20200130 Yujun Kim IE541 HW3

#1 $\hat{F}_n(x) \rightsquigarrow F(x)$.

$$\hat{F}_n(x) = \frac{1}{n}\sum_{i=1}^{n} I(X_i \le x)$$

Let $\mu = \mathbb{E}[I(X \le x)] = F(x)$, $\sigma^2 = \mathbb{V}(I(X \le x)) = F(x)(1-F(x))$ for $X \sim F$.

By CLT, $\dfrac{\sqrt{n}(\hat{F}_n(x) - \mu)}{\sigma} \rightsquigarrow N(0,1)$.

$\Rightarrow \hat{F}_n(x) \rightsquigarrow \mu$.

(or use weak law of large nbrs so that $\hat{F}_n(x) \xrightarrow{P} F(x)$ and thus $\hat{F}_n(x) \rightsquigarrow F(x)$).

Hence, F is the limiting distribution of $\hat{F}_n$.

#2 $\mathbb{E}(\hat{F}_n(x)) = F(x)$

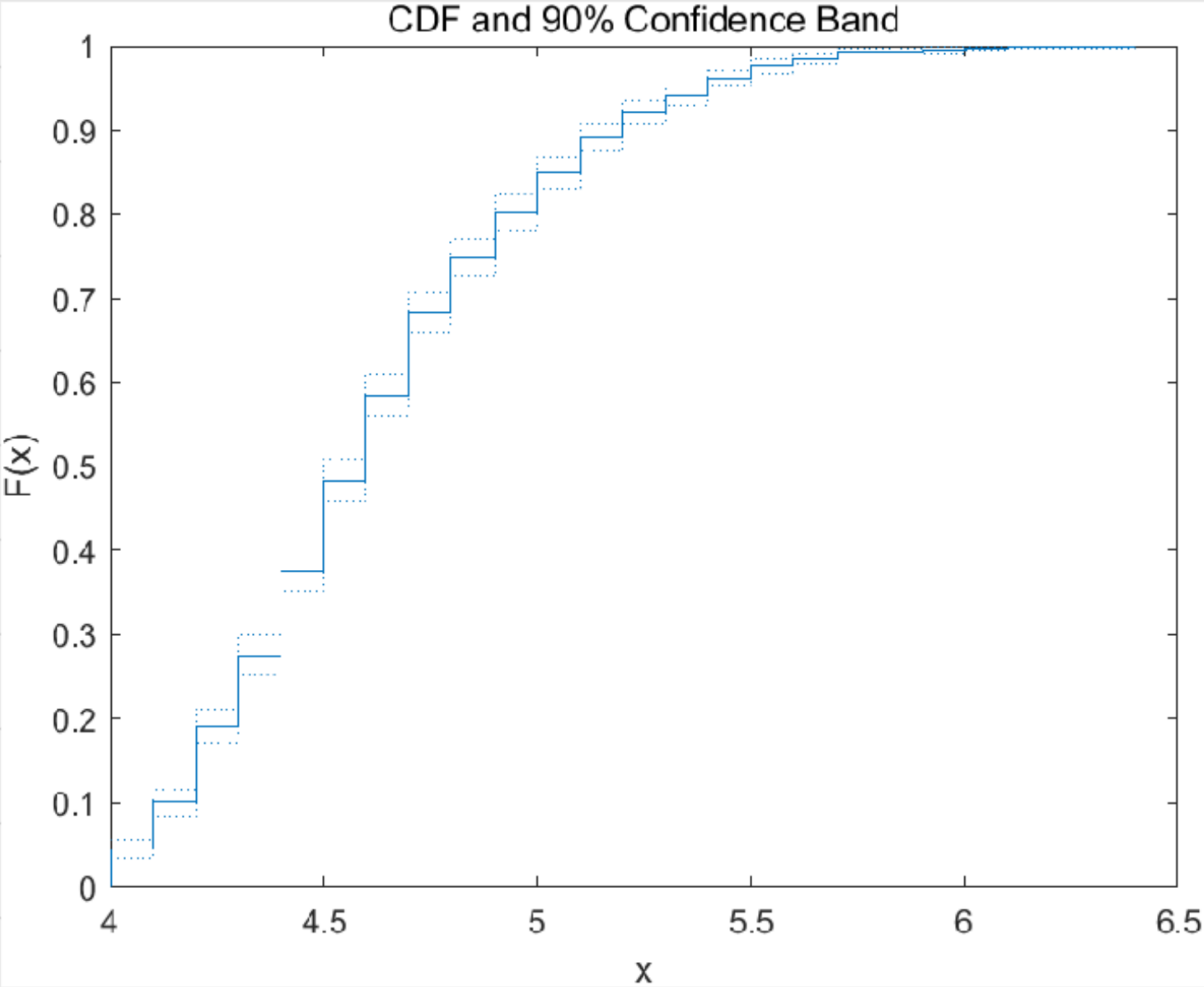
$\mathbb{V}(\hat{F}_n(x)) = \frac{1}{n} F(x)(1-F(x))$.

$\mathrm{Cov}(\hat{F}_n(x), \hat{F}_n(y)) = \mathbb{E}(\hat{F}_n(x)\hat{F}_n(y)) - \mathbb{E}(\hat{F}_n(x))\,\mathbb{E}(\hat{F}_n(y))$.

$$\begin{aligned}
\mathbb{E}(\hat{F}_n(x)\hat{F}_n(y)) &= \mathbb{E}\left(\frac{1}{n}\sum_i I(X_i \le x)\,\frac{1}{n}\sum_j I(X_j \le y)\right) \\
&= \frac{1}{n^2}\sum_i\sum_j \mathbb{E}\left(I(X_i \le x)\,I(X_j \le y)\right) \\
&= \frac{1}{n^2}\left[\sum_{i\ne j}\mathbb{E}\left(I(X_i \le x)I(X_j \le y)\right) + \sum_{i=j}\mathbb{E}\left(I(X_i \le x)I(X_j \le y)\right)\right] \\
&= \frac{1}{n^2}\left[\sum_{i\ne j}\mathbb{E}(I(X_i \le x))\,\mathbb{E}(I(X_j \le y)) + \sum_i \mathbb{E}(I(X_i \le \min\{x,y\}))\right] \\
&= \frac{1}{n^2}\left[(n^2-n)F(x)F(y) + nF(\min\{x,y\})\right] \\
&= F(x)F(y) + \frac{1}{n}F(\min\{x,y\}) - \frac{1}{n}F(x)F(y).
\end{aligned}$$

Thus, $\mathrm{Cov}(\hat{F}_n(x), \hat{F}_n(y)) = \frac{1}{n}\left[F(\min\{x,y\}) - F(x)F(y)\right]$.

#3 The following is empirical CDF $\hat{F}_n$ of given data.
Lower dotted line and upper dotted line gives 90% confidence band for F as well.
The figure is obtained by MATLAB.



#4 From #3, I obtained empirical CDF "f" in the code below. By running below MATLAB code,

```
% Problem 4
f1 = f(x == 5.0);
f2 = f(x == 4.5);
mu = f1 - f2
se = sqrt((f1*(1-f1) + f2*(1-f2))/n)
k = norminv([0.05 0.95]);
low = mu + se*k(1)
high = mu + se*k(2)
```

I obtained

```
mu = 0.3650
se = 0.0194

low = 0.3330
high = 0.3970
```

i.e. Approximate 90% confidence interval for $|F(5.0) - F(4.5)|$ is

$$(\hat{\mu} - 1.96\hat{se}, \hat{\mu} + 1.96\hat{se}) = (0.333, 0.397)$$

#5 $\mathbb{E}(X_i^* \mid X_1, \cdots, X_n) = \frac{1}{n}X_1 + \cdots + \frac{1}{n}X_n = \overline{X}_n$,

where $\overline{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$

$$\mathbb{E}(\overline{X}_n^* \mid X_1, \cdots, X_n) = \frac{1}{n}\sum \mathbb{E}(X_i^* \mid X_1, \cdots, X_n)$$

$$= \frac{1}{n} \cdot n \cdot \overline{X}_n = \overline{X}_n .$$

$$\mathbb{V}(X_i^* \mid X_1, \cdots, X_n) = \frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2 = \frac{n-1}{n}S^2,$$

where $S^2 = \frac{1}{n-1}\sum_{i=1}^{n}(X_i - \overline{X}_n)^2$

$$\mathbb{V}(\overline{X}_n^* \mid X_1, \cdots, X_n) = \frac{1}{n}\mathbb{V}(X_1^* \mid X_1, \cdots, X_n)$$

$$= \frac{n-1}{n^2}S^2$$

Let $X_1, \cdots, X_n \sim F$, $X \sim F$, $\mathbb{E}(X) = \mu$, $\mathbb{V}(X) = \sigma^2$.

$$\mathbb{E}(\overline{X}_n^*) = \mathbb{E}(\mathbb{E}(\overline{X}_n^* \mid X_1, \cdots, X_n))$$

$$= \mathbb{E}(\overline{X}_n) = \frac{1}{n}\sum_{i=1}^{n}\mathbb{E}(X_i) = \frac{1}{n} \cdot n \cdot \mu = \mu .$$

$$\mathbb{V}(\overline{X}_n^*) = \mathbb{E}(\mathbb{V}(\overline{X}_n^* \mid X_1, \cdots, X_n)) + \mathbb{V}(\mathbb{E}(\overline{X}_n^* \mid X_1, \cdots, X_n))$$

$$= \mathbb{E}\left(\frac{n-1}{n^2}S^2\right) + \mathbb{V}(\overline{X}_n)$$

$$= \frac{n-1}{n^2}\sigma^2 + \frac{\sigma^2}{n} = \frac{\sigma^2}{n}\left(2 - \frac{1}{n}\right)$$ ($\because S^2$ is unbiased estimator of σ^2)

#6 These are MATLAB code for Problem 6

```
% Problem 6
mu = 5;
n = 100;
alpha = 0.05;
dist = makedist('Normal', "mu", mu, "sigma",1);
x = random(dist, n, 1);
y = exp(x);
theta = exp(5)
theta_h = exp(mean(x))

% Bootstrap
B = 100000;
T = zeros(B, 1);
for i = 1:B
    idx = unidrnd(n, n, 1);
    T(i) = exp(mean(x(idx)));
end
% (a)
se_boot = std(T)
k = norminv([alpha/2, 1-alpha/2]);
low = (theta_h + se_boot*k(1))
high = (theta_h + se_boot*k(2))
```

```
%(b)
binwidth = 30
figure
y_sample = exp(random(dist, 1000, 1));
hold on
histogram(T, 'BinWidth',binwidth, 'Normalization', 'pdf')
histogram(y_sample, 'BinWidth',binwidth, 'Normalization', 'pdf')
stem(exp(5), 0.01)
legend('Bootstrap', 'Sample', 'True')
hold off


figure
hold on
histogram()
```

(a) Under $n=100$, $\mu=5$, $B=100{,}000$

$\theta = e^{\mu}$, $\hat{\theta} = e^{\bar{X}}$, $X_1, \cdots, X_n \sim N(\mu, 1)$,

$se_{boot} = 14.120$

95% confidence interval $= (110.03, 165.38)$

(b) The following figure shows normalized histogram to compare distribution of bootstrap replication and sample distribution, and the true value.

It follows from the figure that bootstrap replications (blue) has less variance, compared to true sample distribution (red), as B is large