מעבדה בסטטיסטיקה 52568 - 2020-21, מטלה 3. להגשה והצגה ב-8.11

תיאור המשימה:

במשימה זו נבצע סימולציה של ההצבעה בבחירות כאשר כל בוחרת מחליטה באופן מקרי אם להצביע למפלגה המועדפת עליה או לא להצביע כלל, ונשתמש בסימולציה לבדיקת טיב התיקון שלנו תחת הנחות שונות. בכל השאלות יש להשתמש רק באנליזה לפי קלפיות (כלומר קובץ הקלפיות), וכן נעבוד רק על נתוני 8 המפלגות הגדולות ביותר, וללא המעטפות החיצוניות.

1. כתבו פונקציה המבצעת סימולציה של ההצבעה בבחירות על פי המודל שנלמד בכיתה: הפונקציה מקבלת data-frame עבורו $\tilde{n}_{ij}$ מספר המצביעים הפוטנציאלים למפלגה i בקלפי j ו-v_{ij} ההסתברות שמצביע פוטנציאלי למפלגה j בקלפי i אכן יצביע, ומחשבת את מספר המצביעים בפועל המתפלג בינומית:
$$n_{ij} \sim Binom(\tilde{n}_{ij}, v_{ij})$$
ומספרי המצביעים בקלפיות שונות ובמפלגות שונות הם בלתי תלויים.
תוכלו להשתמש בפקודת random.binomial של numpy.

2. כעת השתמשו בפונקציה זו עבור סימולציה של מספרי המצביעים בקלפיות n_{ij}. השתמשו בערכים הבאים:
   - עבור $\tilde{n}_{ij}$ השתמשו בערכים האמיתיים של מספרי המצביעים עבור כל מפלגה בקובץ נתוני **הקלפיות**, מוכפלים במספר בעלי זכות הבחירה הכללי במדינה ומחולקים במספר הקולות הכשרים הכללי במדינה . כלומר אנחנו מניחים כאן שמספר הקולות הפוטנציאליים לכל מפלגה **בסימולציה** פרופורציוני למספר הקולות שאכן נצפו **בתוצאות האמת**, ומתקנים רק כדי להגדיל את מספר הקולות הפוטנציאלים בסימולציה שיתאים למספר בעלי זכות הבחירה הכללי בישראל. עגלו את ערכי $\tilde{n}_{ij}$ למספרים שלמים. (ההנחה כאן אינה בהכרח סבירה עבור נתוני האמת ונועדה רק לצורך הסימולציה)
   - עבור v_{ij} השתמשו בשלושה ערכים:
   א. $v_{ij}=v_i$ כאשר v_i הוא שיעור ההצבעה בפועל בקלפי i הנתון בקובץ (מספר הקולות הכשרים מחולק במספר בעלי זכות הבחירה בקלפי זו)
   ב. $v_{ij}=u_j$ כאשר u_j עבור 8 המפלגות הגדולות הם הערכים 0.2,0.3,...,0.9. (בחרו כרצונכם איזה ערך מתאים לאיזו מפלגה).
   ג. ערכי v_{ij} כמו בסעיף ב אבל בכל ישוב אנו בוחרים את ערכי u_j מחדש באופן אקראי - כלומר הערכים הם עדיין 0.2,0.3,...,0.9 כאשר בכל הקלפיות באותו ישוב הערך 0.2 למשל מתאים לאותה מפלגה j, אבל בישוב אחר ערך זה יכול להתאים למפלגה אחרת.

2.1 עבור כל אחת מהאפשרויות א,ב,ג בצעו 50 סימולציות. בכל סימולציה חשבו על הנתונים המסומלצים n_{ij} את התיקון שעשינו במעבדה 2.

2.2. עבור כל מפלגה j וכל אחת מ-3 הדרכים לסימולציה חשבו עבור כל אחת מ-50 הסימולציות אומדן ל-q_j (כלומר לשכיחות בעלי זכות הבחירה התומכים במפלגה מבין כלל בעלי זכות הבחירה).

2.3 חשבו בעזרת ערכי האומדנים את התוחלת, ההטיה, השונות, והשגיאה הריבועית של האומד ל- q_j.
עבור כל גודל הציגו את הנתונים בטבלה (8×3=24 ערכים) . מהי מסקנתכם לגבי התיקון בכל אפשרות?

2.4 עבור כל אחת מ-3 הדרכים לסימולציה ציירו bar-plot ובו שלוש עמודות לכל מפלגה המשווה את: השכיחות האמיתית באוכלוסיה לכל מפלגה q_j , השכיחות שנצפתה בנתוני הסימולציה של הבחירות p_j, והשכיחות על פי התיקון (כלומר האומדן ל-q_j). עבור 2 הגדלים האחרונים, ציירו בר המייצג את הממוצע וכן error-bars המייצגים את סטיית התקן (בעזרת האופציה yerr של הפונקציה bar).
(דוגמא לציור ברים עם סטיית תקן מופיעה כאן:
https://pythonforundergradengineers.com/python-matplotlib-error-bars.html)

הערות:

- חשבו על עיצוב הגרפים. תנו כותרת לצירים, שימו לב לאורך הצירים.
- השתמשו בצבעים, עובי נקודה, וכו׳ כדי להדגיש נקודות חשובות.
- מותר להיות יצירתיים; חשבו על ערכי $\tilde{n}_{ij}$ ו-v_{ij} אחרים בסימולציה ועל שיטות תיקון אחרות.