MSE(\hat{\theta}) = E\left[(\hat{\theta}-\theta)^2\right]$

$= E\left[\hat{\theta}^2 - 2\hat{\theta}\theta + \theta^2\right]$

$= E\left[\hat{\theta}^2\right] - E\left[2\hat{\theta}\theta\right] + E\left[\theta^2\right]$ by linearity of Expectation $E(X+Y) = E(X) + E(Y)$

$= E\left[\hat{\theta}^2\right] - 2\theta E\left[\hat{\theta}\right] + \theta^2$ .... as θ = Constant, $E(\text{constant}) = \text{Constant}$

$= E\left[\hat{\theta}^2\right] - \left(E[\hat{\theta}]\right)^2 + \left(E[\hat{\theta}]\right)^2 - 2\theta \times E\left[\hat{\theta}\right] + \theta^2$

by adding and subtracting $\left(E[\hat{\theta}]\right)^2$

$= \left(E\left[\hat{\theta}^2\right] - \left(E[\hat{\theta}]\right)^2\right) + \left(\left(E[\hat{\theta}]\right)^2 - 2\theta \times E\left[\hat{\theta}\right] + \theta^2\right)$ ——— (1)

We know $Var(\hat{\theta}) = E(\hat{\theta}^2) - (E(\hat{\theta}))^2$

On putting the above value in Equation ①, we get :-

or, $MSE(\hat{\theta}) = Var(\hat{\theta}) + (E(\hat{\theta}) - \theta)^2$ ——— ②

..... $(a-b)^2 = a^2 - 2ab + b^2$

We know $Bias(\hat{\theta}) = E(\hat{\theta}) - \theta$ ——— ③

$\therefore = Var(\hat{\theta}) + (Bias(\hat{\theta}))^2$ { Putting the value of Equation ③ in Equation ② }

$MSE(\hat{\theta}) = Var(\hat{\theta}) + (Bias(\hat{\theta}))^2$

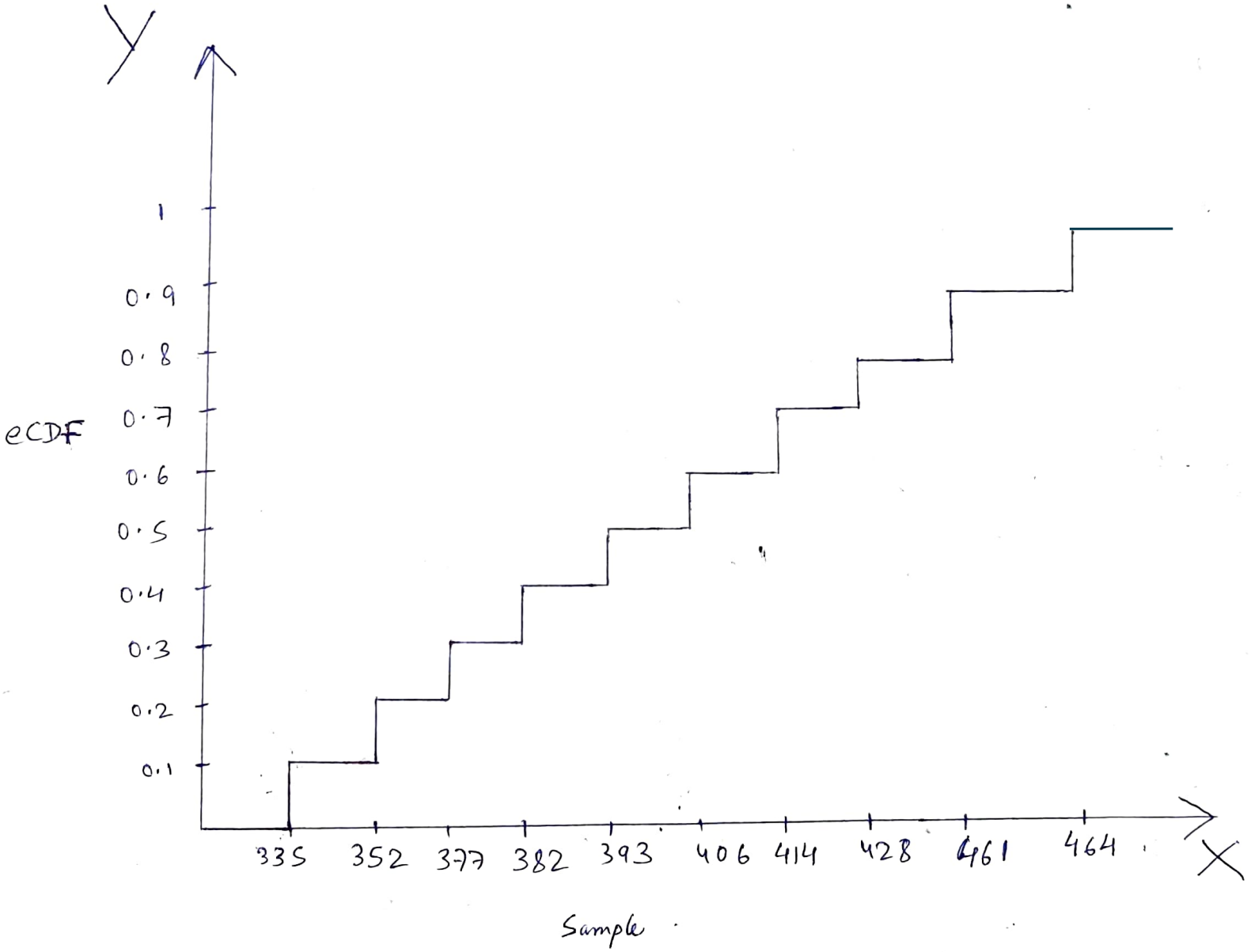
$$\boxed{MSE(\hat{\theta}) = Var(\hat{\theta}) + (Bias(\hat{\theta}))^2}$$

Hence Proved

2. Practice with empirical CDF (eCDF) **(Total 5 points)**

Using the first 10 samples from the collisions.csv file on the class website, carefully draw the eCDF by hand. Make sure the x- and y-axis clearly indicate the sample points and their corresponding eCDF. Your plot must have y-limits from 0 to 1, and x-limits from smallest sample to the largest sample.

Ans.



3. Programming fun with $\hat{F}$ (Total 15 points)

For this question, we require some programming; you should only use Python. You may use the scripts provided on the class website as templates. Do not use any libraries or functions to bypass the programming effort. Please submit your code as usual in your zip/tar file repo on BB. Provide sufficient documentation so the code can be evaluated. **Also attach each plot** as a separate sheet (or image) to your submission upload. All plots must be neat, legible (large fonts), with appropriate legends, axis labels, titles, etc.

(a) Write a program to plot $\hat{F}$ (empirical CDF or eCDF) 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. Show the input points as crosses on the x-axis. (2 points)

(b) Use an integer random number generator with range [1, 99] to draw n=10, 100, and 1000 samples. Feed these as input to (a) to **draw three plots**. What do you observe? (3 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 (as in (a)). The program should now compute the average $\hat{F}$ across the rows and plot it. That is, for a given x point, first compute the $\hat{F}$ for each row (student), then average them all out across rows, and plot the average $\hat{F}$ for x. Repeat for all input points, x. Show all input points as crosses on the x-axis. (2 points)

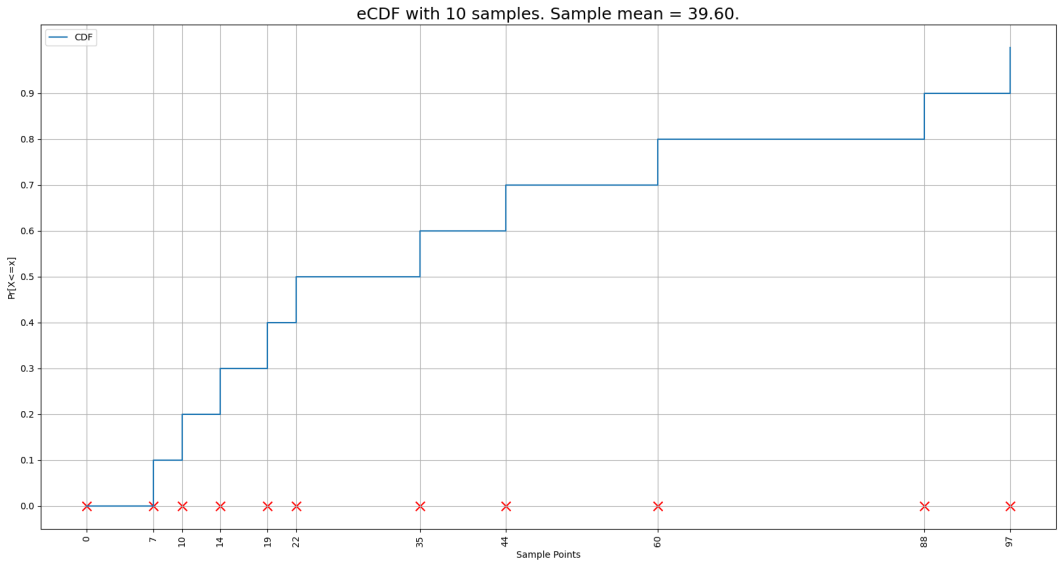
(d) Use the same integer random number generator from (b) to draw n=10 samples for m=10, 100, 1000 rows. Feed these as input to (d) to **draw three plots**. What do you observe? (3 points)

(e) Modify the program from (a) to now also add 95% Normal-based CI lines for $\hat{F}$, given a list of samples as input. **Draw a plot** showing $\hat{F}$ and the CI lines for the a3_q3.dat data file (799 samples) on the class website. Use x-limits of 0 to 2, and y-limits of 0 to 1. (2 points)

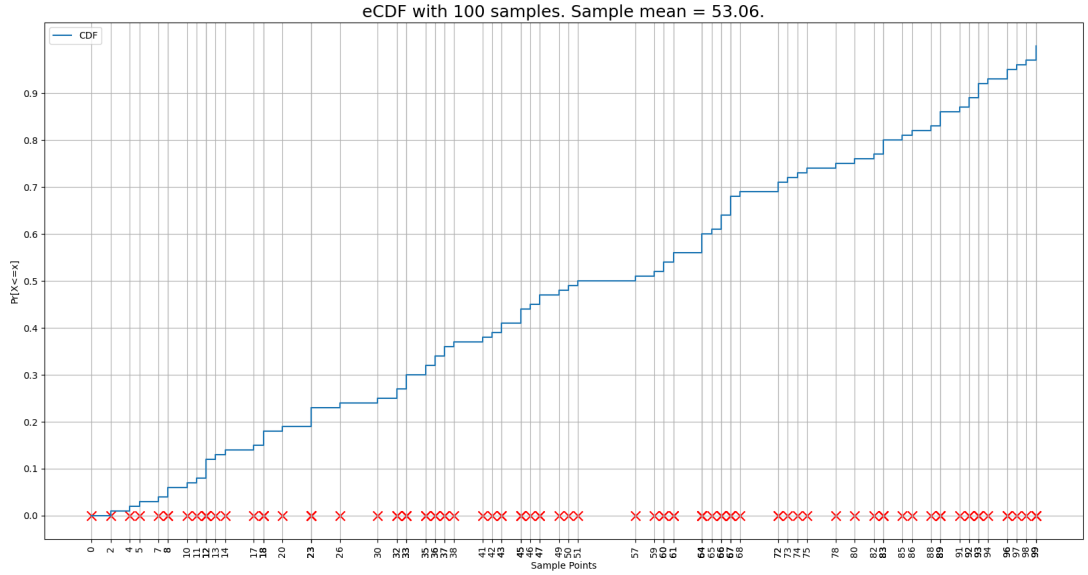
(f) Modify the program from (e) to also add 95% DKW-based CI lines for $\hat{F}$. **Draw a single plot** showing $\hat{F}$ and both sets of CI lines (Normal and DKW) for the a3_q3.dat data. Which CI is tighter? (3 points)

Q3) Filename : a3_q3.py
b)

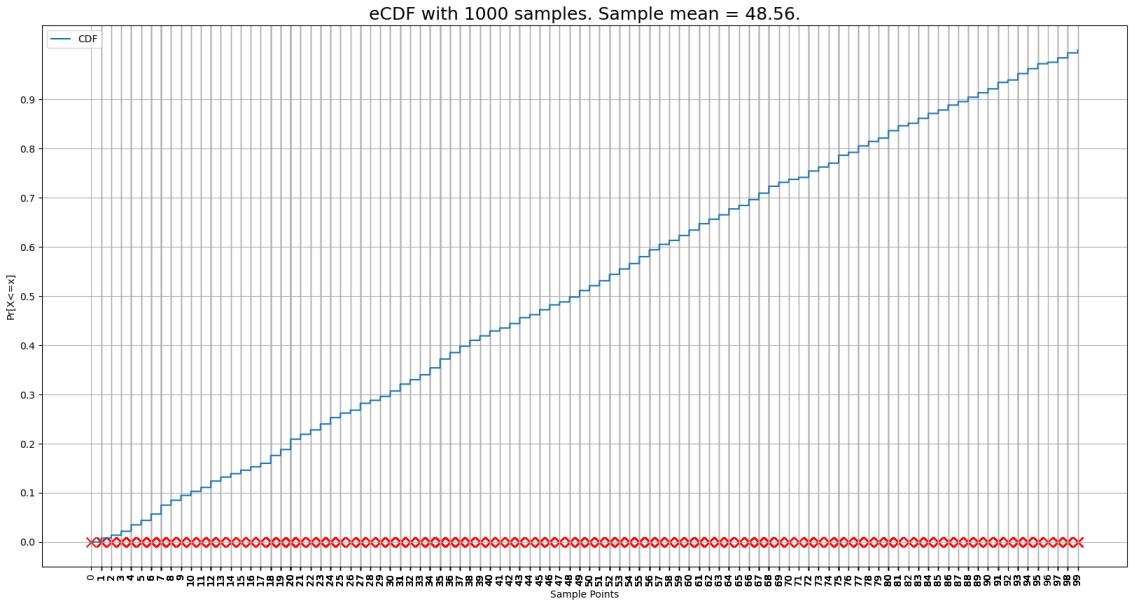
Filename = ecdf10.png



filename = ecdf100.png



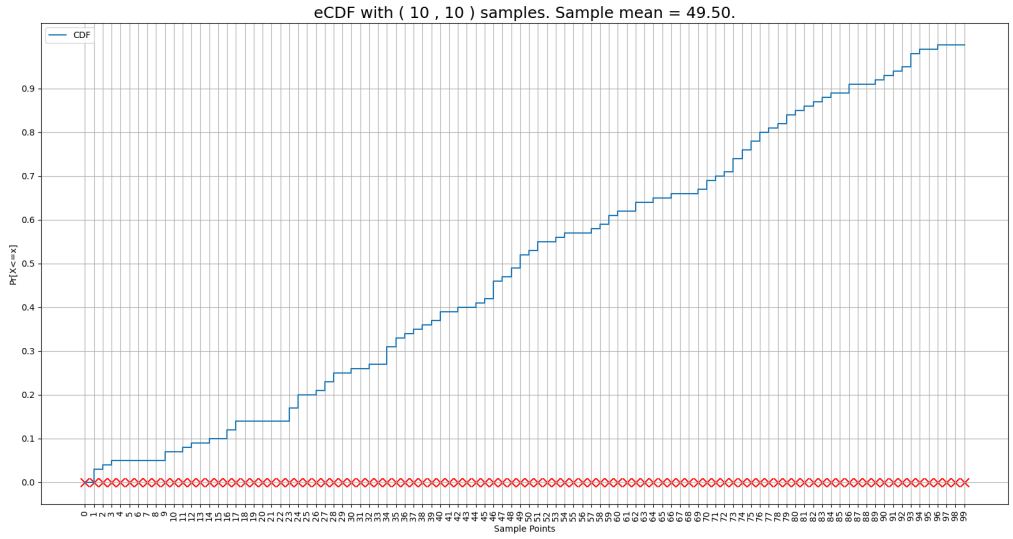
Filename = ecdf1000.png



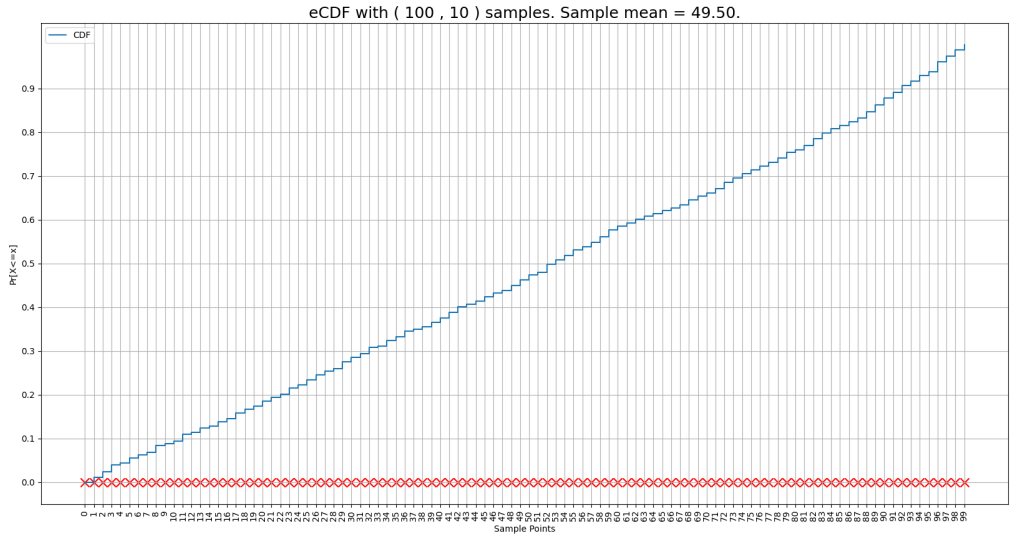
As the number of sample increases (for n = 10, 100, 1000), the step function tends to look like a straight line with a positive slope

d)

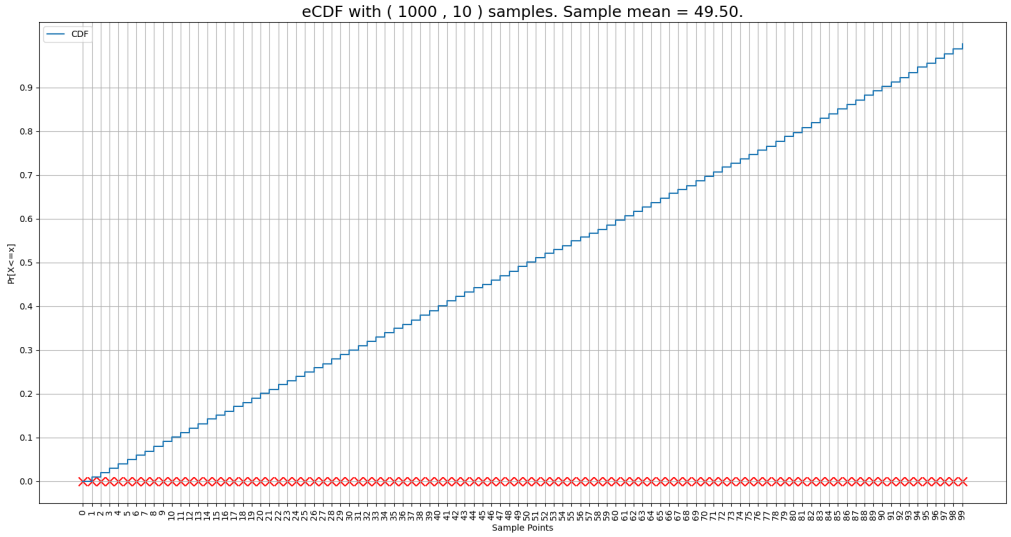
Filename : ecdf(10, 10)samples.png



Filename : ecdf(100, 10)samples.png

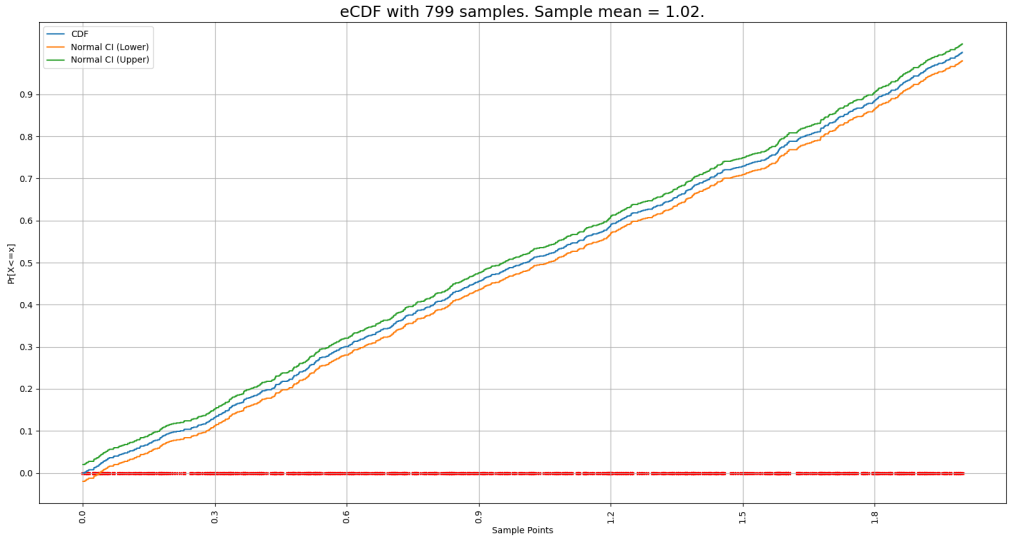


Filename : ecdf(1000, 10)samples.png

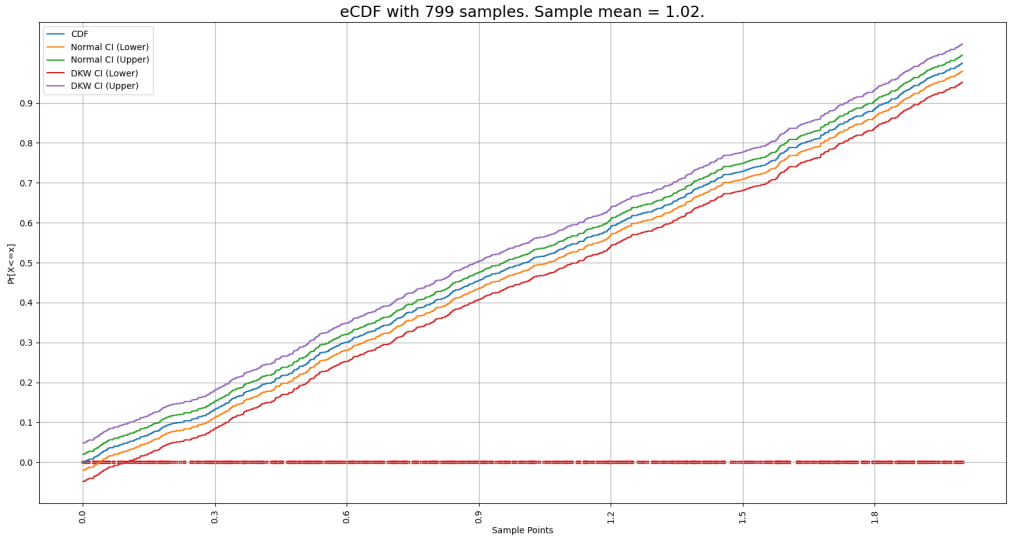


d) As the number of sample increases (for n = 10, 100, 1000), the step function tends to look like a straight line with a positive slope

e)filename : ecdf799.png



f) The Normal CI is Tighter
filename : ecdf799-dkw.png



4. 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, $\bar{X}_n = \frac{1}{n}\sum_{i=1}^{n} X_i$. (2 points)

(b) Show that the bias of $\hat{\sigma}^2$ is $-\sigma^2/n$, where σ^2 is the true variance. (3 points)

(c) The kurtosis for a RV X with mean μ and variance σ^2 is defined as $Kurt[X] = E[(X-\mu)^4] / \sigma^4$. Derive the plug-in estimate of the kurtosis in terms of the sample data. (3 points)

(d) The plug-in estimator idea also extends to two RVs. Consider $\rho = E[XY] - E[X]E[Y]/\sigma_X\sigma_Y$, where σ_X is the standard deviation for RV X. Assuming n i.i.d. observations for X and Y that appear in pairs as $\{(X_1, Y_1), (X_2, Y_2), \ldots, (X_n, Y_n)\}$, derive the plug-in estimator for ρ. (Hint: What is the ePMF for the event $X=X_1$ AND $Y=Y_1$? What about for the event $X=X_1$ AND $Y=Y_2$?) (2 points)

a)

$$Var(X) = E(X^2) - (E(X))^2$$

Let $\sigma^2 = Var(X)$

$$\therefore \sigma^2 = \sum_{\alpha \in \Omega} \alpha^2 \times P_X(\alpha) - \left(\sum_{\alpha \in \Omega} \alpha P_X(\alpha)\right)^2$$

Let Data D be $\{X_1, X_2 \ldots, X_n\} \sim$ i.i.d from true distribution

For Plug-in estimation

$$\hat{\sigma}^2 = \sum_{i=1}^{n} X_i^2 \times \hat{P_X}(X_i) - \left(\sum_{i=1}^{n} X_i \times \hat{P_X}(X_i)\right)^2$$

We know $\hat{P_X}(X_i) = \hat{P_X}(X = X_i) = \frac{1}{n}$

$$\hat{\sigma}^2 = \sum_{i=1}^{n} \frac{X_i^2}{n} - \left(\sum_{i=1}^{n} \frac{X_i}{n}\right)^2$$

$$= \sum_{i=1}^{n} \frac{X_i^2}{n} - (\bar{X}_n)^2 \quad \ldots\ldots \quad \bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - n(\bar{X}_n)^2\right) \quad \text{——— ①}$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - 2n(\bar{X}_n)^2 + n(\bar{X}_n)^2\right)$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - 2n\,\bar{X}_n\,\bar{X}_n + \sum_{i=1}^{n}(\bar{X}_n)^2\right) \quad \ldots\ldots \sum_{i=1}^{n}(\bar{X}_n)^2 = n(\bar{X}_n)^2$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - 2n\,\frac{\sum_{i=1}^{n} X_i}{n} \times \bar{X}_n + \sum_{i=1}^{n}(\bar{X}_n)^2\right) \quad \ldots\ldots \bar{X}_n = \frac{\sum_{i=1}^{n} X_i}{n}$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - 2\sum_{i=1}^{n} X_i \times \bar{X}_n + \sum_{i=1}^{n}(\bar{X}_n)^2\right)$$

$$= \frac{1}{n}\left(\sum_{i=1}^{n} \left(X_i^2 - 2\,X_i\bar{X}_n + \bar{X}_n^2\right)\right) \quad \ldots\ldots \Sigma a + \Sigma b = \Sigma(a+b)$$

$$\therefore \hat{\sigma}^2 = \frac{1}{n}\left(\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right)$$

Which is Sample Variance

..... Hence Proved

b) To Prove $\text{bias}(\hat{\sigma}^2) = -\sigma^2/n$

We know $\text{bias}(\hat{\sigma}^2) = E(\hat{\sigma}^2) - \sigma^2$ ——— (a)

From (1) we know

$$\hat{\sigma}^2 = \frac{1}{n}\left(\sum_{i=1}^{n} X_i^2 - n(\bar{X}_n)^2\right)$$

Taking Expectation on Both Sides

$$E(\hat{\sigma}^2) = \frac{1}{n} E\left(\sum_{i=1}^{n} X_i^2 - n(\bar{X}_n)^2\right)$$

$$= \frac{1}{n} \times \left(E\left(\sum_{i=1}^{n} X_i^2\right) - n E(\bar{X}_n)^2\right) \quad \text{..... Linearity of Expectation. } E(a+b) = E(a) + E(b)$$

$$= \frac{1}{n} \left(\sum E(X_i^2) - n E(\bar{X}_n)^2\right) \quad \text{..... Linearity of Expectation } E(\sum x) = \sum E(x)$$

——— (2)

We know

$$Var(X) = E(X^2) - (E(X))^2$$

$$\therefore E(X^2) = Var(X) + (E(X))^2$$

$$E(X^2) = \sigma^2 + \mu^2 \quad \text{..... as } X \text{ is taken from true distribution with mean } \mu \text{ and variance } \sigma^2.$$

——— (3)

We also know

$$Var(\bar{X}_n) = Var\left(\sum_{i=1}^{n} \frac{X_i}{n}\right)$$

$$= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} X_i\right)$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} Var(X_i) \quad \text{..... by Linearity of Variance as } X_i \text{ are independent}$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \sigma \quad \text{..... as } X_1 \dots X_n \text{ are iid taken from true distribution } X \text{ with Variance } \sigma. \text{ Therefore } Var(X_i) = \sigma$$

$$Var(\bar{X}_n) = \frac{n\sigma}{n^2}$$

$$Var(\bar{x}_n) = \frac{\sigma^2}{n} \qquad \text{——— (4)}$$

Also $E(\bar{x}_n) = E\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)$

$$= \frac{1}{n} E\left(\sum_{i=1}^{n} x_i\right)$$

$$= \frac{1}{n} \sum_{i=1}^{n} E(x_i) \quad \text{.... by Linearity of Expectation}$$

$$= \frac{1}{n} \sum_{i=1}^{n} \mu \quad \text{..... as } x_1 \ldots x_n \text{ are iid from } X \text{ which has true mean } \mu. \; E(x_i) = \mu$$

$$= \frac{n\mu}{n}$$

$$E(\bar{x}_n) = \mu \qquad \text{——— (5)}$$

Now we know,

$$E(\bar{x}_n^2) = Var(\bar{x}_n) + (E(\bar{x}_n))^2 \quad \text{.... as } Var(\bar{x}_n) = E(\bar{x}_n^2) - (E(\bar{x}_n))^2$$

$$E(\bar{x}_n^2) = \frac{\sigma^2}{n} + \mu^2 \quad \text{.... from (4) and (5)}$$

——— (6)

Now putting (6) and (3) in. (2)

$$E(\hat{\sigma}^2) = \frac{1}{n}\left(\sum_{i=1}^{n} E(x_i^2) - n E((\bar{x}_n)^2)\right)$$

After putting (6) and (3)

$$E(\hat{\sigma}^2) = \frac{1}{n}\left(\sum_{i=1}^{n} (\sigma^2 + \mu)^2 - n\left(\frac{\sigma^2}{n} + \mu^2\right)\right)$$

$$= \frac{1}{n}\left(n\sigma^2 + n\mu^2 - \frac{n\sigma^2}{n} - n\mu^2\right)$$

$$E(\hat{\sigma}^2) = \frac{1}{n}\left(n\sigma^2 - \sigma^2\right) = \left(\frac{n-1}{n}\right)\sigma^2 \quad \text{——— (7)}$$

put (7) in (a)

$$\text{bias}(\hat{\sigma}^2) = E(\hat{\sigma}^2) - \sigma^2$$

• Now put (7)

$$\text{bias}(\hat{\sigma}^2) = \left(\frac{n-1}{n}\right)\sigma^2 - \sigma^2$$

$$= \frac{(n-1)\sigma^2 - n\sigma^2}{n}$$

$$= \frac{n\sigma^2 - \sigma^2 - n\sigma^2}{n}$$

$$\therefore \text{bias}(\hat{\sigma}^2) = -\frac{\sigma^2}{n}$$

Hence Proved

c) Let $K = \text{Kurt}[X] = E[(X-\mu)^4]/\sigma^4$

$\therefore$ Plug in estimator of K

$$\hat{K} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)\, P[(X-\bar{X}_n)^4]}{\hat{\sigma}^4} \quad \text{as } E(x) = \Sigma x P(x) \quad \text{——— (8)}$$

$$P[(X-\bar{X}_n)^4] = \frac{1}{n} \quad \text{——— (9)}$$

we also know $\hat{\sigma}^2 = \frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2$

$$\therefore \hat{\sigma}^4 = \left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right)^2 \quad \text{——— (10)}$$

putting (9) and (10) in (8)

$$\hat{K} = \frac{\sum_{i=1}^{n}(X_i - \bar{X}_n)\, P\left[(X - \bar{X}_n)^4\right]}{\hat{\sigma}^4}$$

Now put (9) and (10)

$$\hat{K} = \frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^4}{\left(\frac{1}{n}\sum_{i=1}^{n}(X_i - \bar{X}_n)^2\right)^2}$$

$$\therefore \hat{K} = \frac{n\sum_{i=1}^{n}(X_i - \bar{X})^4}{\left(\sum_{i=1}^{n}(X_i - \bar{X})^2\right)^2}$$

d) Let $\rho = \frac{E[XY] - E[X]E[Y]}{\sigma_x \times \sigma_y}$

For Plug in Estimation

$$\hat{\rho} = \frac{\sum_{i=1}^{n} X_i Y_i \hat{P}_{XY}(X_i Y_i) - \sum X_i \hat{P}_x(X_i) \sum Y_i \hat{P}(Y_i)}{\sigma_x \times \sigma_y}$$

Now for Plug in estimation

$$= \sum_{i=1}^{n} X_i Y_i \hat{P}_{XY}(X_i Y_i) - \frac{\sum X_i \hat{P}_X(X_i) \sum Y_i \hat{P}_Y(Y_i)}{\hat{\sigma}_x \times \hat{\sigma}_y}$$

$$= \sum_{i=1}^{n} \frac{X_i Y_i}{n} - \frac{\sum_{i=1}^{n} X_i}{n} \times \frac{\sum_{i=1}^{n} Y_i}{n}$$

$$\frac{1}{\sqrt{n}} \times \sqrt{\sum (X_i - \bar{X}_n)^2} \times \frac{1}{\sqrt{n}} \times \sqrt{\sum (Y_i - Y_n)^2}$$

..... $\hat{P}_{XY}(X_i Y_i) = \hat{P}_X(X_i) = \hat{P}_Y(Y_i)$

$$= \frac{1}{n}$$

and $\hat{\sigma}_x = \sqrt{\frac{\sum (X_i - \bar{X}_n)^2}{n}}$

$$= \sum_{i=1}^{n} \frac{X_i Y_i}{n} - \frac{\left(\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i\right) \times n}{n^2 \times \sqrt{\sum (X_i - \bar{X}_n)^2 \sum (Y_i - \bar{Y}_n)^2}}$$

$$\hat{\rho} = \sum_{i=1}^{n} \frac{X_i Y_i}{n} - \frac{\sum_{i=1}^{n} X_i \sum_{i=1}^{n} Y_i}{n \sqrt{\sum (X_i - \bar{X}_n)^2 \sum (Y_i - \bar{Y}_n)^2}}$$

... which is plug in estimator of ρ

5. Consistency of eCDF **(Total 10 points)**

Let D={$X_1, X_2, \ldots, X_n$} be a set of i.i.d. samples with true CDF F. Let $\hat{F}$ be the eCDF for D, as defined in class.

(a) Derive $E(\hat{F})$ in terms of F. Start by writing the expression for $\hat{F}$ at some α. (3 points)

(b) Show that bias($\hat{F}$) = 0. (2 points)

(c) Derive se($\hat{F}$) in terms of F and n. (3 points)

(d) Show that $\hat{F}$ is a consistent estimator. (2 points)

Q5.(a) Let $\hat{F}$ be the eCDF for D.

$$\therefore \hat{F} = F_x\hat{}(\alpha) = \hat{P}(X \leq \alpha) = \frac{\sum_{i=1}^{n} I(X_i \leq \alpha)}{n}$$

Taking Expectation on both sides, we get,

$$E(\hat{F}) = E\left[\frac{\sum_{i=1}^{n} I(X_i \leq \alpha)}{n}\right]$$

$$= \frac{1}{n} E\left[\sum_{i=1}^{n} I(X_i \leq \alpha)\right]$$

Taking constant out

$$\overset{\text{LoE}}{=} \frac{1}{n} \sum_{i=1}^{n} E[I(X_i \leq \alpha)]$$

By Linearity of Expectation.

$$\overset{\text{iid}}{=} \frac{1}{n} \times n \; E[I(X_1 \leq \alpha)]$$

[∵ If $X \sim$ Indicator (E), then $E[X] = Pr(E)$]

$$= E[I(X_1 \leq \alpha)]$$

$$= P_r(X_1 \leq \alpha)$$

[$P_r(X_1 \leq \alpha) = Pr(X \leq \alpha)$]

$X_1 \sim \ldots X_n \} \overset{iid}{\sim} X$

$$\doteq P_r(X \leq \alpha)$$

$$= F_X(\alpha)$$

$$\therefore E[\hat{F}] = F_X(\alpha) \quad \text{Ans}$$

5

(b) $\text{bias}(\hat{F}) = 0$ ⇐ To Prove.

We know, $\text{bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$

where $\hat{F}$ = eCDF

$F = F_X(\alpha)$
= True CDF

$\therefore \text{bias}(\hat{F}) = E(\hat{F}) - F_*(\alpha)$

$= F_X(\alpha) - F_X(\alpha)$

$= 0$

[∵ From previous part, $E(\hat{F}) = F_X(\alpha)$]

(c) Standar Error

$se(\hat{\theta}) = \sqrt{Var(\hat{\theta})}$

$\therefore se(\hat{F}) = \sqrt{Var(\hat{F})}$

$\hat{F} = \frac{\sum_{i=1}^{n} I(X_i \leq \alpha)}{n}$

Taking Variance on both sides,

$Var(\hat{F}) = Var\left(\frac{\sum_{i=1}^{n} I(X_i \leq \alpha)}{n}\right)$

$= \frac{1}{n^2} Var\left(\sum_{i=1}^{n} I(X_i \leq \alpha)\right)$ [Taking Constant Out]

$\overset{LOV}{=} \frac{1}{n^2} * Var\left(I(X_1 \leq \alpha)\right)$

$= \frac{Var(I(X_1 \leq \alpha))}{n} =$

$= \frac{P_X(X \leq \alpha)\left(1 - P_Y(X \leq \alpha)\right)}{n}$

[By property of iid.
$P(X_1 \leq \alpha) = P(X \leq \alpha)$
Var of Indicator Variable
$= P(1-P)$
where P is the Probability of Event]

$\therefore se(\hat{F}) = \sqrt{\frac{F_X(\alpha) * (1 - F_X(\alpha)}{n}}$ Ans

5)

(d) As per theorem,

If $\text{bias}(\hat{\theta}) \xrightarrow{n \to \infty} 0$ and $se(\hat{\theta}) \xrightarrow{n \to \infty} 0$.

the $\hat{\theta}$ is a consistent estimator of θ.

In part B we proved $\text{bias}(\hat{F}) = 0$

In part C, we saw

$$se(\hat{F}) = \sqrt{\frac{F_x(\alpha)\ (1 - F_x(\alpha))}{n}}$$

If $n \to \infty$, $se(\hat{F}) = 0$

∴ Therefore $\hat{F}$ is a consistent estimator. Hence Proved.

6. Properties of estimators **(Total 10 points)**

(a) Find the bias, se, and MSE in terms of θ for $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$, where X_i are i.i.d. ~ Bernoulli(θ). Hint: Follow the same steps as in class, assuming the true distribution is unknown. Only at the end use the fact that the unknown distribution is Bernoulli(θ) to get the final answers in terms of θ. (5 points)

(b) Derive the Normal-based (1-α) CI for $\hat{\theta}$. Explain why Normal-based CIs are applicable here.(5 points)

6 (a) $\text{Bias}(\hat{\theta}) = E[\hat{\theta}] - \theta$

$\Rightarrow \text{Bias}(\hat{\theta}) = E\left[\frac{1}{n}\sum_{i}^{n} X_i\right] - \theta$

$\Rightarrow \text{Bias}(\hat{\theta}) \overset{\text{LoE}}{=} \frac{1}{n}\sum_{i}^{n} E[X_i] - \theta$

$\Rightarrow \text{Bias}(\hat{\theta}) = \frac{1}{n} \times n\theta - \theta$ [Replacing expectation of Bernoulli]

$\Rightarrow \text{Bias}(\hat{\theta}) = 0$ Ans.

$\text{Var}(\hat{\theta}) = \text{Var}\left(\frac{1}{n}\sum_{i=1}^{n} X_i\right)$

$\Rightarrow \text{Var}(\hat{\theta}) = \frac{1}{n^2}\sum \text{Var}(X_i)$ $\because X_i$ are iid

$= \frac{1}{n^2} \times n \times \theta(1-\theta) = \frac{\theta(1-\theta)}{n}$ [Replacing variance of Bernoulli]

$\text{se}(\hat{\theta}) = \sqrt{\text{Var}(\hat{\theta})} = \sqrt{\frac{\theta(1-\theta)}{n}}$ Ans. ——— (1)

$\text{MSE}(\hat{\theta}) = \text{Bias}^2(\hat{\theta}) + \text{Var}(\hat{\theta})$ [From Question]

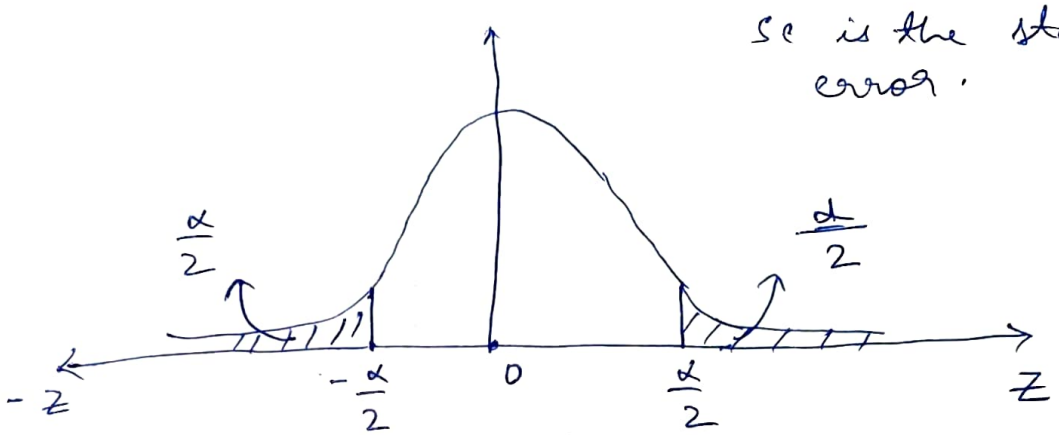
$= 0 + \frac{\theta(1-\theta)}{n}$

$= \frac{\theta(1-\theta)}{n}$ Ans.

(b) Ans. On using $\hat{\theta}$ = Normal (θ, se^2), for converting this into Standard Normal, we have :-

$$Z = \frac{\hat{\theta} - \theta}{se} \quad \text{---------- (1)}$$

where θ is the mean and se is the standard error.



From the symmetric Nature of the graph above :-

$$Pr(Z > Z_{\frac{\alpha}{2}}) = \frac{\alpha}{2} \quad ; \quad Pr(Z < Z_{-\frac{\alpha}{2}}) = \frac{\alpha}{2}$$

$$\Rightarrow Pr\left[Z \in \left(-Z_{\frac{\alpha}{2}}, Z_{\frac{\alpha}{2}}\right)\right] = 1 - \frac{\alpha}{2} - \frac{\alpha}{2}$$

$$= 1 - \alpha$$

$$\Rightarrow Pr\left[\frac{\hat{\theta} - \theta}{se} \in \left[-Z_{\frac{\alpha}{2}}, Z_{\frac{\alpha}{2}}\right)\right] = 1 - \alpha$$

$$\Rightarrow Pr\left[\theta \in \left(\hat{\theta} - Z_{\frac{\alpha}{2}} \cdot se, \hat{\theta} + Z_{\frac{\alpha}{2}} \cdot se\right)\right] = 1 - \alpha$$

{ on Multiplying by se in $\pm Z_{\frac{\alpha}{2}}$ and subtracting $\hat{\theta}$ we get, $\theta \in (\hat{\theta} \mp Z_{\frac{\alpha}{2}} \cdot se)$ }

$$\therefore (1-\alpha)\ CI \text{ for } \hat{\theta} = \left\{\hat{\theta} - Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{\theta(1-\theta)}{n}},\ \hat{\theta} + Z_{\frac{\alpha}{2}} \sqrt{\frac{\theta(1-\theta)}{n}}\right\}$$

(using $se = \sqrt{\frac{\theta(1-\theta)}{n}}$ from Equation (1))

It is given that X_i are i.i.d, hence by Central limit theorem when independent random variables are added their sum tends towards a normal distribution even if variables are not normally distributed. For that reason Normal Based Confidence Intervals are applicable here.

7. Kernel density estimation (Total 15 points)

This question asks you to implement Kernel density estimator (KDE) from scratch and evaluate it for a sample dataset, a3_q7.csv. Do not use inbuilt KDE functions. But, you can use inbuilt pdf functions to estimate pdf at a point. The formal definition of KDE, which estimates pdf, is:

$$\widehat{f_h}(x) = \frac{1}{nh}\sum_{i=1}^{n} K\left(\frac{x-x_i}{h}\right) \quad (1)$$

where K(.) is called the kernel function which should be a smooth, symmetric and a valid density function. Parameter h > 0 is called the smoothing bandwidth that controls the amount of smoothing.

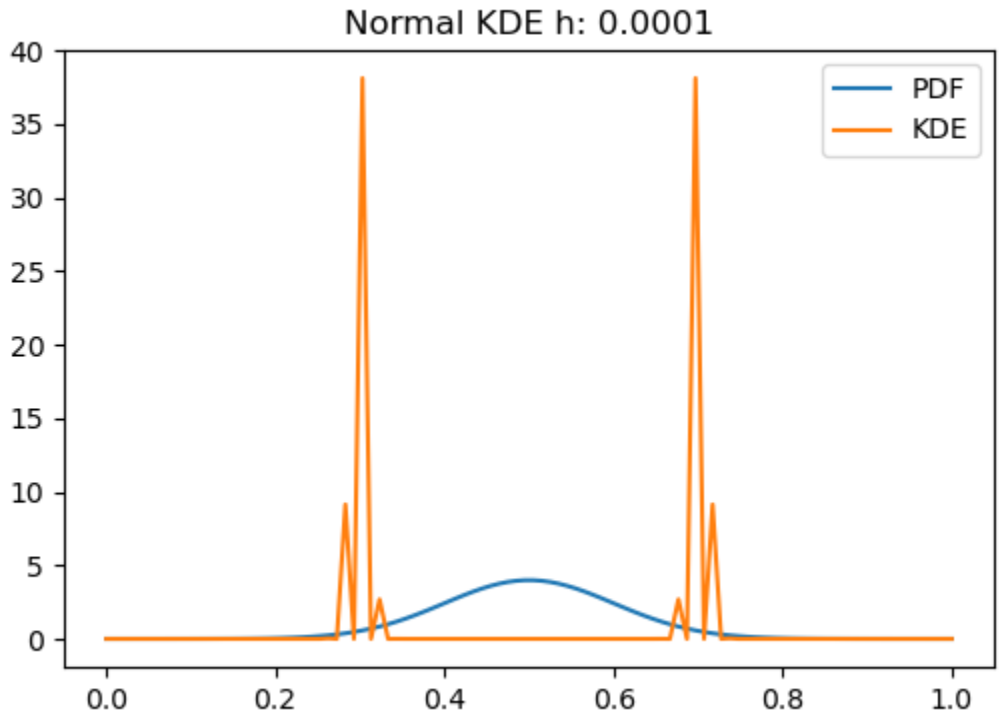
(a) For the a3_q7.csv dataset, the true distribution is Normal(0.5, 0.01) (the mean value μ is 0.5 , the variance σ^2 is 0.01). The task here is to implement a KDE function using the Normal distribution as the kernel, ***normal_kde(x,h,D)*** in python, where x is the point at which the pdf is to be estimated, h is the bandwidth and D is the list of data points. Implement the function as normal_kde.py by first computing $K\left(\frac{x-x_i}{h}\right)$ for all data points x_i in given dataset, where K(u) is the pdf of the standard Normal at point **u** = $\frac{x-x_i}{h}$, and then summing up all K() values and dividing by nh, where n is number of data points, as in Equation (1) above. Submit your code. (3 points)

(b) Obtain the p.d.f. for x = {0, 0.01, 0.02, ...,1} and compute the sample mean and sample variance (use result of Q4(a) as needed) for h=0.0001, 0.0005, 0.001, 0.005, 0.05. Report the deviation (as a percentage difference with respect to true mean or variance) of the estimates from the original distribution (Normal(0.5, 0.01)) in each of the 5 cases. Show on a single plot the pdf of the original Normal and the KDE estimates of the pdf for all 5 bandwidths. Include this plot in your submission. Which of the h values performs best? (6 points)

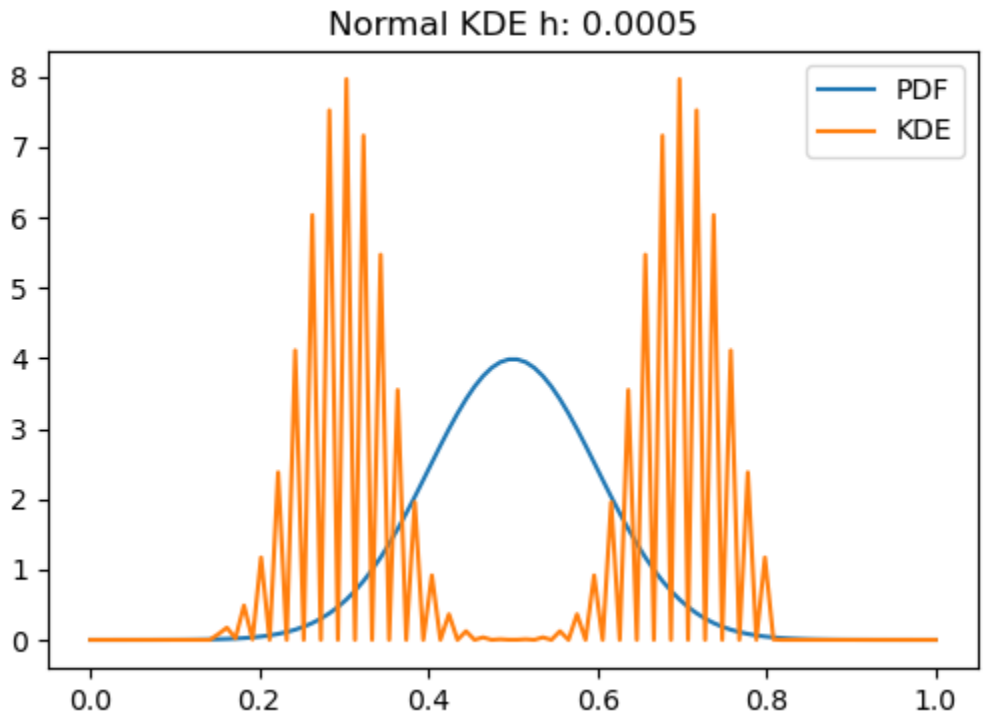
(c) Repeat (a) and (b) above when using the uniform kernel (implement as ***uniform_kde(x,h,D)*** as uniform_kde.py) with the function **K(u)** described as $\mathrm{K(u)} = \begin{cases} ½ & for -1 \le u \le 1 \\ 0 & otherwise \end{cases}$, where **u** = $\frac{x-x_i}{h}$, and Triangular distribution, ***triangular_kde(x,h,D)*** (implement as triangular_kde.py), using triangle kernel described as **K(u)** = 1-|u| for |u| ≤ 1 (and K(u) = 0 otherwise), where **u** = $\frac{x-x_i}{h}$. Repeat all parts of (b) for these two kernels for all 5 bandwidth values and report the percentage deviation from original mean and variance, plot the KDE estimates, and report the best bandwidth for each kernel choice. (6 points)

Q7) File Name : a3_q7.py

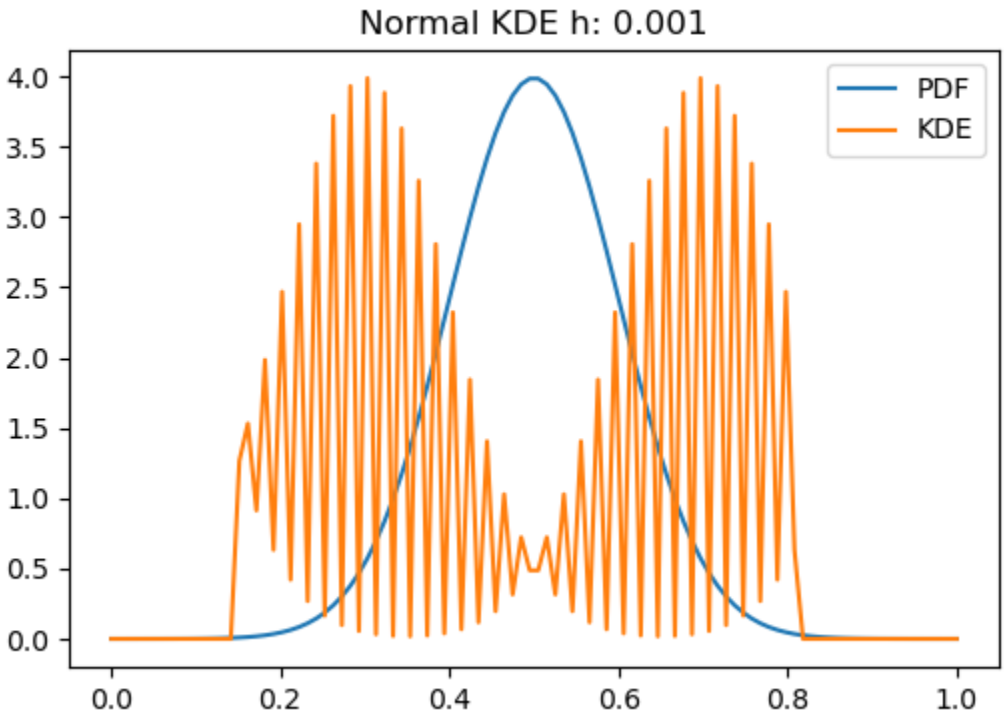
****
Normal KDE
H : 0.0001
Sample Mean: 0.9992336886836867 Sample Variance: 29.852148447119685
True Mean: 0.9899995686339825 True Variance: 1.8126392926634731
Percentage Deviation from Mean: 0.9327398053764338
Percentage Deviation from Variance: 1546.8885215025464



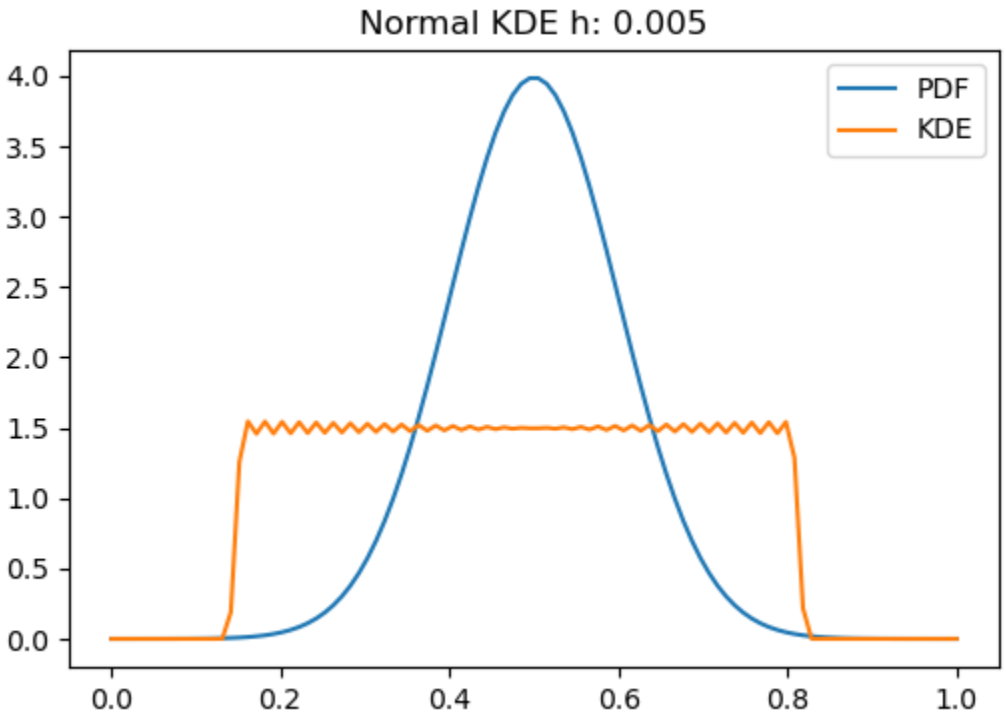
****
Normal KDE
H : 0.0005
Sample Mean: 0.9831549930837393 Sample Variance: 4.616210220891831
True Mean: 0.9899995686339825 True Variance: 1.8126392926634731
Percentage Deviation from Mean: -0.6913715689480007
Percentage Deviation from Variance: 154.66788894931327



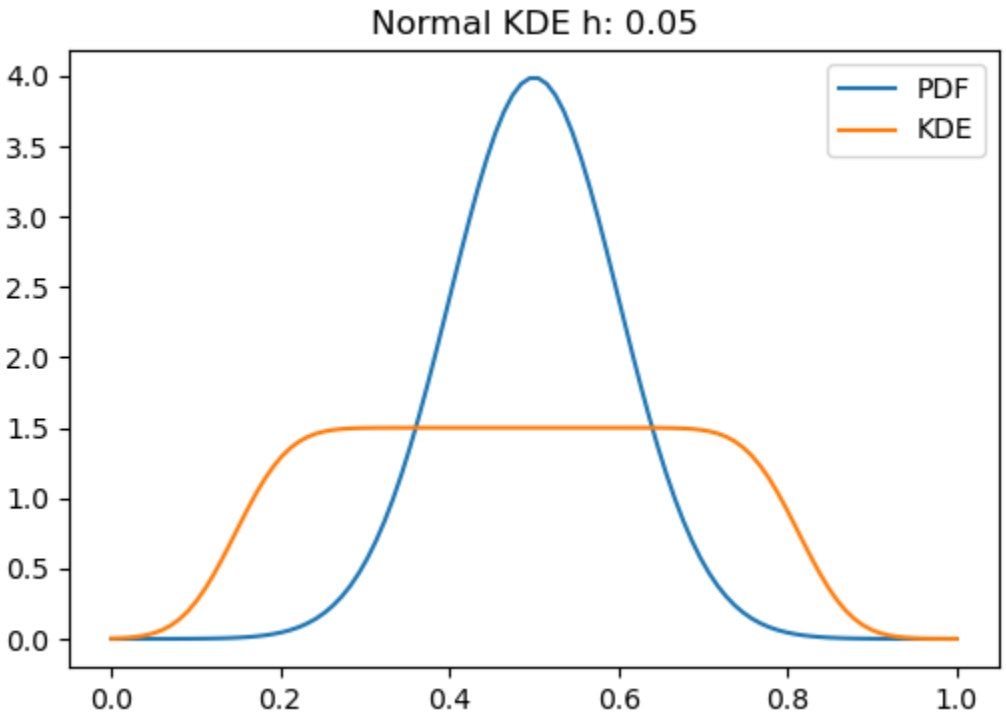
****
Normal KDE
H : 0.001
Sample Mean: 0.942803519158494 Sample Variance: 1.830197110986151
True Mean: 0.9899995686339825 True Variance: 1.8126392926634731
Percentage Deviation from Mean: -4.767279802011467
Percentage Deviation from Variance: 0.9686327772845867



****
Normal KDE
H : 0.005
Sample Mean: 0.9891703384046019 Sample Variance: 0.49416231868421945
True Mean: 0.9899995686339825 True Variance: 1.8126392926634731
Percentage Deviation from Mean: -0.08376066572683714
Percentage Deviation from Variance: -72.73796719047712

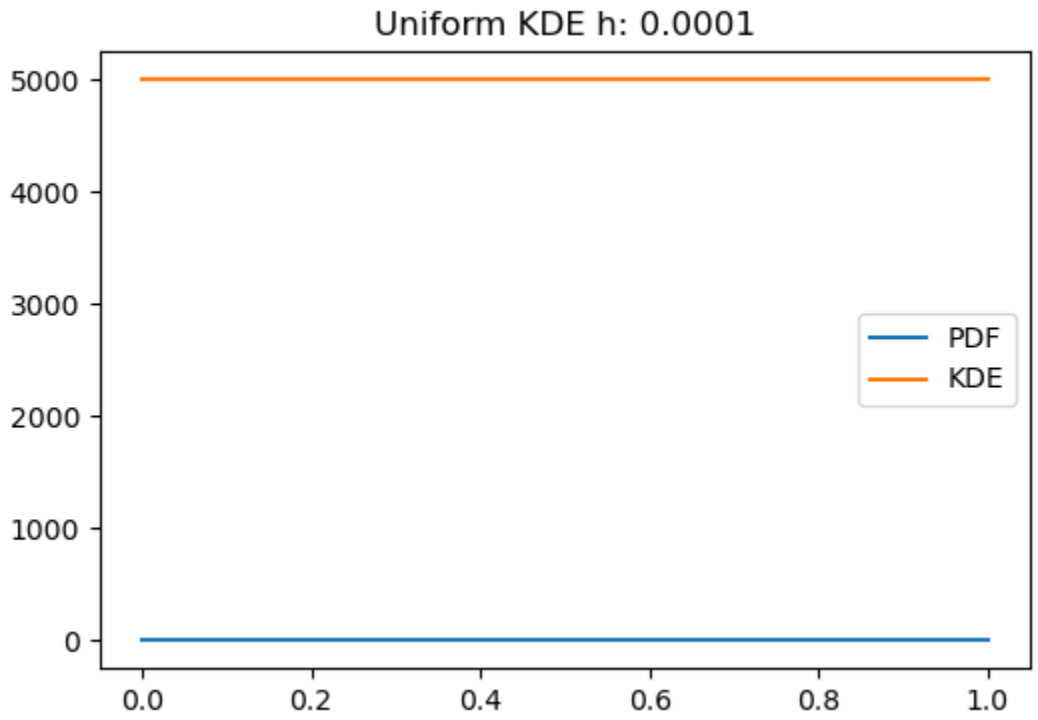


****
Normal KDE
H : 0.05
Sample Mean: 0.9899746161164065 Sample Variance: 0.3794055986346408
True Mean: 0.9899995686339825 True Variance: 1.8126392926634731
Percentage Deviation from Mean: -0.002520457419036049
Percentage Deviation from Variance: -79.06888589636904

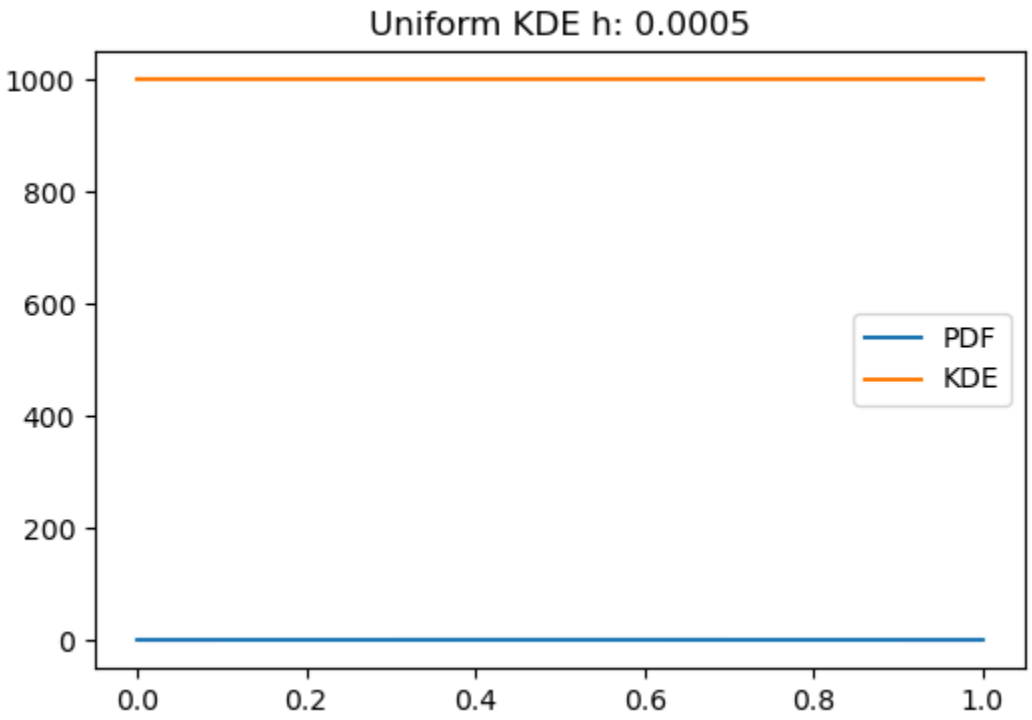


Best MSE for Normal KDE = 0.3794055992572689 with H = 0.05

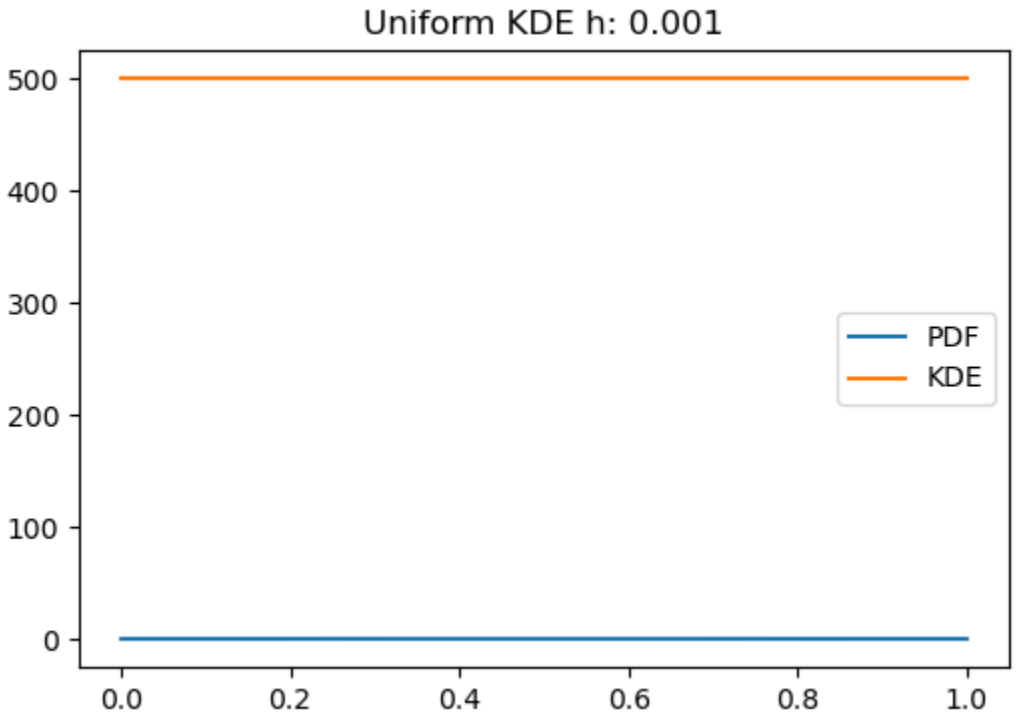
c)
****
Uniform KDE
H : 0.0001
Sample Mean: 5000.0 Sample Variance: 0.0
True Mean: 0.3826133630634787 True Variance: 0.000206429233868056
Percentage Deviation from Mean: 1306702.2402475418
Percentage Deviation from Variance: -100.0



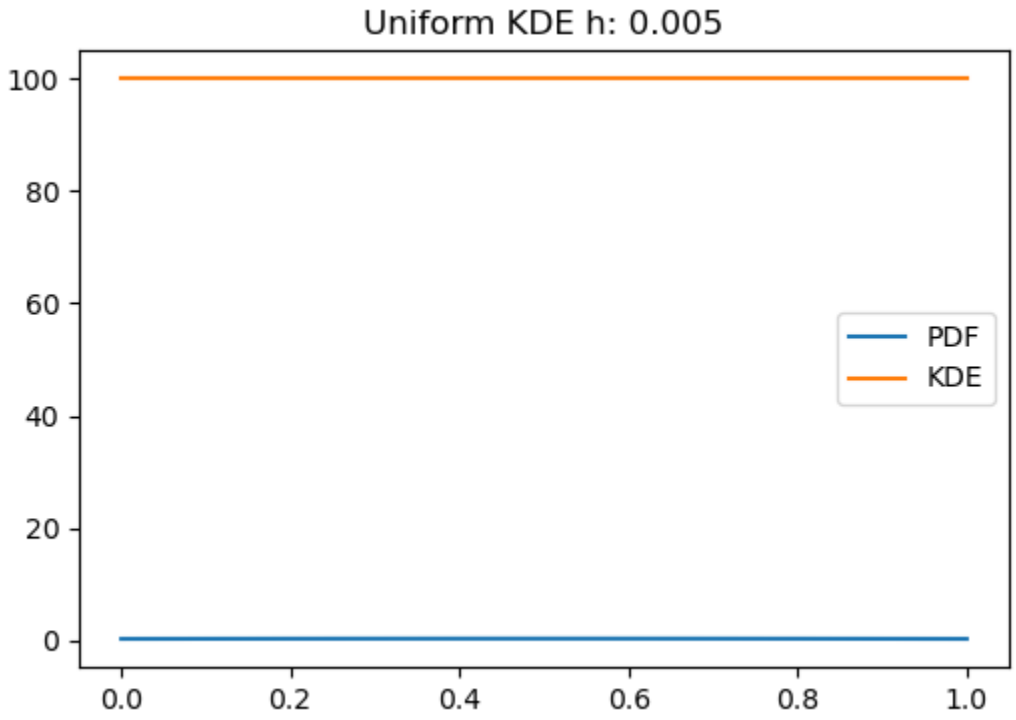
****
Uniform KDE
H : 0.0005
Sample Mean: 1000.0 Sample Variance: 0.0
True Mean: 0.3826133630634787 True Variance: 0.000206429233868056
Percentage Deviation from Mean: 261260.44804950836
Percentage Deviation from Variance: -100.0



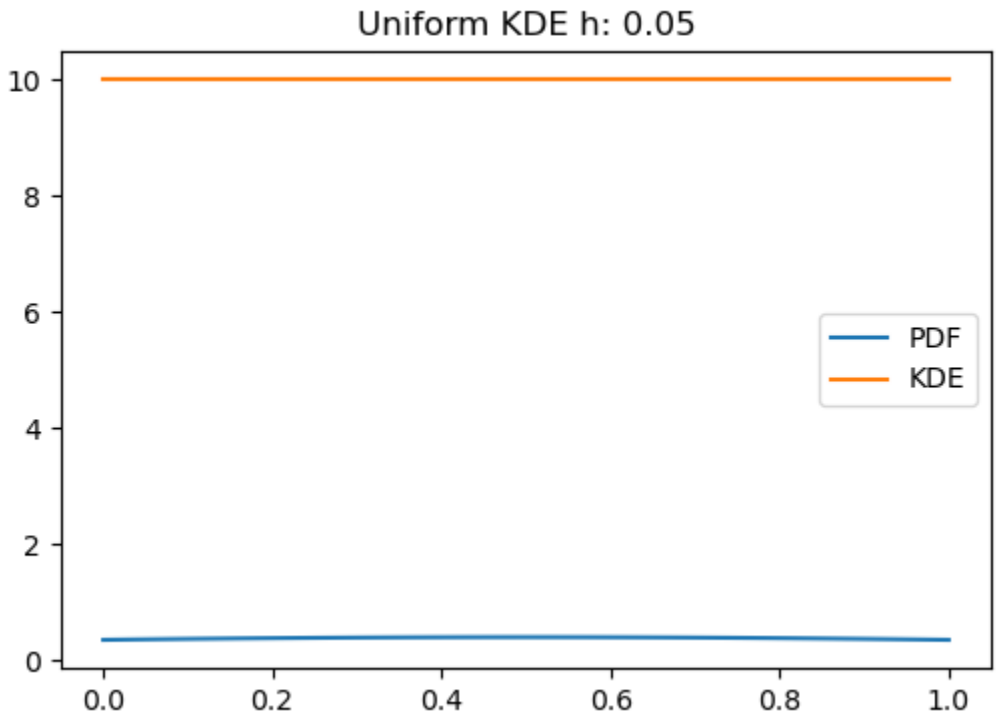
****
Uniform KDE
H : 0.001
Sample Mean: 500.0 Sample Variance: 0.0
True Mean: 0.3826133630634787 True Variance: 0.000206429233868056
Percentage Deviation from Mean: 130580.22402475418
Percentage Deviation from Variance: -100.0



****
Uniform KDE
H : 0.005
Sample Mean: 100.0 Sample Variance: 0.0
True Mean: 0.3826133630634787 True Variance: 0.000206429233868056
Percentage Deviation from Mean: 26036.044804950834
Percentage Deviation from Variance: -100.0

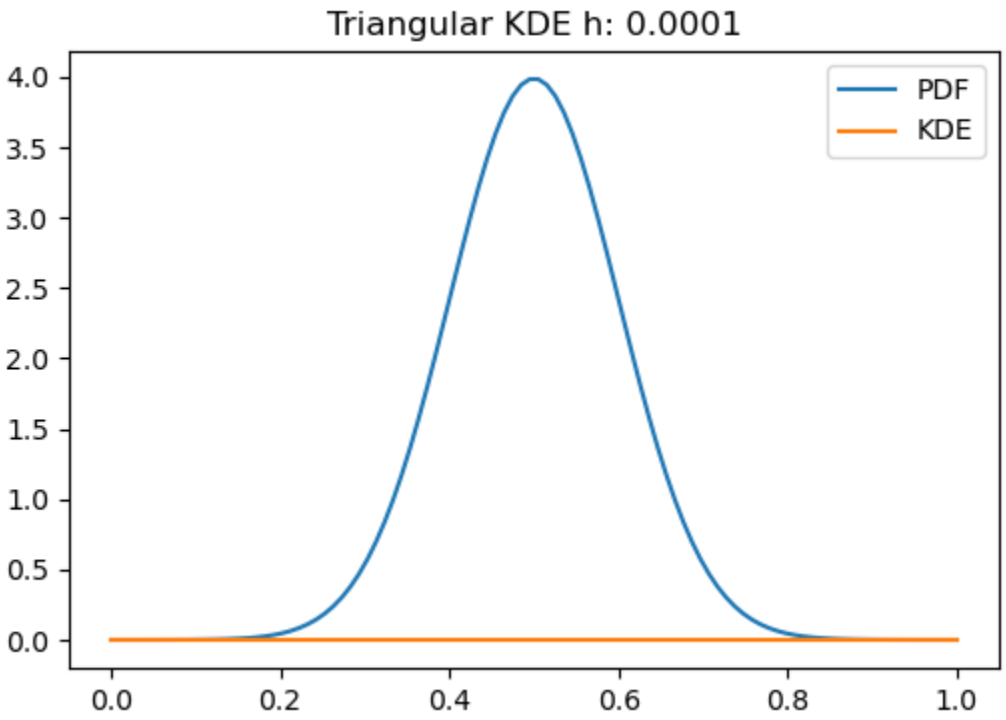


****
Uniform KDE
H : 0.05
Sample Mean: 10.0 Sample Variance: 0.0
True Mean: 0.3826133630634787 True Variance: 0.000206429233868056
Percentage Deviation from Mean: 2513.6044804950834
Percentage Deviation from Variance: -100.0

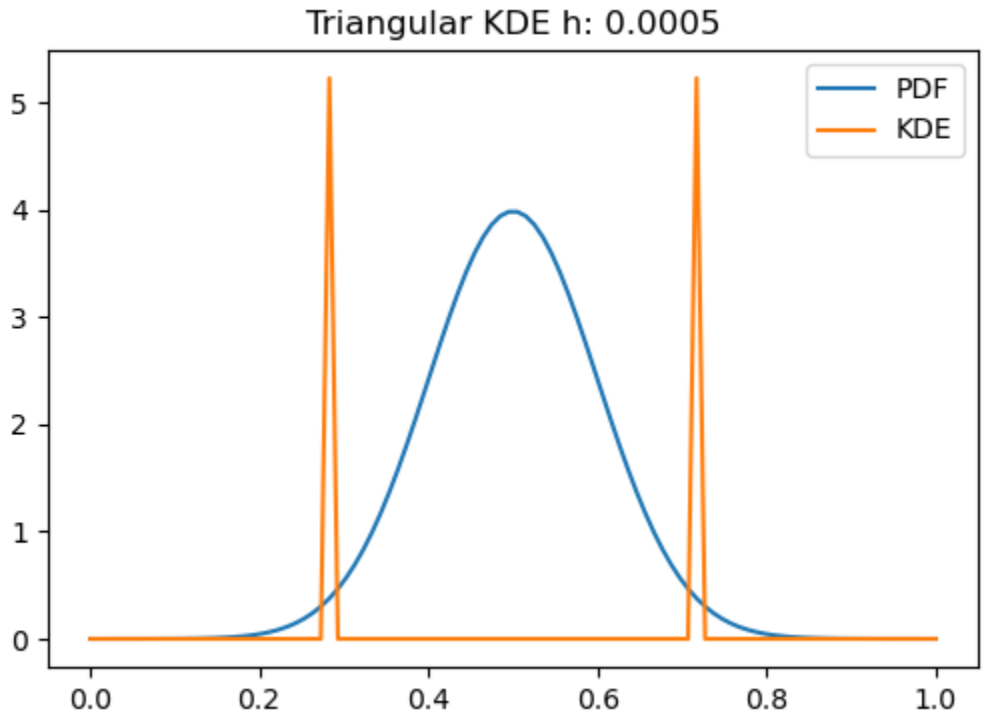


Best MSE for Uniform KDE = 92.49412572432517 with H = 0.05

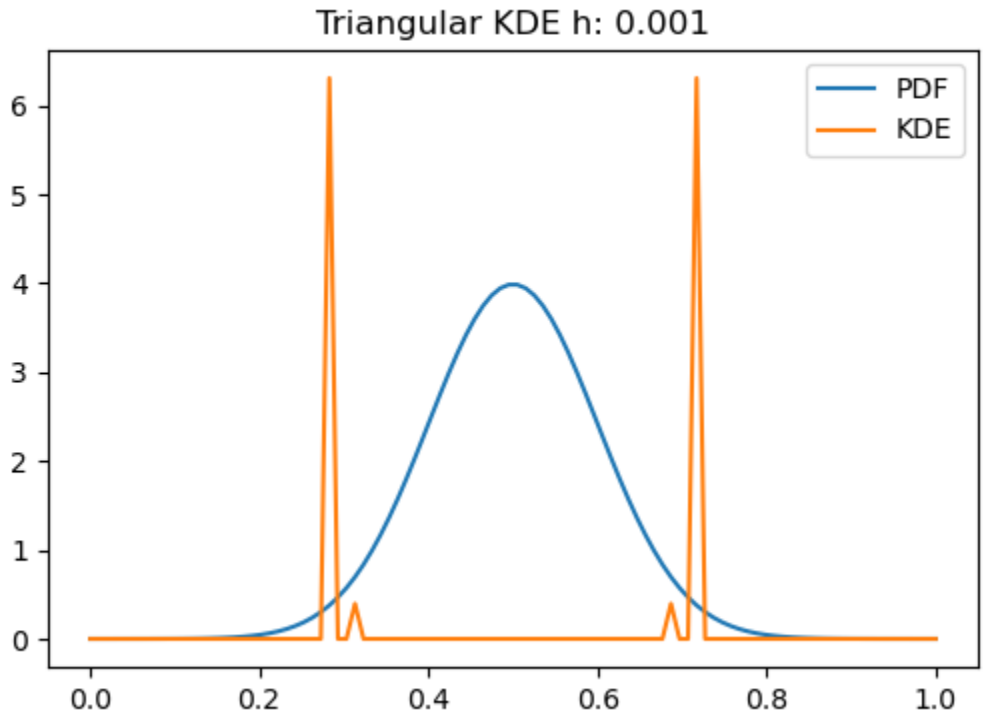
****
Triangular KDE
H : 0.0001
Sample Mean: 0.0 Sample Variance: 0.0
True Mean: 0.9899995686339825 True Variance: 1.8126392926634731
Percentage Deviation from Mean: -100.0
Percentage Deviation from Variance: -100.0



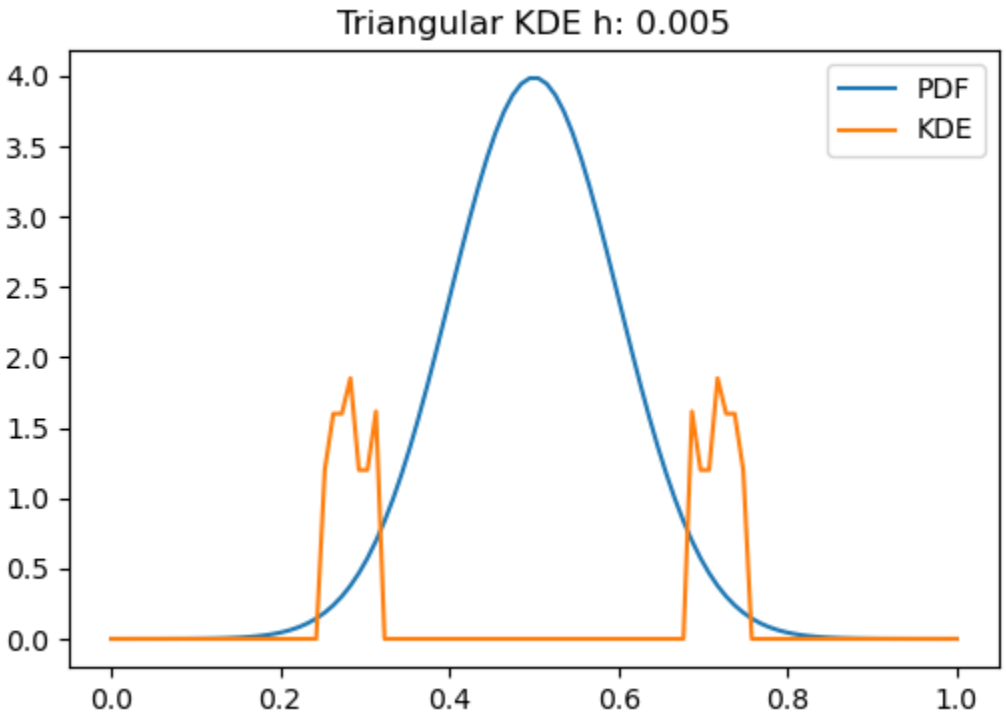
****
Triangular KDE
H : 0.0005
Sample Mean: 0.1046148748805594 Sample Variance: 0.5362693302674788
True Mean: 0.9899995686339825 True Variance: 1.8126392926634731
Percentage Deviation from Mean: -89.43283631679671
Percentage Deviation from Variance: -70.41500024643676



****
Triangular KDE
H : 0.001
Sample Mean: 0.1340705033716024 Sample Variance: 0.7808969114325771
True Mean: 0.9899995686339825 True Variance: 1.8126392926634731
Percentage Deviation from Mean: -86.45751901118552
Percentage Deviation from Variance: -56.91934326960686

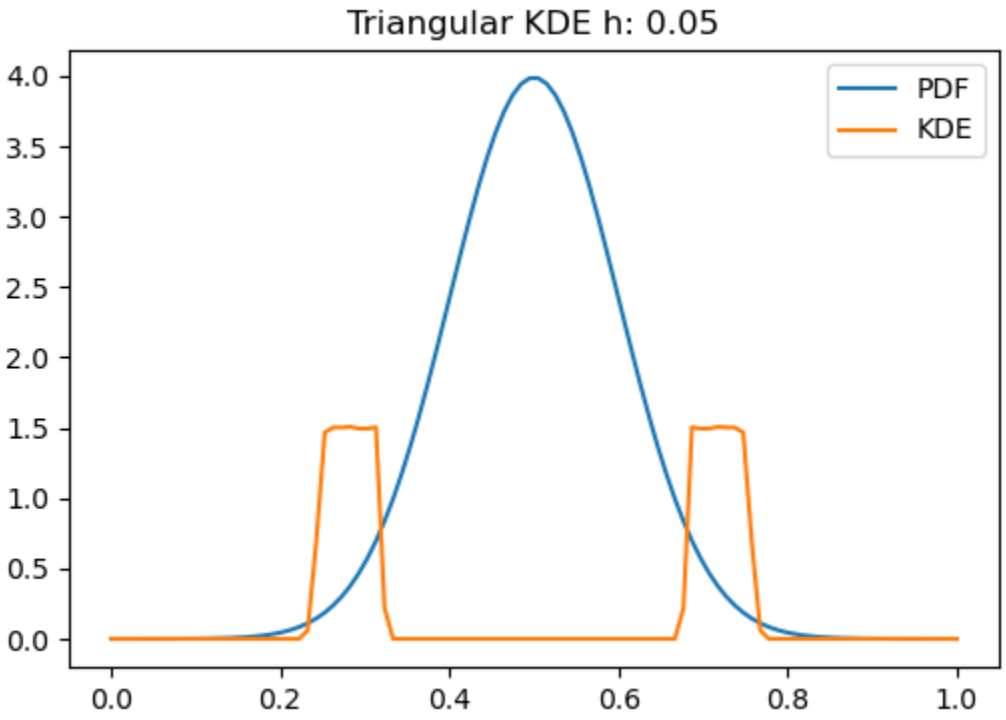


****
Triangular KDE
H : 0.005
Sample Mean: 0.20536282013486293 Sample Variance: 0.2674653314359089
True Mean: 0.9899995686339825 True Variance: 1.8126392926634731
Percentage Deviation from Mean: -79.2562717559336
Percentage Deviation from Variance: -85.24442604116244



****

Triangular KDE
H : 0.05
Sample Mean: 0.22872594922891884 Sample Variance: 0.27139243955155995
True Mean: 0.9899995686339825 True Variance: 1.8126392926634731
Percentage Deviation from Mean: -76.89635869796199
Percentage Deviation from Variance: -85.027774657098



****

Best MSE for Triangular KDE = 0.8509299631536458 with H = 0.05