סטטיסטיקה למדעי המחשב – תרגיל בית שבוע 7

שאלה 1 (25 נקודות)

בשאלה זו נדגים תכונות של רווחי סמך על ידי סימולציה בפייתון:

עבור כל סט ערכים של α , n, יש להגריל $K = 10^4$ מדגמים מיקריים בגודל n מהתפלגות נורמלית עם תוחלת 175 ושונות 100. עבור כל מדגם יש למצוא:

1. גבול עליון ותחתון של רווח סמך (עם שונות ידועה) **לתוחלת** ברמת סמך של $1 - \alpha\%$.
2. אורך הקטע של רווח הסמך הנ"ל.
3. האם רווח הסמך מכיל את התוחלת האמיתית.
4. מהו הסיכוי שדגימה מקרית נוספת מתוך ההתפלגות האמיתית "תיפול" בתוך רווח הסמך שחושב למדגם זה?

א. בצעו את הניסוי הנ"ל עבור $\alpha = 0.05$, $n \in \{10,20,40,80\}$. ומצאו על סמך K המדגמים:

1. מהו אורך הקטע של רווח הסמך בממוצע? מהי סטיית התקן?
2. מהו הסיכוי שרווח הסמך מכיל את התוחלת האמיתית.
3. הסיכוי שדגימה מקרית נוספת מתוך ההתפלגות האמיתית "תיפול" בתוך רווח הסמך - בממוצע על פני K המדגמים.

הסבירו את הקשר שמצאתם בין גודל המדגם לכל אחד מהערכים הנ"ל.

ב. בצעו את הניסוי הנ"ל עבור $n = 30$, $\alpha \in \{0.05,0.1,0.2\}$. ומצאו על סמך K המדגמים:

1. מהו אורך הקטע של רווח הסמך בממוצע? מהי סטיית התקן?
2. מהו הסיכוי שרווח הסמך מכיל את התוחלת האמיתית.
3. הסיכוי שדגימה מקרית נוספת מתוך ההתפלגות האמיתית "תיפול" בתוך רווח הסמך - בממוצע על פני K המדגמים.

הסבירו את הקשר שמצאתם בין רמת הסמך לכל אחד מהערכים הנ"ל.

שאלה 2 (25 נקודות, 5 נקודות לכל סעיף)

נתון מדגם מקרי מהתפלגות נורמלית $X_1, \dots, X_n \sim N(\mu, \sigma^2)$, כאשר השונות ידועה ושווה ל-4.

נתון רווח סמך ברמת סמך 95% לתוחלת μ:

$$[L, U] \equiv \left[\overline{X}_n - 1.96\frac{\sigma}{\sqrt{n}}, \overline{X}_n + 1.96\frac{\sigma}{\sqrt{n}}\right] = \overline{X}_n \pm 1.96\frac{\sigma}{\sqrt{n}}$$

כאשר L הגבול התחתון (lower bound) ו-U הגבול העליון (upper bound) של רווח הסמך. נגדיר "טווח השגיאה" כחצי רוחב רווח הסמך: $d \equiv \frac{U-L}{2}$.

א. במדגם של 12 תצפיות, מהו טווח השגיאה d לרווח הסמך?

ב. מהו גודל המדגם המינימלי עבורו d לא גדול מ-1?

ג. הראו את התוצאה של סעיף ב' בצורה גרפית: ציירו גרף של d כפונקציה של n והראו שהתוצאה שקיבלתם בסעיף ב׳ היא אכן גודל המדגם המינימלי עבורו $d \leq 1$.

ד. במדגם של 12 תצפיות, כאשר סטיית התקן **לא** ידועה אבל סטיית התקן המדגמית s יצאה 2, מהו טווח השגיאה d לרווח סמך ברמת סמך 95%?

ה. האם אפשר לחזור על סעיף ב' כשסטיית תקן לא ידועה? אם כן, חזרו עליו. אם לא, הסבירו מדוע.

שאלה 3 (25 נקודות, 5 נקודות לכל סעיף)

הציון של מבחני IQ מנורמל ביחס לאוכלוסייה, כך שתוחלת הציונים באוכלוסייה הכללית תהיה 100 נקודות, וסטיית התקן תהיה 10 נקודות. ההנחה היא שההתפלגות באוכלוסייה הכללית היא נורמלית.

פסיכולוג א' ערך מבחן IQ ל-15 נבדקים שהגיעו אליו באופן בלתי תלוי מתוך <u>אוכלוסיית הלקוחות שלו</u>. להלן הציונים:

113, 105, 102, 104, 117, 123, 110, 108, 93, 96, 99, 107, 112, 82, 96

נגדיר μ_1 תוחלת הציונים של אוכלוסיית הלקוחות של פסיכולוג א' (הניחו כי סטיית התקן ידועה ושווה גם היא ל-10).

הפסיכולוג מבקש מכם עזרה בניתוח הנתונים.

א. סכמו את הנתונים. האם לדעתכם ההנחה שהתצפיות מתפלגות נורמלי סבירה במקרה זה? צרפו גרף מתאים.

ב. בעזרת ההנחות על התפלגות האוכלוסייה, חשבו רווח סמך של 80% לתוחלת μ_1.

ג. הסבירו לפסיכולוג את רווח הסמך ומשמעותו (התייחסו למשמעות של רמת ביטחון 80% ולסיכוי שהפרמטר נמצא בתוך רווח הסמך שחישבתם).

ד. האם לדעתכם אוכלוסיית הנבדקים חריגה ביחס לאוכלוסייה הכללית?

ה. חשבו רווח סמך לתוחלת μ_1 רמת סמך של 95%. האם תשובתכם לסעיף ד' תשתנה על בסיס רווח הסמך הזה?

ו. חזרו על סעיף ה' בלי להניח שסטיית התקן ידועה. האם תשובתכם לסעיף ד' תשתנה על בסיס רווח הסמך הזה?

ז. (**בונוס** 5 נקודות) מהי הרמת הסמך הגבוהה ביותר שבה תוכלו לטעון שאוכלוסיית הנבדקים חריגה ביחס לאוכלוסייה הכללית כאשר סטיית התקן אינה ידועה?

שאלה 4 (25 נקודות)

הקבצים Keshet12.csv ו-Reshet13.csv מכילים תוצאות מדגמים של סוקרים מטעם הערוצים קשת 12 ורשת 13 מיד עם סגירת הקלפיות בבחירות א2019. בערוץ 12 דגמו באקראי 60 קלפיות ברחבי הארץ, מתוכן ספרו באקראי שליש מהקולות. בערוץ 13 עשו אותו הדבר עם 70 קלפיות (אחרות).

בשני הקבצים כל שורה מייצגת קלפי, וכל עמודה מייצגת אחת מ-14 המפלגות הגדולות. הקולות שניתנו לשאר המפלגות שלא מופיעות בקובץ - ניתנים להזנחה. עבור כל אחד מהמדגמים בנפרד:

א. מצאו את פרופורציית הקולות שכל מפלגה קיבלה.

ב. מצאו את גבולות רווח סמך שמרני של 95% לפרופורציית הקולות של כל מפלגה.

ג. ציירו תרשים המתאר את פרופורציית הקולות שקיבלה כל מפלגה בשילוב הגבולות של רווח הסמך. הציגו את התוצאות לפי גודל המפלגה, כך שהמפלגה שקיבלה הכי הרבה קולות תופיע בצד ימין של התרשים.

❖ ניתן להוסיף קווים אנכיים שמתארים את גבולות רווחי הסמך U, L בצורה הבאה:

```
plt.vlines(x=party_names, ymin=L, ymax=U, colors='black', ls
='-', lw=2)
```

כאשר party_names הוא וקטור של מזהי המפלגות, L הוא וקטור של הקצוות השמאליים של רווחי הסמך, ו-U הוא וקטור של הקצוות הימנים של רווחי הסמך.

ד. בהתייחס לשתי המפלגות הגדולות, האם ניתן לקבוע על סמך אחד הסקרים שאחת מהן קיבלה יותר קולות מהשנייה?

ה. האם התשובה לסעיף ד' תשתנה במידה ונבקש רמת סמך של 99%?