א

$$E[\hat{F}_n(x)] = E\left[\frac{1}{n}\sum_{i=1}^{n} I_{\{x_i \le x\}}\right] = \frac{1}{n}E\left[\sum_{i=1}^{n} I_{\{x_i \le x\}}\right] = \frac{1}{n}\sum_{i=1}^{n} \overset{\text{iid}}{E}\left[I_{\{x_i \le x\}}\right] =$$

$$= \frac{1}{n}\sum_{i=1}^{n} F(x) = F(x)$$

ב

$$Var[\hat{F}_n(x)] = Var\left[\frac{1}{n}\sum_i \mathbb{1}_{\{x_i \le x\}}\right] = \frac{1}{n^2}Var\left[\sum_i \mathbb{1}_{\{x_i \le x\}}\right] =$$

$$= \frac{1}{n}Var\left[\mathbb{1}_{\{x_i \le x\}}\right] \underset{\text{שונות של ברנולי}}{=} \frac{1}{n}\cdot\left(F(x)(1-F(x))\right) = \frac{F(x)(1-F(x))}{n}$$

ג

כדי שהאומד יהיה עקיב, נדרוש כי יתקיים $MSE \xrightarrow{n\to\infty} 0$

$$MSE_{F(x)}(F_n(x)) = Var[F_n(x)] + bias[F_n(x)]^2 =$$

$$= \frac{F(x)(1-F(x))}{n} + 0 = \frac{1}{n}\cdot \overbrace{F(x)(1-F(x))}^{\text{קבוע}} = \frac{1}{n}\cdot C =$$

$$\frac{1}{n}\cdot C \xrightarrow{n\to\infty} 0$$

ולכן האומד עקיב.

2)

א)

נזכיר כי $I\{X_i \le x\} \sim ber(F(x))$ לכל i

מכאן נובע כי $\hat{F}_n(x)$ הינו ממוצע של מ"מ ש.ה ב.ת

בעלי שונות $F(x)(1-F(x))$, ותוחלת $F(x)$ ולכן

עפ"י משפט הגבול המרכזי

$$\sqrt{n}\,\frac{\hat{F}_n(x)-F(x)}{\sqrt{F(x)(1-F(x))}} \xrightarrow{D} N(0,1)$$

לאחר ליפה הכפלה והעברת אגפים נגיע ל

$$\hat{F}_n(x) \xrightarrow{D} N\left(F(x), \frac{F(x)(1-F(x))}{n}\right)$$

ב)

$$T(F) = P(a < X \le b) = P(X \le b) - P(X \le a) = F(b) - F(a) =$$

$$\int_a^b f(x) = \int_{dom(x)} I\{a < x \le b\} f(x)dx = \int_{dom(x)} I\{a < x \le b\} dF(x)$$

ולכן לינארי לפי הגדרה.

המשך בעמוד הבא

המשך שאלה 2 סעיף ב'

אומד ה-plugin של θ יהיה $\hat{\theta} = T(\hat{F}_n) = \frac{1}{n}\sum_i 1\{a < x_i \le b\}$

$$T(\hat{F}_n) = \frac{1}{n}\sum_i P(x_i \le b) - P(x_i < a) = \hat{F}_n(b) - \hat{F}_n(a)$$

ולכן לפי סעיף הקודם מתקיים

$$\hat{F}_n(b) - \hat{F}_n(a) \xrightarrow{D} N\left(F(b) - F(a), \frac{F(b)(1-F(b))}{n} + \frac{F(a)(1-F(a))}{n}\right)$$

ומכאן, רווח הסמך הוא:

$$T(\hat{F}_n) \pm Z_{\frac{\alpha}{2}} \cdot \sqrt{\frac{F(b)(1-F(b)) + F(a)(1-F(a))}{n}}$$

שאלה 3

$$\operatorname{cov}(\hat{F}_n(x), \hat{F}_n(y)) = E(\hat{F}_n(x)\cdot \hat{F}_n(y)) - E[\hat{F}_n(x)]\cdot E[\hat{F}_n(y)] =$$

$$= E\left(\left(\frac{1}{n}\sum 1\{x_i \le x\}\right)\cdot\left(\frac{1}{n}\sum 1\{x_i \le y\}\right)\right) - F(x)\cdot F(y) =$$

$$= \frac{1}{n^2}\left(E\left(\sum 1\{x_i \le \min\{x,y\}\}\right) + \sum_{i\neq j} E\, 1\{x_i \le x\} 1\{x_j \le y\}\right) - F(x)F(y) =$$

$$\frac{1}{n}\cdot F(\min\{x,y\}) + \frac{1}{n^2}\sum_{i\neq j} F(x)F(y) - F(x)F(y) =$$

$$= \frac{1}{n}\cdot F(\min\{x,y\}) + \left(\frac{1}{n^2}\cdot\frac{n(n-1)}{2} - 1\right) F(x)F(y) =$$

$$= \frac{1}{n} F(\min\{x,y\}) - \frac{n+1}{2n} F(x)F(y)$$

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

<u>Need to show:</u>

$$\frac{\hat{g}_{1-\frac{\alpha}{2}}}{\sqrt{n}} = \hat{\theta}^*_{1-\frac{\alpha}{2}} - \hat{\theta}_n$$

$$\hat{g}_{\frac{\alpha}{2}} = \inf\{x : G_n(x) \geq \tfrac{\alpha}{2}\} = \inf\{x : \tfrac{1}{B}\sum_{b=1}^{B} I[\sqrt{n}(\hat{\theta}_n^{*b} - \hat{\theta}_n) \leq x] \geq \tfrac{\alpha}{2}\}$$

$$= \inf\{x : \sum_{b=1}^{B} I[\sqrt{n}(\hat{\theta}_n^{*b} - \hat{\theta}_n) \leq x] \geq \tfrac{B\alpha}{2}\}$$

$$= \sqrt{n} \cdot \inf\{x : \tfrac{1}{B}\sum_{b=1}^{B} I[\hat{\theta}_n^{*b} - \hat{\theta}_n \leq x] \geq \tfrac{\alpha}{2}\}$$

$$= \sqrt{n}\left(\inf\{x : \tfrac{1}{B}\sum_{b=1}^{B} I[\hat{\theta}_n^{*b} \leq x] \geq \tfrac{\alpha}{2}\} - \hat{\theta}_n\right) = \sqrt{n} \cdot (\hat{\theta}^*_{\frac{\alpha}{2}} - \hat{\theta}_n)$$

Therefore:

$$\begin{cases} \hat{g}_{\frac{\alpha}{2}} = \sqrt{n}\,(\hat{\theta}^*_{\frac{\alpha}{2}} - \hat{\theta}_n) \\ \hat{g}_{1-\frac{\alpha}{2}} = \sqrt{n}\,(\hat{\theta}^*_{1-\frac{\alpha}{2}} - \hat{\theta}_n) \end{cases} \Rightarrow CI = [2\hat{\theta}_n - \hat{\theta}^*_{1-\frac{\alpha}{2}},\ 2\hat{\theta}_n - \hat{\theta}^*_{\frac{\alpha}{2}}]$$

Question 5:

Notice:

$$std\big(med(X) - med(Y)\big) = \sqrt{var\big(med(X) - med(Y)\big)} = \sqrt{var\big(med(X)\big) + var\big(med(Y)\big)}$$

Therefore, we are required only to estimate the variance of the median.

In order to do so, we calculate the empirical CDF of the data, denoted $\widehat{F_n^X}$.

For $i = 1, \dots, B$:

Sample n data points from X, denoted $X_1^b, X_2^b, \dots$

Calculate $\widehat{T_b} = med\big(X_1^b, \dots, X_n^b\big)$

The estimator for $Var(T)$ is hence:

$$\frac{1}{B}\sum_{b=1}^{B}\big(\widehat{T_b}\big)^2 - \left(\frac{1}{B}\sum_{b=1}^{B}\widehat{T_b}\right)^2$$

The same goes for Y, so we plug back into the original formula:

$$std\big(med(X) - med(Y)\big) = \sqrt{var\big(med(X)\big) + var\big(med(Y)\big)}$$

```
import pandas as pd
import numpy as np
```

(a)

```
def empirical_dist(data, x):
    f = np.sum(data == x) / data.shape[0]
    return f

def plug_in_mean(data):
    return sum([val * empirical_dist(data, val) for val in data])

def plug_in_var(data):
    return sum([(val ** 2) * empirical_dist(data, val) for val in data]) - (plug_
```

```
df = pd.read_csv("ex5.csv")
lsat_mean = df['LSAT'].mean()
gpa_mean = df['GPA'].mean()
```

```
# calculate plug-in estimator using formula from tutorial
def corr_estimator(df):
    plug_in_corr = ((df['LSAT'] - lsat_mean) * (df['GPA'] - gpa_mean)).sum() / \
                np.sqrt(((df['LSAT'] - lsat_mean) ** 2).sum() * ((df['GPA'] - gpa
    return plug_in_corr
```

```
corr = corr_estimator(df)
print(f"The plug-in estimate for the correlation coefficient between LSAT score a
```

```
The plug-in estimate for the correlation coefficient between LSAT score and GPA
```

(b)

```python
B = 1000
corr_df = np.zeros(B)
for i in range(B):
    boot = df.sample(n=15, replace=True)
    corr_df[i] = corr_estimator(boot)

se_boot = np.std(corr_df)
print(f"Std: {se_boot}")
```

```
Std: 0.13374187005828042
```

(c)

```python
# Gaussian approximation
print(f"Confidence interval for correlation coefficient under Gaussian assumption
print("[{corr - 2*se_boot}, {corr + 2*se_boot}]")
# Pivotal approximation
low = 2*corr - np.quantile(corr_df, 0.975)
high = 2*corr - np.quantile(corr_df, 0.025)
print(f"Pivotal confidence interval for correlation coefficient: ")
print(f"[{low}, {high}]")
# Quantile based approximation
print(f"Quantile-based confidence interval for correlation coefficient: ")
print(f"[{2*corr - high}, {2*corr - low}]")
```

```
Confidence interval for correlation coefficient under Gaussian assumption:
[{corr - 2*se_boot}, {corr + 2*se_boot}]
Pivotal confidence interval for correlation coefficient:
[0.5954261057122939, 1.0970028227440025]
Quantile-based confidence interval for correlation coefficient:
```

```
[0.4557461598348116, 0.9573228768665202]
```