סטטיסטיקה למדעי המחשב – תרגיל בית שבוע 6

שאלה 1 (25 נקודות)

בשאלה זו נשווה בין הביצועים של אומדים שונים על ידי סימולציה בפייתון:

חזרו על התהליך עבור כל אחת מההתפלגויות המופיעות מטה:

1. הגרילו $K = 10^4$ מדגמים מיקריים (**x**) בגודל 10 כל אחד. עבור כל מדגם יש לחשב שני אומדנים בהתאם לאומדים שיוגדרו לפרמטר מסוים.
2. לאחר מכן יש לחשב את רכיבי השגיאה הריבועית של האומדים ביחס לפרמטר האמיתי, על סמך *ה"אוכלוסייה"* בגודל K של התוצאות שהתקבלו. עבור התפלגות אחידה, השוו את התוצאות האמפיריות לחישוב תיאורטי על מנת לוודא שאין טעויות בקוד.
3. הציגו היסטוגרמה חלקה (פונקציית kdeplot() של seaborn) של שני האומדים (ביחד), הוסיפו גם את ערך הפרמטר האמיתי בקו אנכי. הסבירו את התוצאות במילים שלכם.
4. הגרילו **מדגם מקרי מתוך האוכלוסייה** בגודל 20 וחשבו אומדנים לתוחלת ולשונות של אחד האומדים על סמך מדגם זה. הסבירו כיצד ביצעתם את הדגימה ואת החישובים הרצויים.

עבור התפלגות אחידה: $X \sim U(0,\theta)$ (12.5 נקודות)

א. קבעו את הערך האמיתי להיות: $\theta = 5$.

ב. השוו בין שני האומדים הבאים: $\hat{\theta}^{(1)} = 2mean(\boldsymbol{x})$; $\hat{\theta}^{(2)} = max(\boldsymbol{x})$.

עבור התפלגות מעריכית: $X \sim exp(\lambda)$ (12.5 נקודות)

א. קבעו את הערך האמיתי להיות: $\lambda = 1$.

ב. השוו בין: $\hat{\lambda}^{(1)} = 1/mean(\boldsymbol{x})$; $\hat{\theta}^{(2)} = \ln(2)/median(\boldsymbol{x})$.

עבור התפלגות נורמלית: $X \sim N(\mu, 25)$ (**בונוס** 5 נקודות), שימו לב שסטיית התקן של ההתפלגות היא 5 והיא ידועה.

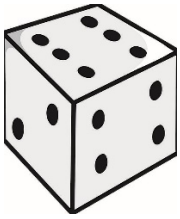
א. קבעו את הערך האמיתי להיות: $\mu = 62$.

ב. השוו בין: $\hat{p}^{(1)}$, $\hat{p}^{(2)}$, כאשר p זה הסיכוי ש: X≤60.

1. האומד $\hat{p}^{(1)}$ צריך להיות חסר הטיה ומבוסס על התפלגות ברנולי.
2. האומד $\hat{p}^{(2)}$ מבוסס על הערך $mean(\boldsymbol{x})$.

שאלה 2 (25 נקודות, 5 נקודות לכל סעיף)

נתון הצילום של הקוביה הבאה:



ידוע שהקוביה הינה הוגנת (סיכוי שווה לכל אחת מהפאות) אך לא ידוע מהם המספרים שכתובים בפאות שאינן מופיעות בתמונה. ידוע שהספרה '1' כתובה לפחות פעם אחת ומעוניינים ללמוד את הסיכוי p שתתקבל התוצאה '1' בקוביה.

א. מהם ערכים אפשריים ל- p? רשמו ביטוי לפונקציית הנראות של p בהינתן מדגם בגודל n של תוצאות בינאריות - '0' אם לא התקבל '1' ו- '1' אחרת.

ב. במדגם מקרי של 12 תצפיות התקבלה התוצאה '1' 3 פעמים. מצאו אומדן ל-p בשיטת נראות מקסימלית בהתבסס על מדגם זה.

ג. כעת, נתון גם שבפאות המוסתרות שבהן לא מופיע '1', מופיע המספר '5', ויש לפחות פאה אחת כזו. מצאו אומדן לתוחלת של תוצאת הקוביה בהתבסס על המדגם מסעיף ב' בשיטת נראות מקסימלית.

ד. כעת התקבלו תוצאות הטלת הקוביות באופן מלא: {1,1,1,6,4,5,2,5,4,5,2,5}. מצאו אומדן ל-p ולתוחלת תוצאת הקוביה עבור המדגם בשיטת נראות מקסימלית (בהתבסס על מדגם המלא).

ה. הגדירו אומד אחר לתוחלת, שאינו משתמש באומד ל-p, וחשבו את האומדן לפי מדגם זה. הסבירו במילים שלכם איזה מהאומדים לתוחלת עדיף (האומד שהוצע בסעיף זה או האנ"מ). השתמשו בשיקולי הטיה ושונות.

ו. (**בונוס** 5 נקודות) הראו באמצעות סימולציה בפייתון איזה אומד עדיף לתוחלת. הסבירו את התוצאות במילים שלכם.

שאלה 3 (30 נקודות, 6 נקודות לכל סעיף)

בנקודה מסוימת ב-DNA האנושי יכולה להופיע האות "A" (בסיכוי $0<p<1$,p) או האות "G" (בסיכוי 1-p). לכל אדם יש 2 כרומוזומים, שבכל אחד מהם קיימת נקודה זו. כלומר לכל אדם יש 4 אפשרויות:

AA – בשני הכרומוזומים מופיע "A"

AG – בכרומוזום א' מופיע "A" ובכרומוזום ב' מופיע "G"

GA – בכרומוזום א' מופיע "G" ובכרומוזום ב' מופיע "A"

GG – בשני הכרומוזומים מופיע "G"

א. הניחו שאין תלות בין שני הכרומוזומים. בטאו באמצעות p את ההסתברויות הבאות:

1. p1 - ההסתברות שבשני הכרומוזומים של אדם כלשהו מופיע "A"
2. p2 - ההסתברות שאצל אדם כלשהו מופיעה גם האות "A" וגם האות "G"
3. p3 - ההסתברות שבשני הכרומוזומים של אדם כלשהו מופיע "G"

ב. רשמו בכתיב מתמטי של כפולות וחזקות את פונקציית ההתפלגות עבור מ"מ X שיכול לקבל 3 ערכים בהתאם לשלושת האפשרויות הנ"ל.

ג. נדגמו n אנשים. מתוכם, אצל y1 נבדקים האות "A" הופיעה פעמיים, אצל y2 נבדקים הופיעו גם "A" וגם "G", ואצל y3 נבדקים האות "G" הופיעה פעמיים. רשמו את פונקציית הנראות של p כפונקציה של y1, y2, y3 וn-, ומצאו אומד נראות מקסימלית ל-p.

ד. אם האומד חסר הטיה? נמקו.

ה. הסיכוי שצבע שערו של אדם, שנושא גם את האות "A" וגם את האות "G", יהיה סגול, הוא 0.3. אם אדם נושא פעמיים "G" שערו יהיה סגול בוודאות, ואחרת הוא לא יהיה סגול. נסמן בq את הסיכוי שצבע שערו של אדם יהיה סגול. בטאו את q במונחים של p. השתמשו בכך כדי למצוא אומד נראות מקסימלית עבור q בהינתן y1, y2, y3 וn-.

שאלה 4 (20 נקודות, 5 נקודות לכל סעיף)

ידוע שהזמן בין רעידות אדמה בחבל ארץ מסוים מתפלג מעריכית. חוקר א' החליט לאסוף נתונים באופן אקראי על מרווחי זמן (בשנים) בין רעידות אדמה: $X_1 \cdots X_{20}$. חוקר ב' החליט לאסוף לבדוק את כמות רעידות אדמה בשנים ספציפיות שנדגמו אקראית: $Y_1 \cdots Y_{20}$ (רמז: איזה התפלגות מתארת מספר אירועים בפרק זמן מסוים?).

א. כיצד מתפלגים הנתונים בכל מדגם? כתבו את פונקציית הנראות כפונקציה של פרמטר הקצב λ עבור כל אחד מהמדגמים ($X_1 \cdots X_{20}$ ו- $Y_1 \cdots Y_{20}$).

ב. כתבו ביטוי לאנ"מ בכל עבור כל אחד מהמדגמים. האם הם חסרי הטיה?

ג. השוו בין הביצועים (MSE) של שני האנ"מים על ידי סימולציה בפייתון עבור $\lambda \in \{0.1,10\}$. קבעו איזו שיטת דגימה עדיפה בכל אחד המקרים. הסבירו באופן אינטואיטיבי מדוע.

ד. שני החוקרים מעוניינים לאמוד את הסיכוי שלא תהיה רעידת אדמה בשנה הקרובה. מהו הביטוי לאנ"מ לסיכוי זה על סמך כל אחד מהמדגמים.