מעבדה לסטטיסטיקה 52568 - 2019-20, מטלה 3. להגשה והצגה ב-17.11

תיאור המשימה:

במשימה זו נציג שיטות שונות לתיקון אחוז ההצבעה למפלגות שונות בבחירות לכנסת ה-22 (ספטמבר 2019) וכן נבצע סימולציה לבדיקת טיב התיקון שלנו תחת הנחות שונות. בכל השאלות יש להשתמש רק באנליזה לפי קלפיות (כלומר קובץ הקלפיות), וכן נעבוד רק על נתוני 10 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-p_{ij} ההסתברות שמצביע פוטנציאלי למפלגה j בקלפי i אכן יצביע, ומחשבת את מספר המצביעים בפועל המתפלג בינומית:

$$n_{ij} \sim Binom(\tilde{n}_{ij}, p_{ij})$$

ומספרי המצביעים בקלפיות שונות ובמפלגות שונות הם בלתי תלויים.

תוכלו להשתמש בפקודת random.binomial של numpy.

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 פרופורציוני למספר הקולות שאכן נצפו, ומתקנים רק כדי להגדיל את מספר הקולות הפוטנציאלים 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 p_{ij} השתמשו בשני ערכים:

א. $p_{ij}=v_i$ כאשר v_i הוא שיעור ההצבעה בפועל בקלפי i הנתון בקובץ (מספר הקולות הכשרים מחולק במספר בעלי זכות הבחירה בקלפי זו)

ב. $p_{ij}=\alpha_j$ כאשר α_j עבור 10 המפלגות הגדולות הם הערכים 0.50,0.55,...,0.95. (בחרו כרצונכם איזה ערך מתאים לאיזו מפלגה).

עבור כל אחת מהאפשרויות א,ב בצעו 25 סימולציות. בכל סימולציה חשבו על הנתונים המסומלצים n_{ij} את התיקון שעשינו במעבדה 2. ציירו bar-plot ובו שלוש עמודות לכל מפלגה המשווה את: השכיחות האמיתית לכל מפלגה, השכיחות שנצפתה, והשכיחות על פי התיקון. עבור 2 הגדלים האחרונים, ציירו בר המייצג את הממוצע על פני 25 הסימולציות וכן error-bars המייצגים את סטיית התקן על פני 25 הסימולציות (בעזרת האופציה yerr של הפונקציה bar). מהי מסקנתכם לגבי התיקון בכל אפשרות? (דוגמא לציור ברים עם סטיית תקן מופיעה כאן:
https://pythonforundergradengineers.com/python-matplotlib-error-bars.html)

2. כתבו פונקציה שעושה תיקון על ידי פתרון בעיית הרגרסיה הלינארית הבאה על פי שיטת הריבועים הפחותים. תוכלו להשתמש בפונקציה OLS מתוך מודול הפייתון statsmodel (זו פונקציה מקבילה מהרבה בחינות לפונקציה lm ב-R).
   - ראשית, אומדים את ערכי α_j^{-1} ע"י מיזעור הביטוי:

$$\Sigma_i(\Sigma_j n_{ij}\alpha_j^{-1}-\tilde{n}_{i\bullet})^2$$

   - שנית, מחשבים את האומדים ל- $\tilde{n}_{ij}$ ע"י הכפלת n_{ij} באומדים ל-α_j^{-1} ומכאן מחשבים את האומדים לשכיחויות ההצבעה המתוקנות.

כעת חזרו על הסימולציות בא,ב מהשאלה הקודמת, והציגו באותו אופן את תוצאות התיקון החדש עבור סימולציות אלו. מה מסקנותיכם? באיזה תיקון כדאי להשתמש ומתי?

3. השתמשו בתיקון המבוסס על רגרסיה לינארית משאלה 2 עבור קובץ התוצאות של הבחירות על פי קלפיות (כלומר על הנתונים האמיתיים, לא נתוני סימולציה) והשוו לתיקון אותו חישבנו עבור אותו קובץ במעבדה 2 : הציגו bar-plot המראה לכל מפלגה את השכיחות בפועל והשכיחות על פי כל אחד משני התיקונים. מהי מסקנתכם? איזה תיקון נראה יותר סביר עבור נתוני האמת ולמה?

הערות:

- חשבו על עיצוב הגרפים. תנו כותרת לצירים, שימו לב לאורך הצירים.
- השתמשו בצבעים, עובי נקודה, וכו׳ כדי להדגיש נקודות חשובות.
- השתמשו בשמות המפלגה הבאים בכל פעם שתדרשו לכך במקום באותיות: (ניתן להמיר באמצעות בניית dictionary בפייתון)

פתק	פה	מחל	ודעם	שס	ל	ג	טב	אמת	מרץ	כף
מפלגה	כחול לבן	הליכוד	הרשימה המשותפת	שס	ישראל ביתנו	יהדות התורה	ימינה	העבודה גשר	המחנה הדמוקרטי	עוצמה יהודית

- מותר לכם להיות יצירתיים; נסו לחשוב על שיטות אחרות לתיקון אחוזי ההצבעה.