סטטיסטיקה למדעי המחשב – תרגיל בית שבוע 5

שאלה 1 (30 נקודות, 5 נקודות לכל סעיף)

ידוע כי ההתפלגות של משקל של עציץ היא נורמלית בעלת תוחלת של 81.5 ק"ג וסטיית תקן של 13 ק"ג. חשבו: (ניתן להשתמש בפייתון)

א. מה ההסתברות שעציץ בודד ישקול מעל 82 ק"ג?

ב. מה ההסתברות ש-50 עציצים ישקלו כולם מעל 82 ק"ג?

ג. מה ההסתברות שהממוצע של 50 עציצים יהיה גבוה מ82 ק"ג?

ד. מה ההסתברות שמשקלם הממוצע של 25 עציצים יהיה בין 80 ל-82 ק"ג?

ה. מה המשקל ש-10% עציצים גדולים ממנו?

ו. בבדיקה חוזרת התבררה טעות במדידת הקודמות של העציצים שעליהם התבססה ההתפלגות עד כה. המידע היחיד שנותר הוא שמשקלם מתפלג נורמאלית, האחוזון ה-70 הוא 90 ק"ג והאחוזון ה-20 הוא 70 ק"ג. מהי התוחלת וסטיית התקן של התפלגות משקלי העציצים כעת?

שאלה 2 (30 נקודות, 5 נקודות לכל סעיף)

הדגימו את משפט הגבול המרכזי בפייתון:

א. הראו בטבלה את פונקציית ההתפלגות של תוצאת הטלת קובייה (מ"מ X). מה התוחלת וסטיית התקן של X? נסמן ב- X_2 ממוצע מקרי של **2** הטלות קובייה. מה התוחלת וסטיית התקן של X_2?

ב. הגרילו את X_2 עשרות אלפי פעמים וציירו את התפלגות שלו בעזרת הפונקציה kdeplot של seaborn.

ג. מה הממוצע וסטיית התקן האמפיריות של ההתפלגות שיצרתם? האם זה מתאים לחישוב מסעיף א'?

ד. הוסיפו לשרטוט את הצפיפות של התפלגות נורמלית עם אותם פרמטרים (תוחלת ושונות) כמו X_2. האם ניתן להגיד ש X_2 מתפלג בקירוב נורמלית?

ה. הגדילו בהדרגה את גודל המדגם עבורו מחושב הממוצע X_n (זאת אומרת, את מספר הטלות הקובייה המרכיבות את הממוצע) וחיזרו על סעיפים ב' ו-ג'. הראו לפחות 4 גדלי מדגם שונים.

ו. מאיזה גודל מדגם (בערך) הייתם אומרים שהממוצעים מתפלגים נורמאלית? עבור המקרה הזה, השתמשו ב CLT על מנת למצוא את השברון ה 0.95 באופן מקורב.

ז. (5 נקודות בונוס) בחרו משתנה מיקרי רציף המתפלג לא נורמלית וחיזרו על סעיפים א' עד ה' (לא כולל הצגת טבלה של פונקציית ההתפלגות).

שאלה 3 (20 נקודות, 10 לכל סעיף)

א. נתונות מדגם בגודל 2: X_1, X_2 בלתי תלויים ושוות התפלגות המתפלגות לפי פונקציית הצפיפות הבאה:

$$f_\theta(x) = \theta e^{-\theta x}, \qquad x \geq 0, \qquad else\ 0$$

נתונים שני האומדים הבאים לפרמטר ההתפלגות $\theta > 0$:

$$\widehat{\theta_1} = \frac{2X_1 + 4X_2}{6}$$

$$\widehat{\theta_2} = \frac{4X_1 + 5X_2}{6}$$

מצאו את ה-MSE של שני האומדים. האם אחד מהם עדיף?

ב. נתון מדגם בגודל n עם דגימות בלתי תלויות ושוות התפלגות המתפלגות לפי פונקציית הצפיפות הבאה:

$$f_\theta(x) = \begin{cases} \frac{3x^2}{\theta^3} & 0 \leq x \leq \theta \\ 0 & else \end{cases}$$

נתון האומד הבא לפרמטר ההתפלגות $\theta > 0$:

$$\hat{\theta} = \frac{4}{3}\bar{X}$$

מצאו את השונות וההטיה של האומד.

שאלה 4 (20 נקודות, 5 נקודות לכל סעיף)

הציון במבחן מסוים (X) הינו משתנה מקרי עם תוחלת μ (לא ידוע) וסטיית תקן 5. חוקרים מעוניינים ללמוד את התוחלת באמצעות מדגם מקרי המונה n=10 ציונים. לשם כך הם הציעו את שלושה האומדים הבאים:

$$\hat{\mu}^{(1)} = 60 \quad ; \quad \hat{\mu}^{(2)} = \bar{X} \quad ; \quad \hat{\mu}^{(3)} = \frac{\sum X_i + 120}{n+2}$$

א. אילו מהאומדים חסרי הטיה? לאיזה מהם שונות מינימלית?

ב. בטאו את תוחלת השגיאה הריבועית של כל אחד מהאומדים באמצעות הפרמטר μ.

ג. ציירו על גרף את ערכי השגיאה הריבועית כפונקציה של μ עבור ערכים בין 50 ל-70. הגבילו את ציר y ל-10 לכל היותר. מה ניתן ללמוד מההשוואה הזו? באילו מצבים נעדיף להשתמש ב $\hat{\mu}^{(3)}$ על פני $\hat{\mu}^{(2)}$?

ד. במדגם מסוים של 10 תצפיות התקבלו הציונים הבאים:
81,95,85,75,98,100,85,86,92,91. חשבו את שלושת האומדנים על סמך מדגם זה. באיזה מהם הייתם בוחרים כדי להעריך את μ האמיתי?