מעבדה בסטטיסטיקה 52568 - 2020-21, מטלה 7. להגשה והצגה ב-13.12

תיאור המשימה:

המעבדה עוסקת בחיזוי תוצאות הבחירות על ידי דגימה.

יש להשתמש בקבצי תוצאות הבחירות על פי **קלפיות** בבחירות מרץ 2020 וספטמבר 2019.

1. בצעו מדגם אקראי של **קלפיות** באופן הבא:
   - בחרו 100 פעמים מדגם אקראי עם b=10 של קלפיות בבחירות 2020 וחשבו את שכיחות ההצבעה $\widehat{p_j}$ ל-8 המפלגות הגדולות עבור כל מדגם כזה.
   - לאחר מכן חשבו עבור בעזרת 100 האומדנים שקיבלתם לכל מפלגה אומדנים להטיה, לשונות ולשגיאה הריבועית הממוצעת של תוצאות קלפיות המדגם עבור מפלגה זו, ביחס ל- p_j, שכיחות ההצבעה למפלגה בכלל המדינה. האם ההטיה קרובה לאפס כמצופה על פי התאוריה?
   - הראו בגרף bar-plot עם error-bar את הממוצע פלוס/מינוס סטיית תקן של האומד עבור כל מפלגה.
   - השוו ע"י bar-plot את ה-MSE שקיבלתם עבור כל מפלגה לשגיאה הריבועית הממוצעת התאורטית $p_j\left(1-p_j\right)/m$ עבור המפלגה אם היינו עושים מדגם בו בוחרים **מצביעים** באופן אקראי, כאשר m מספר הבוחרים הממוצע במדגם. (מכיוון שאנו דוגמים קלפיות מספר הבוחרים הממוצע משתנה קצת עבור כל סימולציה גם עבור b קבוע. חשבו את המספר הממוצע - כלומר מספר הבוחרים הממוצע בקלפי כפול מספר הקלפיות).
2. כעת בצעו **מדגם שכבות** באופן הבא:
   - השתמשו בקובץ האשכולות החברתיים כלכליים כדי לקבוע עבור כל קלפי את האשכול של הישוב אליו היא שייכת.
   - כעת עבור b=10 בחרו קלפי אחת באופן אקראי מכל אשכול חברתי כלכלי. חשבו את שכיחות ההצבעה $\widehat{p_j^{(l)}}$ לכל מפלגה j במדגם של כל אשכול l. לאחר מכן חשבו את האומדן המשוקלל לשכיחות ההצבעה למפלגה במדינה על ידי ממוצע משוקלל של האשכולות, כאשר המשקלות N_l ניתנים ע"י מספר הקולות הכשרים הכללי בכל אשכול: $\widehat{p_j} = \sum_l N_l \widehat{p_j^{(l)}} / \sum_l N_l$
   - חזרו על מדגם זה 100 פעמים כמו בשאלה 1 והשוו את השגיאה הריבועית הממוצעת המתקבלת ממדגם השכבות לשגיאה עבור למדגם האקראי של 10 קלפיות משאלה 1 עבור כל מפלגה ע"י bar-plot. האם יש עדיפות לאחת השיטות? חשבו והשוו את סכום ה-MSE על פני כל המפלגות בשתי השיטות.
3. נשתמש כעת בתיקון שהוצג בכיתה עם פרמטר α ושקלול עם תוצאות הבחירות הקודמות, כלומר האומדן למפלגה j הוא : $\overline{p_j}(\alpha) = \alpha\widehat{p_j} + (1-\alpha)q_j$ כאשר q_j אחוז ההצבעה למפלגה j בבחירות ספטמבר 2019.

   א. עבור $\widehat{p_j}$ המתקבל מהמדגם מסעיף 1 בגודל b=10 קלפיות השתמשו בשונות שחישבתם מסעיף 1 וחשבו עבור כל ערך של α בין הערכים 0,0.01,...,0.99,1 את השונות, ה-$bias^2$ וה-MSE של האומדן המשוקלל $\overline{p_j}(\alpha)$ בעזרת הנוסחאות שהוצגו בכיתה.

   הציגו עבור כל אחת מ-8 המפלגות הגדולות את השונות, ה-$bias^2$ וה-MSE כפונקציה של α (2x4 גרפים, כאשר בכל גרף 3 עקומות המתארות את 3 הגדלים עבור המפלגה).

   נניח שהיינו בוחרים להשתמש ב- $\alpha = 0.5$ עבור כל 8 המפלגות הגדולות. האם השגיאה הריבועית הממוצעת היתה משתפרת ביחס לתוצאות המדגם בלבד? (כלומר $\alpha = 1$)

ב. נניח ש- $\widehat{p_j}$ אומד בלתי מוטה ל- p_j עם שונות σ^2. גיזרו את הביטוי ל-MSE וכתבו נוסחא לפרמטר α שימזער את ה- MSE של $\overline{p_j}(\alpha)$ כפונקציה של p_j, q_j ושל σ^2. בדקו שהנוסחא אכן נותנת לכם את ה- α הממזער את עקומות ה-MSE בסעיף הקודם עבור כל מפלגה.
האם ניתן להשתמש בנוסחא זו כדי לקבוע מראש באיזה אומדן כדאי להשתמש עבור המדגם? אם כן, כיצד? אם לא, מדוע?

הערות:

- חשבו על עיצוב הגרפים. תנו כותרת לצירים, שימו לב לאורך הצירים.
- השתמשו בצבעים, עובי נקודה, וכו׳ כדי להדגיש נקודות חשובות.
- מותר להיות יצירתיים. באיזה עוד דרכים אפשר לדגום או לאמוד את התוצאות מתוך תוצאות המדגם ומתוצאות עבר? איך משתנה השגיאה שלנו כאשר משנים את גודל המדגם?