סטטיסטיקה למדעי המחשב – תרגיל בית 5

הצעה לפתרון

שאלה 1:

א.

$X \sim U(1,6)$ ולכן נשתמש בנוסחאות של התפלגות אחידה בדידה. נקבל:

הממוצע (תוחלת) של X – $\frac{1+6}{2} = 3.5$

סטיית התקן של X – $\sqrt{\frac{(6-1+1)^2-1}{12}} = 1.7078$

נשים לב שהטלות שונות של הקובייה הן בלתי תלויות. נסמן Y1, Y2 את שתי ההטלות ועבור X2 נקבל:

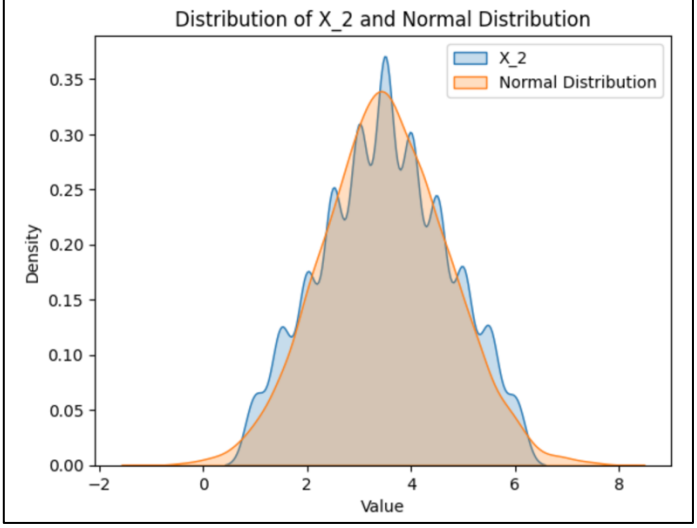
הממוצע (תוחלת) של X2 – $E[X_2] = E\left[\frac{Y_1+Y_2}{2}\right] = \frac{1}{2} * (E[Y_1] + E[Y_2]) = \frac{1}{2} * (3.5 + 3.5) = 3.5$

סטיית התקן של X2 – $\sqrt{Var(X_2)} = \sqrt{Var\left(\frac{Y_1+Y_2}{2}\right)} = \sqrt{\frac{1}{4}\left(Var(Y_1) + Var(Y_2)\right)} = \sqrt{\frac{1}{4} * 2 * 1.7078} = 1.2076$

ב', ג', ד':

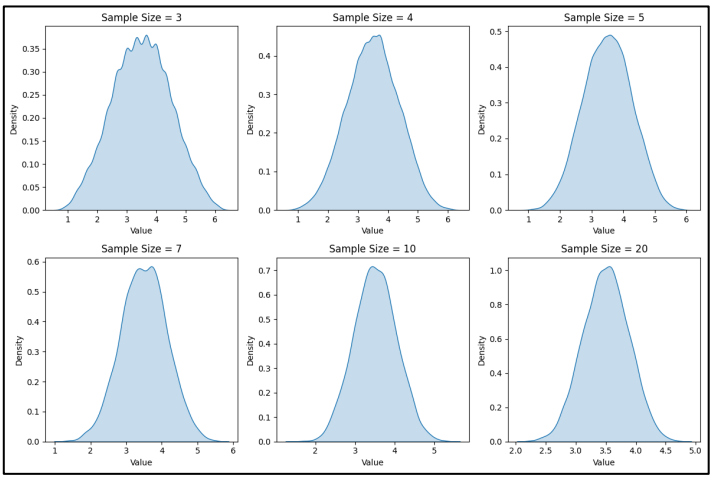
```python
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# Define number of trials
num_trials = 10000

# Rolling dice and calculating values of X and X_2
X = np.random.randint(1, 7, num_trials)
X_2 = (np.random.randint(1, 7, num_trials) + np.random.randint(1, 7, num_trials)) / 2

# Calculating mean and standard deviation of X
mean_X = np.mean(X)
std_X = np.std(X)

# Calculating mean and standard deviation of X_2
mean_X_2 = np.mean(X_2)
std_X_2 = np.std(X_2)

print("Mean of X: {:.2f}, Standard Deviation of X: {:.2f}".format(mean_X, std_X))
print("Mean of X_2: {:.2f}, Standard Deviation of X_2: {:.2f}".format(mean_X_2, std_X_2))

# Plotting distribution of X_2
sns.kdeplot(X_2, fill=True)

# Adding normal distribution with mean and standard deviation of X_2
normal_dist = np.random.normal(mean_X_2, std_X_2, num_trials)
sns.kdeplot(normal_dist, fill=True)

plt.xlabel('Value')
plt.ylabel('Density')
plt.title('Distribution of X_2 and Normal Distribution')
plt.legend(['X_2', 'Normal Distribution'])
plt.show()
```

```
Mean of X: 3.53, Standard Deviation of X: 1.71
Mean of X_2: 3.50, Standard Deviation of X_2: 1.21
```



ה. להלן כמה דוגמאות:

```python
import numpy as np
import seaborn as sns
import matplotlib.pyplot as plt

# Define number of trials
num_trials = 10000
sample_sizes = [3, 4, 5, 7, 10, 20]

# Create subplots
fig, axs = plt.subplots(2, 3, figsize=(12, 8))
axs = axs.flatten()

for i, size in enumerate(sample_sizes):
    X_n = np.zeros(num_trials)
    for _ in range(size):
        X_n += np.random.randint(1, 7, num_trials)
    X_n /= size  # Calculate the average

    # Plot KDE
    sns.kdeplot(X_n, ax=axs[i], fill=True)
    axs[i].set_title(f"Sample Size = {size}")
    axs[i].set_xlabel("Value")
    axs[i].set_ylabel("Density")

plt.tight_layout()
plt.show()
```



ו.

נניח ש – n גדול דיו כך שמשפט הגבול המרכזי מתקיים –

$$\bar{X} = \frac{1}{n}\sum_{i=1}^{n} X_i \sim N\left(E[X], \frac{V(X)}{n}\right)$$

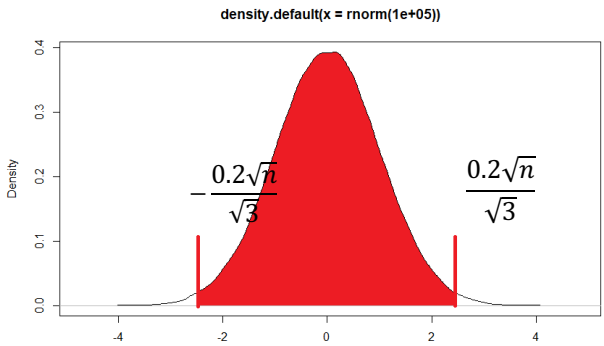
נדרוש

$$P(|\bar{X} - E[X]| < 0.1) \geq 0.99$$

נפתח את הביטוי משמאל

$$P(-0.1 < \bar{X} - E[X] < 0.1) = P\left(-\frac{0.1}{\sqrt{\frac{V(X)}{n}}} < \frac{\hat{p} - 0.25}{\sqrt{\frac{V(X)}{n}}} < \frac{0.1}{\sqrt{\frac{V(X)}{n}}}\right)$$
$$= P\left(-\frac{0.1\sqrt{n}}{\sqrt{V(X)}} < Z < \frac{0.1\sqrt{n}}{\sqrt{V(X)}}\right) = 1 - 2 * P\left(Z < -\frac{0.1\sqrt{n}}{\sqrt{V(X)}}\right)$$

המעבר האחרון הוא מטעמי סימטריה של ההתפלגות הנורמלית הסטנדרטית:



$$\Rightarrow 1 - 2 * P\left(Z < -\frac{0.1\sqrt{n}}{\sqrt{V(X)}}\right) \geq 0.99 \Rightarrow P\left(Z < -\frac{0.1\sqrt{n}}{\sqrt{V(X)}}\right) \leq 0.005$$

$$\Rightarrow -\frac{0.1\sqrt{n}}{\sqrt{V(X)}} \leq Z_{0.005} = -2.575$$

$$\sqrt{n} \geq \frac{2.575 * 1.708}{0.1} \rightarrow n \geq 43.91 \Rightarrow n = 44$$

שאלה 2:

א. אוכלוסיית המחקר היא תושבי סנטה קלרה, מהם נדגמו 3,300 איש. הפרמטר הוא הימצאות הנוגדנים, או ההסתברות להמצאות נוגדנים באוכלוסייה. האומד לפרמטר הוא פרופורציית האנשים במדגם להם הבדיקה הראתה שיש נוגדנים. האומדן הוא 50/3300 = 0.15.

ב. אוכלוסיית המחקר היא עכברים, מתוכם נלקח מדגם של 12. יש שני פרמטרים מבוקשים - תוחלת המרחק המצטבר באוכלוסייה ("רמת פעילות"), וההסתברות לשהות ליד הקיר ("רמת חרדה"). האומדים הם ממוצע המרחקים המצטברים וממוצע ההסתברויות במדגם. האומדנים הם 371.3 ס"מ ו-0.63 בהתאמה.

ג. יש כאן שתי אוכלוסיות שמשווים ביניהן - משתמשי טינדר ואנשים שלא משתמשים בטינדר. מהאוכלוסייה הראשונה דגמו 182 אנשים, ומהשנייה 89, כאשר סה"כ נדגמו 271 אנשים. הפרמטר הנבדק הוא תוחלת ה"פסיכופתיה" (מידת ההסכמה עם המשפטים). האומדן הוא 2.83 במשתמשי טינדר ו-2.63 בשאר.

שאלה 3:

א. $\hat{\mu}^{(1)} = 60$ מוטה כיוון לא מתקיים $\mu = 60 = E[60]$ לכל ערך פרמטר μ.

$\hat{\mu}^{(2)} = \bar{X}$ בלתי מוטה, ראינו שהממוצע תמיד אומד חסר הטיה לתוחלת.

עבור $\hat{\mu}^{(3)} = \frac{\sum X_i + 120}{n+2}$ נקבל:

$$E\left[\hat{\mu}^{(3)}\right] = E\left[\frac{\sum X_i + 120}{n+2}\right] \underset{\text{ליניאריות התוחלת}}{=} \frac{\sum E[X_i] + 120}{n+2} = \frac{n\mu + 120}{n+2} \neq \mu$$

ולכן מוטה.

$$Var\left(\hat{\mu}^{(1)}\right) = Var(60) = 0$$

$$Var\left(\hat{\mu}^{(2)}\right) = Var(\bar{X}) = Var\left(\frac{X_1 + \cdots + X_n}{n}\right) = \frac{n \cdot Var(X_i)}{n^2} = \frac{n \cdot \sigma^2}{n^2} = \frac{\sigma^2}{n} = \frac{25}{10} = 2.5$$

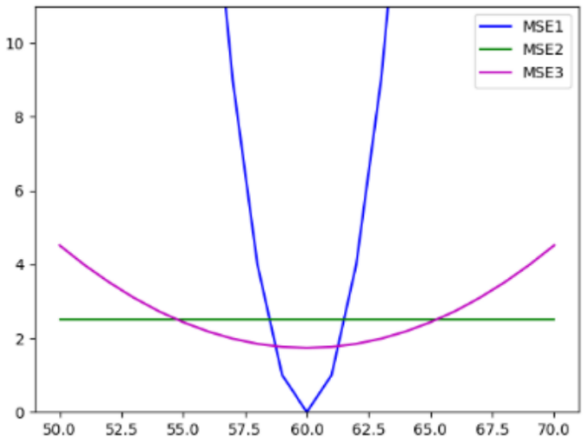
$$Var\left(\hat{\mu}^{(3)}\right) = Var\left(\frac{\sum X_i + 120}{n+2}\right) = \frac{1}{(n+2)^2} \cdot Var\left(\sum X_i\right) \overset{\text{ב"ת}}{=} \frac{1}{(n+2)^2} \cdot nVar(X)$$
$$= \frac{1}{(n+2)^2} \cdot n\sigma^2 = \frac{1}{144} \cdot 10 \cdot 25 = 1.736$$

ולכן ל-$\hat{\mu}^{(1)}$ שונות מינימלית.

ב.

$$MSE\left(\mu, \hat{\mu}^{(1)}\right) = E[(60-\mu)^2] = \boldsymbol{(60-\mu)^2}$$
$$MSE\left(\mu, \hat{\mu}^{(2)}\right) = Var\left(\hat{\mu}^{(2)}\right) + \left(E\left[\hat{\mu}^{(2)}\right] - \mu\right)^2 = 2.5 + (\mu-\mu)^2 = \mathbf{2.5}$$
$$MSE\left(\mu, \hat{\mu}^{(3)}\right) = Var\left(\hat{\mu}^{(3)}\right) + \left(E\left[\hat{\mu}^{(3)}\right] - \mu\right)^2 =$$
$$= \frac{n\sigma^2}{(n+2)^2} + \left(\left(\frac{n\mu + 120}{n+2}\right) - \mu\right)^2 =$$
$$= \frac{n\sigma^2}{(n+2)^2} + \left(\frac{n\mu + 120 - n\mu - 2\mu}{n+2}\right)^2 =$$
$$= \frac{n\sigma^2}{(n+2)^2} + \left(\frac{120 - 2\mu}{n+2}\right)^2 = \frac{n\sigma^2 + (120-2\mu)^2}{(n+2)^2} =$$
$$= \mathbf{1.736} + \left(\mathbf{10} - \frac{\boldsymbol{\mu}}{\mathbf{6}}\right)^2$$

ג.



בכל קטע של ערכי טטא נעדיף את האומד בעל ה-MSE המינימלית. נראה שאם יש לנו סיבה להאמין שהתוחלת סביב 60 (אך לא בדיוק) יתכן והאומד המוטה $\hat{\mu}^{(3)}$ עדיף. אחרת, הממוצע עדיף.

ד. נחשב:

```
arr = np.array([81, 95, 85, 75, 98, 100, 85, 86, 92, 91])
m1 = 60
m2 = arr.mean()
m3 = (sum(arr) + 120) / 12
print("m1 =", m1)
print("m2 =", m2)
print("m3 =", m3)
```

```
m1 = 60
m2 = 88.8
m3 = 84.0
```

<u>שאלה 4:</u>

א. פתרון:

$$E(T_1) = \theta \rightarrow Var(T_1) = E(T_1^2) - E(T_1)^2 = E(T_1^2) - \theta^2 = \sigma^2$$

$$E(T_1^2) = \theta^2 + \sigma^2 \rightarrow Bias(T_1^2, \theta^2) = \sigma^2 + \theta^2 - \theta^2 = \sigma^2$$

ב. פתרון:

$$MSE\left(\hat{\mu}^{(1)}\right) = Var\left(\hat{\mu}^{(1)}\right) = \frac{\mu^2 + 1}{n}$$

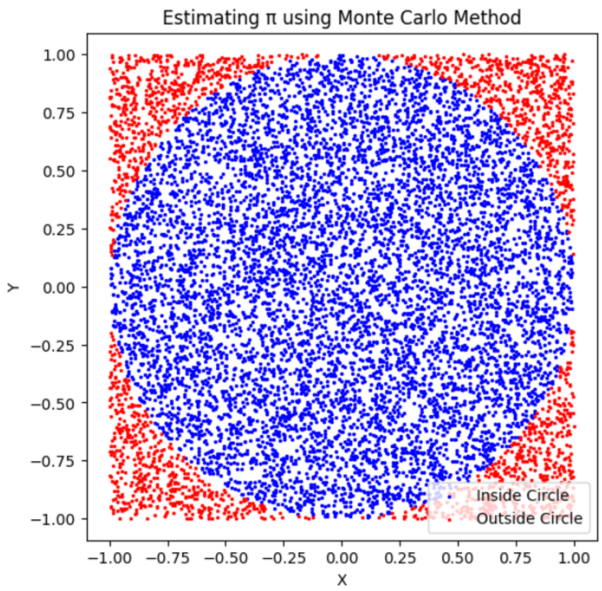
$$MSE\left(\hat{\mu}^{(2)}\right) = Var\left(\hat{\mu}^{(2)}\right) = \mu^2 + 1$$

$$MSE\left(\hat{\mu}^{(3)}\right) = \left(E[\hat{\mu}^{(3)}] - \mu\right)^2 = \mu^2$$

פתרון לשאלת בונוס:

נשתמש בעובדה שהסיכוי שנקודה אקראית שנבחרת בתחום הדו מימדי $[-1,1]^2$ תיפול בדיוק במעגל שרדיוס 1 ומרכזו 0, היא שטח המעגל חלקי שטח הצורה כולה (ריבוע עם צלע באורך 2). כלומר הסיכוי הוא $p = \frac{\pi}{4}$. כך, אם נאמוד את p נוכל ע״י פרופורציית התצפיות במדגם שנפלו בתוך העיגול, נוכל לאמוד גם את π ע״י הכפלה ב4. האומד יהיה בלתי מוטה. נקודה אקראית כזו ניתן לייצר ע״י דגימה של שתי תצפיות בלתי תלויות ב-$[-1,1]$. להלן הדגמה:

```python
import random

def estimate_pi(num_points):
    points_inside_circle = 0
    for _ in range(num_points):
        x = random.uniform(-1, 1)
        y = random.uniform(-1, 1)
        if x**2 + y**2 <= 1:
            points_inside_circle += 1
    return 4 * points_inside_circle / num_points

num_points = 1000000  # You can adjust this number for better accuracy
estimated_pi = estimate_pi(num_points)
print("Estimated value of π:", estimated_pi)
```

```
Estimated value of π: 3.144812
```

והמחשה: