CSE 544

PROBABILITY AND STATISTICS FOR DATA SCIENCE

ASSIGNMENT 3

Team members:

Name	SBU ID
Venkata Satwik Chilukuri	113263262
Akshay Somayaji	113322316
Shardul Churi	114403283

1)

A To find MSE of $\hat{\theta}$ = $\text{bias}^2(\hat{\theta}) + \text{var}(\hat{\theta})$

we know that,

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

Now squaring on both sides

$$\Rightarrow (\text{bias}(\hat{\theta}))^2 = (E[\hat{\theta}] - \theta)^2$$

$$\Rightarrow \text{bias}^2(\hat{\theta}) = (E[\hat{\theta}])^2 + \theta^2 - 2E[\hat{\theta}]\cdot\theta \quad — (1)$$

As $\hat{\theta}$ is a Random variable, we can find its variance

$$\text{var}(\hat{\theta}) = E[(\hat{\theta})^2] - (E[\hat{\theta}])^2$$

$$\Rightarrow E[(\hat{\theta})^2] = \text{var}(\hat{\theta}) + (E[\hat{\theta}])^2 \quad — (2)$$

Now,

we know that

MSE of an estimator $\hat{\theta}$ = $E[(\theta - \hat{\theta})^2]$

$$\Rightarrow MSE = E[\theta^2 + (\hat{\theta})^2 - 2\theta\cdot\hat{\theta}]$$

Using linearity of expectation

$$\Rightarrow MSE = E[\theta^2] + E[(\hat{\theta})^2] - E[2\theta\cdot\hat{\theta}]$$

we know that θ is a constant

So, $E[\text{constant}] = \text{constant}$, $E[\text{constant} * X] = \text{Const}\, E[X]$

$$\Rightarrow MSE = \theta^2 + E[(\hat{\theta})^2] - 2\theta E[\hat{\theta}]$$

using ②, substituting $E[(\hat{\theta})^2]$ as $var(\hat{\theta}) + (E[\hat{\theta}])^2$

$$\Rightarrow MSE = \theta^2 + var(\hat{\theta}) + (E[\hat{\theta}])^2 - 2\theta E[\hat{\theta}]$$

using ①, substituting $(E[\hat{\theta}])^2 + \theta^2 - 2E[\hat{\theta}]\cdot\theta$ as $bias^2(\hat{\theta})$

$$\therefore MSE = bias^2(\hat{\theta}) + var(\hat{\theta})$$

Hence proved

2)
a

Given,

$$D = \{\underset{x_1}{393}, \underset{x_2}{377}, \underset{x_3}{414}, \underset{x_4}{382}, \underset{x_5}{335}, \underset{x_6}{461}, \underset{x_7}{428}, \underset{x_8}{406}, \underset{x_9}{464}, \underset{x_{10}}{352}\}$$

we know that ecdf at α is

$$ecdf = \frac{\sum_{i=1}^{10} I(x_i \le \alpha)}{10}$$

where,

$$\text{for } i \in [1,10],\ I(x_i \le \alpha) = \begin{cases} 1 & \text{if } x_i \le \alpha \\ 0 & \text{otherwise} \end{cases}$$

Now

ecdf for $\alpha \le 334$ will be zero because there are no values at D which are $\le \alpha$

So,

At $\alpha = 335$, ecdf will be $\frac{1}{10} = 0.1$

At $\alpha = 352$, ecdf will be $\frac{2}{10} = 0.2$

At $\alpha = 377$, ecdf will be $\frac{3}{10} = 0.3$

⋮

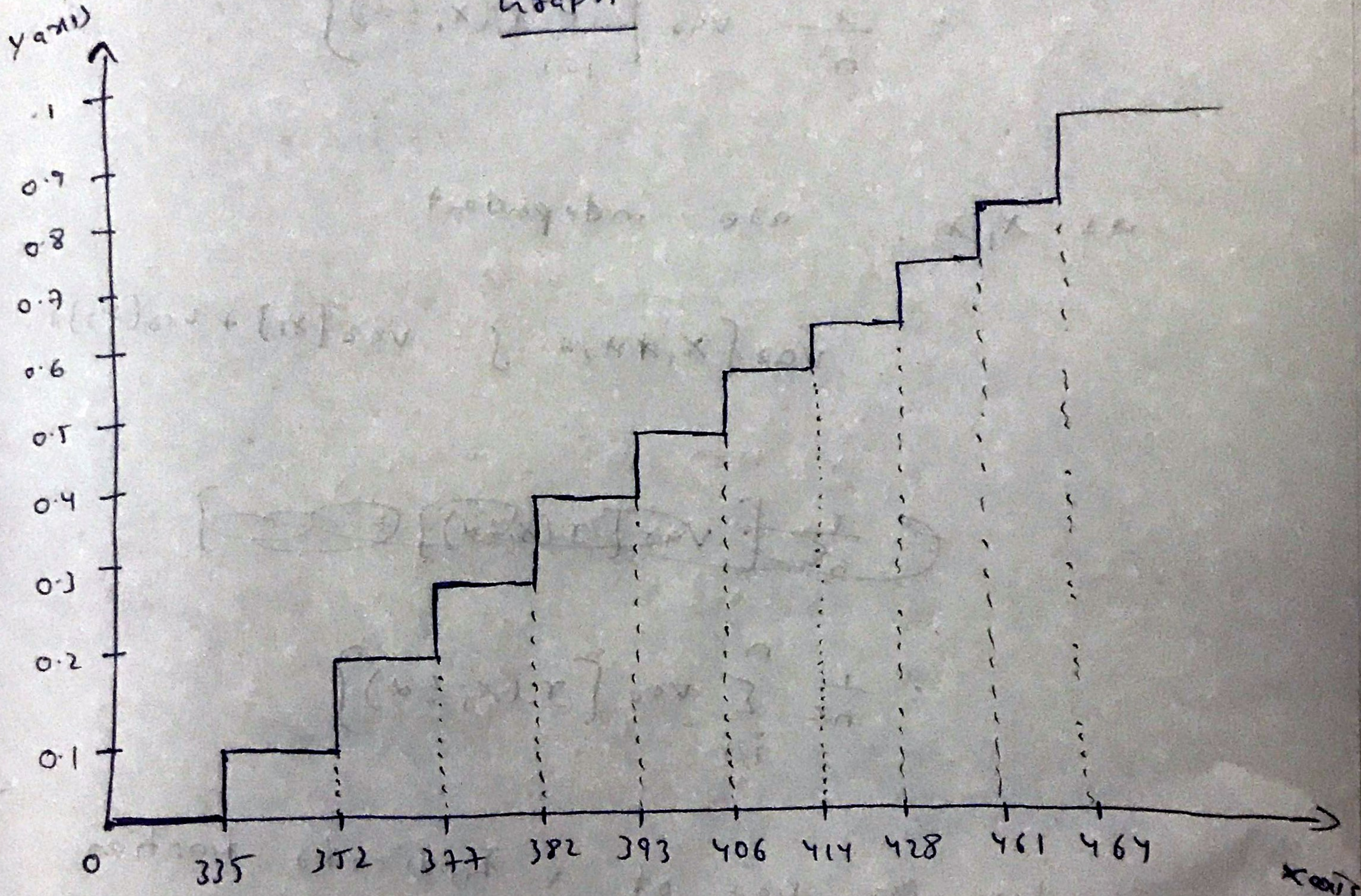
$|||^{ly}$

At $\alpha = 464$, ecdf will be $\frac{10}{10} = 1$

x - axis → sample points

y - axis → ecdf.

Graph



sample points

3) **Programming fun with $\boldsymbol{F}$**

a) Refer png file **3a_n_10.png** under **./Graphs/Q3 path**

Taking arbitrary list of samples

sample_list = [2, 56, 24, 67, 8, 24, 72, 24, 10, 45]

b) Refer png files with the file pattern **3b_*.png** under **./Graphs/Q3 path**

Observation:

As we increase the sample size n, the CDF estimate, i.e., eCDF approaches the true CDF and converges towards becoming a straight line (like the true CDF).

c) Refer png file **3c_n_5_m_5.png** under **./Graphs/Q3 path**

d) Refer png files with the file pattern **3d_*.png** under **./Graphs/Q3 path**

Observation:

As we increase the number of students, i.e, the number of rows m, the CDF estimate, i.e, eCDF approaches the true CDF and converges towards becoming a straight line (like the true CDF), even though the number of samples n remains constant. Also, since each row m has n samples, it is effectively equivalent of sampling over n*m samples, which may explain why the eCDF at m=x appears to be sharper than the corresponding graph in 3b) with n=x, where x = {10, 100, 1000}

e) Refer png file **3e_normal_ci.png** under **./Graphs/Q3 path** for Normal-based CI

f) - Refer png file **3e_kdw_ci.png** under **./Graphs/Q3 path** For KDW-based CI

- Refer png file **3f_normal_kdw_ci.png** under **./Graphs/Q3 path** For comparison of both normal and DKW-based CIs

Observation:

From **3f_normal_kdw_ci.png,** we can see that Normal-based CI is tighter than the DKW-bound one.

4)

a)

$$D = \{x_1, x_2, \ldots, x_n\} \sim X$$

Given,

$$E[X] = \bar{x}_n = \frac{1}{n}\sum_{i=1}^{n} x_i \quad \text{— (1)}$$

we know that

$$\text{Var}[X] = E[X^2] - (E[X])^2$$

$$\Rightarrow \text{Var}[X] = \sum_{x \in D} x^2 \cdot P(x) - \left(\sum_{x \in D} x \cdot P(x)\right)^2$$

plug in estimator of $\text{Var}[X]$ is $\hat{\text{var}}[X]$

$$\Rightarrow \hat{\text{Var}}[X] = \sum_{x \in D} x^2 \cdot \hat{p}(x) - \left(\sum_{x \in D} x \cdot \hat{p}(x)\right)^2$$

let us Assume all are unique $(x_1, x_2, \ldots x_n)$

so, $\hat{p}_x(x)$ will be $\frac{1}{n}$

$$\Rightarrow \hat{\text{var}}[X] = \sum_{x \in \{x_1, x_2 \ldots x_n\}} x^2 \cdot \frac{1}{n} - \left(\sum_{x \in (x_1, x_2 \ldots x_n)} x \cdot \frac{1}{n}\right)^2$$

$$\Rightarrow \hat{\text{Var}}[X] = \sum_{i=1}^{n} x_i^2 \cdot \frac{1}{n} - \left(\sum_{i=1}^{n} x_i \cdot \frac{1}{n}\right)^2$$

$$\Rightarrow \hat{var}[x] = \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - \underbrace{\frac{1}{n}\left[\sum_{i=1}^{n} x_i\right]}_{\bar{x}_n} \cdot \sum_{i=1}^{n} x_i\right]$$

from ①

$$= \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - \bar{x}_n \times \underbrace{\sum_{i=1}^{n} x_i}_{n \times \bar{x}_n}\right]$$

from ①

$$= \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - \bar{x}_n \times n \cdot \bar{x}_n\right]$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - n\bar{x}_n^2\right] \quad — ③$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - 2n \cdot \bar{x}_n^2 + n \cdot \bar{x}_n^2\right]$$

$$= \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - 2\,(n \cdot \bar{x}_n)\,\bar{x}_n + n \cdot \bar{x}_n^2\right]$$

from ① $\quad n \cdot \bar{x}_n = \sum_{i=1}^{n} x_i$

$$= \frac{1}{n}\left[\sum_{i=1}^{n} x_i^2 - 2 \cdot \sum_{i=1}^{n} x_i \cdot \bar{x}_n + n \cdot \bar{x}_n^2\right]$$

$n \cdot (\bar{x}_n)^2$ can be written as $\bar{x}_n + \bar{x}_n + \dots + \bar{x}_n$ for n times

then $(x_1^2 - 2x_1\bar{x}_n + \bar{x}_n^2) + (x_2^2 - 2x_2\bar{x}_n + \bar{x}_n^2) + \dots$

can be written as $\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$

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

$$\Rightarrow \hat{var}[x] = \frac{1}{n}\left[\sum_{i=1}^{n}(x_i - \bar{x}_n)^2\right]$$

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

Hence proved

(b)

For any estimator $\hat{\theta}$,

$$\text{Bias}[\hat{\theta}] = E[\hat{\theta}] - \theta$$

So,

$$\text{Bias}[(\hat{\sigma})^2] = E[\hat{\sigma}^2] - \sigma^2 \quad \text{--- (1)}$$

lets find $E[\hat{\sigma}^2]$

we proved in 4(a) that $(\hat{\sigma}^2) = \frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2$

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

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

$$= \frac{1}{n}\cdot E\left[\sum_{i=1}^{n} x_i^2 - 2\bar{x}_n\cdot\sum_{i=1}^{n} x_i + n\cdot\bar{x}_n^2\right]$$

As, $\bar{x}_n = \sum_{i=1}^{n}\frac{1}{n}\cdot x_i$ then $n\cdot\bar{x}_n = \sum_{i=1}^{n} x_i$

$$= \frac{1}{n}\cdot E\left[\sum_{i=1}^{n} x_i^2 - 2\bar{x}_n\cdot n\cdot\bar{x}_n + n\cdot\bar{x}_n^2\right]$$

$$= \frac{1}{n}\cdot E\left[\sum_{i=1}^{n} x_i^2 - 2n\bar{x}_n^2 + n\cdot\bar{x}_n^2\right]$$

$$= \frac{1}{n}\cdot E\left[\sum_{i=1}^{n} x_i^2 - n\cdot\bar{x}_n^2\right]$$

using linearity of expectation

$$= \frac{1}{n}\left[\underbrace{E\left[\sum_{i=1}^{n} x_i^2\right]}_{①}\right] - \underbrace{E\left[\bar{x}_n^2\right]}_{②}$$

<u>lets solve</u> <u>part ①</u>

$$= \frac{1}{n}\cdot E\left[\sum_{i=1}^{n} x_i^2\right]$$

using linearity of expectation

$$= \frac{1}{n}\cdot\sum_{i=1}^{n} E[x_i^2]$$

as $X_1, X_2, \dots X_n$ are iid then

$$= \frac{1}{n} \times n \times E(X_1^2)$$

$$= E(X_1^2)$$

As $X_1, X_2, \dots X_n \sim X$ so,

$$= E(X^2)$$

<u>lets solve</u> <u>part ②</u> :

$$= E[\bar{X}_n^2]$$

we know that,

$$Var(\bar{X}_n) = E[(\bar{X}_n)^2] - (E[\bar{X}_n])^2$$

$$\Rightarrow E[(\bar{X}_n)^2] = Var(\bar{X}_n) + (E(\bar{X}_n))^2$$

we know that $\bar{X}_n = \sum_{i=1}^{n} X_i \times \frac{1}{n}$

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

As, $X_1, X_2 \dots X_n$ are iid

using linearity of variance and

linearity of expectation.

$$\Rightarrow E\left[(\bar{X}_n)^2\right] = \frac{1}{n^2} \cdot \sum_{i=1}^{n} var(X_i) + \frac{1}{n^2}\left(\sum_{i=1}^{n} E[X_i]\right)^2$$

As, $X_1, X_2, \dots X_n$ are iid

$$E[X_1] = E[X_2] = \dots = E[X_n]$$

$$var[X_1] = var[X_2] = \dots = var[X_n]$$

$$\Rightarrow E\left[(\bar{X}_n)^2\right] = \frac{1}{n^2} \cdot n \cdot var(X_1) + \frac{1}{n^2} \cdot \left(n \cdot E[X_1]\right)^2$$

$$= \frac{1}{n} \cdot var(X_1) + \frac{1}{n^2} \cdot n^2 \left(E[X_1]\right)^2$$

$$= \frac{1}{n} \cdot var(X_1) + \left(E[X_1]\right)^2$$

~~from 4(a) equation~~ 3

we know that,

$$var(X) = E[X^2] - \left(E[X]\right)^2$$

$$= \frac{1}{n} \cdot \left[E(X_1^2) - \left(E[X_1]\right)^2\right] + \left(E[X_1]\right)^2$$

as $(X_1, X_2, \dots X_n) \sim X$

$$= \frac{1}{n} \cdot \left[E(X^2) - \left(E[X]\right)^2\right] + \left(E[X]\right)^2$$

Now putting back the solved ①, ② parts back in the equation

$$\Rightarrow E[\hat{\sigma}^2] = E[x^2] - \left[\frac{1}{n}\cdot\left[E(x^2) - (E(x))^2\right] + (E(x))^2\right]$$

$$= E[x^2] - \frac{1}{n}\cdot\left[E(x^2) - (E(x))^2\right] - (E(x))^2$$

we know that

$$\sigma^2 = E[x^2] - (E(x))^2$$

$$\Rightarrow E[\hat{\sigma}^2] = \sigma^2 - \frac{1}{n}(\sigma^2)$$

Now

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

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

$$= -\sigma^2/n$$

$\therefore$ bias of $\hat{\sigma}^2 = -\sigma^2/n$, where σ^2 is the variance

(i)

$(x_1, x_2, \dots x_n) \sim X$

Sample Mean, $\mu = \frac{1}{n} \cdot \sum_{i=1}^{m} x_i$

plug-in estimator of variance of X i.e $\hat{\sigma}^2$

$$\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^2 \quad \left[\text{from } 4(a)\right]$$

— ①

Given,

$$\text{Kurt}[X] = \frac{E\left[(X-\mu)^4\right]}{\sigma^4}$$

plug-in estimator of $\text{Kurt}[X] = \hat{\text{Kurt}}[X]$

Now,

$$\hat{\text{Kurt}}[X] = \frac{E\left[(X-\mu)^4\right]}{(\hat{\sigma})^4}$$

we know that $E[X] = \frac{1}{n} \cdot \sum_{i=1}^{n} x_i$

$(X-\mu)^4$ is a r.v so, $E\left[(X-\mu)^4\right] = \frac{1}{n} \cdot \sum_{i=1}^{n} (x_i - \bar{x}_n)^4$

$$\Rightarrow \hat{\text{Kurt}}[X] = \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)^4 * \frac{1}{n}}{(\hat{\sigma})^4}$$

$$\Rightarrow \hat{\text{Kurt}}[x] = \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^4}{(\hat{\sigma}^2)^2}$$

from ① $\hat{\sigma}^2 = \frac{1}{n} \sum_{i=1}^{n} (x_i - x_n)^2$

$$= \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^4}{\left[\frac{1}{n} \sum_{i=1}^{n} (x_i - x_n)^2\right]^2}$$

$$= \frac{\frac{1}{n} \sum_{i=1}^{n} (x_i - \bar{x}_n)^4}{\frac{1}{n^2}\left[\sum_{i=1}^{n} (x_i - x_n)^2\right]^2}$$

$$\boxed{\therefore \hat{\text{Kurt}}[x] = n \cdot \frac{\sum_{i=1}^{n} (x_i - \bar{x}_n)^4}{\left[\sum_{i=1}^{n} (x_i - \bar{x}_n)^2\right]^2}}$$

where, $\bar{x}_n = \sum_{i=1}^{n} x_i \times \frac{1}{n}$

d)

Given,

$$\rho = \frac{E[XY] - E[X]E[Y]}{\sigma_x \cdot \sigma_y}$$

plug-in estimator for ρ

$$\Rightarrow \hat{\rho} = \frac{E[XY] - E[X] \cdot E[Y]}{\hat{\sigma}_x \cdot \hat{\sigma}_y}$$

$\hat{\sigma}_X, \hat{\sigma}_y$ is standard deviation of X, Y

$$\hat{\sigma}_x = \sqrt{var(x)} \quad , \hat{\sigma}_y = \sqrt{var(y)}$$

from 4(a)

$$\hat{\sigma}_x = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(x_i - \bar{x}_n)^2} \quad , \hat{\sigma}_y = \sqrt{\frac{1}{n}\sum_{i=1}^{n}(y_i - \bar{y}_n)^2}$$

$$\boxed{\hat{\rho} = \frac{\sum_{i=1}^{n} x_i y_i \times \frac{1}{n} - \bar{x}_n \cdot \bar{y}_n}{\sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n}(x_i - \bar{x}_n)^2} \cdot \sqrt{\frac{1}{n} \cdot \sum_{i=1}^{n}(y_i - \bar{y}_n)^2}}}$$

5)

Given,

$D = \{x_1, x_2, x_3, \ldots, x_n\}$ set of iid samples

Indicator function will be,

$$I(x \le \alpha) = \begin{cases} 1 & \text{if } x \le \alpha \\ 0 & \text{otherwise} \end{cases}$$

True CDF $= F$, eCDF $= \hat{F}$

(a)

$\hat{F}$ at some α , $\hat{F} = \hat{F}_x(\alpha)$

$$= Pr(\hat{x} \le \alpha)$$

$$= \frac{\sum_{i=1}^{n} I(x_i \le \alpha)}{n}$$

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} I(x_i \le \alpha)$$

as x is R.V we can use expectation

So,

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

n is constant so

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

using linearity of expectation

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

as $X_1, X_2, \dots X_n$ are iid

$$= \frac{1}{n}\left[\cancel{\sum_{i=1}} \; n \cdot E(I(X_1 \le \alpha))\right]$$

$$= E(I(X_1 \le \alpha))$$

In Assignment -1, Question 5(a) we proved that Expectation of a Indicator R.V is the probability of the R.V

$$= Pr(X_1 \le \alpha)$$

As, $(X_1, X_2, \dots X_n) \sim X$ so,

$$= Pr(X \le \alpha) = F_X(\alpha)$$

$= F_x(u)$

This is the true CDF F

so,

$$E[\hat{F}] = F$$

b)

we know that,

for some estimator $\hat{\theta}$,

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

so,

$$\text{Bias}(\hat{F}) = E[\hat{F}] - F$$

we just proved in the J(a) that

$$E[\hat{F}] = F$$

$$\Rightarrow \text{Bias}(\hat{F}) = F - F$$

$$\boxed{\therefore \text{Bias}(\hat{F}) = 0}$$

(c)

we know that,

for some estimator $\hat{\theta}$

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

so,

$$se(\hat{F}) = \sqrt{var(\hat{F})}$$

$$= \sqrt{var\left(\frac{\sum_{i=1}^{n} I(X_i \le \alpha)}{n}\right)}$$

we know that $var(kX) = k^2 \cdot var(X)$ where k is constant

$$= \sqrt{\frac{1}{n^2} \cdot var\left(\sum_{i=1}^{n} I(x_i \le \alpha)\right)}$$

As, $X_1, X_2, \ldots X_n$ are iid. Using linearity of variance

$$= \sqrt{\frac{1}{n^2} \cdot \sum_{i=1}^{n} var(I(x_i \le \alpha))}$$

As $X_1, X_2, \ldots X_n$ are iid, $var(X_1) = var(X_2) = \ldots$

$$= \sqrt{\frac{1}{n^2} \cdot n \cdot var(I(X_1 \le \alpha))}$$

$$= \sqrt{\frac{1}{n} \cdot var(I(X_i \le \alpha))}$$

In Assignment-1, Question 5(b) we proved that variance of a Indicator R.V is the

$$pr(x)(1 - pr(x))$$

$$= \sqrt{\frac{1}{n} \cdot pr(X_i \le \alpha)[1 - pr(X_i \le \alpha)]}$$

As, $(X_1, X_2, \ldots X_n) \sim X$ we can say that

$$= \sqrt{\frac{1}{n} \cdot pr(X \le \alpha)[1 - pr(X \le \alpha)]}$$

$$= \sqrt{\frac{1}{n} F[1 - F]} \quad \left[\text{As } pr(X \le \alpha) \text{ is true cdf } (F(\alpha))\right]$$

$$= \sqrt{\frac{F[1-F]}{n}}$$

$$\boxed{\therefore \; se(\hat{F}) = \sqrt{\frac{F[1-F]}{n}}}$$

(d)

for any estimator $\hat{\theta}$ to be consistent

$bias(\hat{\theta})$ as $n \to \infty$ should be zero

$se(\hat{\theta})$ as $n \to \infty$ should be zero

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

Here,

$bias(\hat{f}) = 0$ [proved it in 5(b)]

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

as $n \to \infty$

$se(\hat{f})$ will be equal to 0

So,

$\hat{f}$ is a consistent estimator.

6)

(a)

Given,

$$\hat{\theta} = \frac{1}{n} \sum_{i=1}^{n} x_i \quad , \quad (x_1, x_2, \dots x_n) \sim \text{Bernoulli}(\theta)$$

$x_1, x_2, \dots x_n$ are iid

Now,

$$E[\hat{\theta}] = E\left[\frac{1}{n} \sum_{i=1}^{n} x_i\right]$$

$$= \frac{1}{n} \cdot E\left[\sum_{i=1}^{n} x_i\right]$$

using linearity of expectation

$$= \frac{1}{n} \cdot \sum_{i=1}^{n} E[x_i]$$

as $x_1, x_2, \dots, x_n$ are iid

$$= \frac{1}{n} \cdot n \times E[x_1]$$

$$= E[x_1]$$

Given, $x_1, x_2, x_3 \dots x_n \sim$ bernoulli (θ)

so, $E[x_1] = \theta$

$\therefore$ $E[\hat{\theta}] = \theta$.

Now,

$$\text{bias of } \hat{\theta} = E[\hat{\theta}] - \theta$$

$$= \theta - \theta$$

$$= 0$$

$$\boxed{\therefore \text{bias}[\hat{\theta}] = 0} \quad -(1)$$

we know that,

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

$$= \sqrt{var\left(\frac{\sum_{i=1}^{n} x_i}{n}\right)} \quad (\text{Given } \hat{\theta})$$

$$= \sqrt{\frac{1}{n^2} \cdot var\left(\sum_{i=1}^{n} x_i\right)}$$

As, $x_1, x_2, \dots x_n$ are iid, using linearity of variance

$$= \sqrt{\frac{1}{n^2} \cdot n \cdot var(X_1)}$$

$x_1, x_2, \dots x_n \sim$ Bernoulli (θ)

so, $var(x) = \theta(1-\theta)$

$$= \sqrt{\frac{1}{n} \cdot \theta(1-\theta)}$$

$$\boxed{\therefore \ se(\hat{\theta}) = \sqrt{\frac{\theta(1-\theta)}{n}}} \quad -②$$

In the Assignment -3, Question -1 we proved that

MSE of estimator $\hat{\theta}$ = $bias^2(\hat{\theta}) + var(\hat{\theta})$

So,

using that formulae

$$\text{MSE of } \hat{\theta} = 0 + \left(\sqrt{\frac{\theta(1-\theta)}{n}}\right)^2$$

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

$$\boxed{\therefore \ MSE(\hat{\theta}) = \frac{\theta(1-\theta)}{n}}$$

6b) We have to derive the Normal-based $(1-\alpha)$ CI for $\hat{\theta}$.
Now, to apply normal-based CI on an estimator, it needs to satisfy the below 2 conditions:

i) $\hat{\theta}$ is normally distributed

ii) The distribution for the estimate $\hat{\theta}$ is centered around true value θ.

→ we have already proved ii) in 6a), when we proved that bias of $\hat{\theta} = 0$ since $E[\hat{\theta}] = \theta$ (due to the fact that X_i are i.i.d $\sim$ Bernoulli(θ))

→ Since $\hat{\theta} = \frac{1}{n}\sum_{i=1}^{n} X_i$, we can prove that $\hat{\theta}$ is normally distributed using the Central Limit Theorem. Since $X_1, X_2 \dots X_n$ are iid RVs with some true mean μ and variance σ^2, then the distribution of the sample means will be normally distributed around $(\mu, \frac{\sigma^2}{n})$. $\Rightarrow \hat{\theta} = \frac{\sum X_i}{n} \sim \text{Normal}(\mu, \frac{\sigma^2}{n})$ when $n \to \infty$

From 6a), we have true mean $= \theta$, and true variance $= \theta(1-\theta)$

$$\Rightarrow \hat{\theta} \sim \text{Normal}\left(\theta, \frac{\theta(1-\theta)}{n}\right)$$

Now, to find the $(1-\alpha)$ CI for $\hat{\theta}$, let us first understand the meaning of $1-\alpha$.
From our understanding of standard normal dist^n, we know that $1-\alpha$ is the area between 2 points $(-z_{\alpha/2}, z_{\alpha/2})$ on the x-axis of the standard normal distribution curve. ∴ For any point z, we have $Pr(-z_{\alpha/2} \le z \le z_{\alpha/2}) = 1-\alpha$

Now, to transform a normal distribution $N \sim \text{Normal}(\mu, \sigma^2)$ to standard normal $Z \sim \text{Normal}(0,1)$, we use

$$Z = \frac{N - \mu}{\sigma}, \quad \text{where } \sigma \text{ is the std deviation} = \sqrt{\sigma^2}$$

Hence, we can transform $\hat{\theta} \sim \text{Normal}\left(\theta, \frac{\theta(1-\theta)}{n}\right)$ into standard normal by using

$$Z = \frac{\hat{\theta} - \theta}{\sqrt{\frac{\theta(1-\theta)}{n}}} \quad \left(\because \sigma^2_{\hat{\theta}} = \frac{\theta(1-\theta)}{n}\right)$$

Replacing above z value in ①, we get

$$Pr\left(-z_{\alpha/2} \le \frac{\hat{\theta} - \theta}{\sigma} \le z_{\alpha/2}\right) = 1 - \alpha$$

$$\Rightarrow Pr\left(-\hat{\theta} - z_{\alpha/2} \cdot \sigma \le -\theta \le \sigma \cdot z_{\alpha/2} - \theta\right) = 1 - \alpha$$

$$\Rightarrow Pr\left(\hat{\theta} + z_{\alpha/2} \cdot \sigma \ge \theta \ge \hat{\theta} - z_{\alpha/2} \cdot \sigma\right) = 1 - \alpha$$

$$\therefore Pr\left(\theta \in \left(\hat{\theta} - z_{\alpha/2}, \hat{\theta} + z_{\alpha/2}\right)\right) = 1 - \alpha$$

which is in the form of $Pr(\theta \in (a,b)) = 1 - \alpha$, which defines the $(1-\alpha)$ CI within limits (a,b)

$$\therefore CI(1-\alpha) \text{ for } \hat{\theta} = \hat{\theta} \pm z_{\alpha/2} \cdot \sigma$$

$$= \hat{\theta} \pm z_{\alpha/2} \cdot \sqrt{\frac{\theta(1-\theta)}{n}}, \quad \text{where } \theta \text{ is the true expectation}$$

Now, since $\hat{\theta}$ is a good estimator of θ (it is unbiased and consistent), we can plug-in the estimator $\hat{\theta}$ in place of θ

$$\therefore (1-\alpha) \text{ CI of } \hat{\theta} = \hat{\theta} \pm z_{\alpha/2} \sqrt{\frac{\hat{\theta} \cdot (1 - \hat{\theta})}{n}}$$

$$\boxed{(1-\alpha) \text{ CI of } \hat{\theta} = \frac{\sum_{i=1}^{n} X_i}{n} \pm z_{\alpha/2} \sqrt{\frac{\frac{\sum_{i=1}^{n} X_i}{n}\left(1 - \frac{\sum_{i=1}^{n} X_i}{n}\right)}{n}} \quad \text{(Proved using CLT)}}$$

7) Kernel density estimation

b)

<u>Normal KDE</u>

<u>Refer png file 7b_normal.png under ./Graphs/Q7 path</u>

True PDF Mean - 0.990, True PDF Variance - 0.325

h-value	Sample KDE mean	Sample KDE variance	Deviation of sample mean from True PDF	Deviation of sample mean from True PDF
0.0001	13.430	167.270	1256.57%	51367.69%
0.0005	2.686	4.864	171.31%	1396.62%
0.001	1.372	0.845	38.59%	160.00%
0.005	0.992	0.327	0.20%	0.62%
0.05	0.990	0.325	0.00%	0.00%

Observation: The plot has the smoothest curve and just a single peak at h = 0.05. So we can say that h=0.05 performs the best.

c)

<u>Uniform KDE</u>

<u>Refer png file 7c_uniform.png under ./Graphs/Q7 path</u>

True PDF Mean - 0.500, True PDF Variance - 0.590

h-value	Sample KDE mean	Sample KDE variance	Deviation of sample mean from True PDF	Deviation of sample mean from True PDF
0.0001	0.500	0.590	0.00%	0.00%
0.0005	0.500	0.590	0.00%	0.00%
0.001	0.500	0.590	0.00%	0.00%
0.005	0.480	0.570	-4.00%	-3.39%
0.05	0.495	0.585	-1.00%	-0.85%

Observation: The plot has the widest, smoothest flatline peak at h=0.05 and a point peak at h=0.0001. The sample mean and variance are observed to be equal to true PDF's mean/variance at h=0.0001, 0.0005, 0.001.

Triangular KDE

Refer png file 7c_triangular.png under ./Graphs/Q7 path

True PDF Mean - 0.495, True PDF Variance - 0.585

h-value	**Sample KDE mean**	**Sample KDE variance**	**Deviation of sample mean from True PDF**	**Deviation of sample mean from True PDF**
0.0001	0.697	0.826	40.81%	41.20%
0.0005	0.503	0.593	1.62%	1.37%
0.001	0.495	0.585	0.00%	0.00%
0.005	0.470	0.561	-5.05%	-4.10%
0.05	0.495	0.585	0.00%	0.00%

Observation: The sample mean and variance are observed to be equal to true PDF’s mean/variance at h=0.001 and 0.05. The distribution resembles a true triangular distribution the best at h=0.001. Area under the curve also appears to be greatest and smoothest for h=0.05, so we can say the pdf is denser at this h-value and it is the best h-bandwidth.