מעבדה לסטטיסטיקה 52568 - 2020-21, מטלה 4. להגשה והצגה ב-15.11

תיאור המשימה:

המעבדה עוסקת בתיקונים לאחוז הצבעה וכן בקשר בין דירוג חברתי כלכלי של יישובים לבין תוצאות הבחירות.
בנוסף לקובץ תוצאות הבחירות יש להוריד ולהשתמש במשימה זו בקובץ דירוג חברתי כלכלי לפי ישובים:
HevratiCalcaliYeshuvim.csv או HevratiCalcaliYeshuvim.txt

1. כתבו פונקציה שעושה תיקון לתוצאות הבחירות ואמידת השכיחויות q_j באוכלוסיה של כל מפלגה 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 u_j^{-1} ע"י מיזעור הביטוי: (כאן i סוכם על הישובים ו-j על הקלפיות) $\Sigma_i(\Sigma_j n_{ij}u_j^{-1}-\tilde{n}_{i\cdot})^2$
   - שנית, מחשבים את האומדים ל- $\tilde{n}_{ij}$ ע"י הכפלת n_{ij} באומדים ל-u_j^{-1} ומכאן מחשבים את האומדים $\widehat{q_j}$ לשכיחויות ההצבעה המתוקנות כפי שהודגם בכיתה.
2. כעת חזרו על הסימולציות בשאלה 2 ממעבדה 3. חשבו את האומדנים $\widehat{q_j}$ לכל אחת מ-3 הדרכים לבחירת v_{ij} על פי התיקון בשאלה 1 (כלומר בדומה לשאלה 2.2 במעבדה הקודמת אך בעזרת התיקון החדש). עבור כל אחת מ-3 דרכי הסימולציה ועבור כל מפלגה הוסיפו ל-bar-plot משאלה 2.4 במעבדה הקודמת עמודה חדשה עבור התיקון החדש (ממוצע +/- סטיית תקן) והציגו את ה-bar-plots החדשים (כלומר לכל מפלגה יהיו 4 עמודות). מה מסקנותיכם? באיזה תיקון כדאי להשתמש ומתי?
3. קראו את הנתונים מקובץ הדירוג החברתי הכלכלי והצליבו אותם עם נתוני הבחירות על פי **ישובים** ליצירת data-frame משותף (ניתן להשתמש בפקודת merge של pandas). הציגו את רשימת הישובים שמופיעים בשני הקבצים. כמה ישובים קיבלתם? בשאלה זו ובבאה השתמשו רק בישובים אלו לניתוח הנתונים. חשבו את תוצאות הבחירות (שכיחות הקולות p_j לכל מפלגה מתוך כלל הקולות הכשרים עבור 8 המפלגות הגדולות) רק בישובים אלו והציגו אותן מול התוצאות הכלל ארציות בבר עמודות כפול. האם יש הבדלים משמעותיים בין התוצאות הארציות לתוצאות בישובים שקיבלתם?
   (**הערה:** אין חובה להוסיף ידנית ישובים המופיעים באיות שונה וכן ישובים המופיעים במועצות אזוריות בקובץ הדירוג החברתי הכלכלי.)
4. א. הציגו את תוצאות הבחירות שהיו מתקבלות בכל אחד מעשרת הדירוגים החברתיים כלכליים מ-1 עד 10. כלומר, יש להראות גרף עם 10 subplots מסודרים בתבנית של 2X5 כאשר בכל subplot יש גרף עמודות המתאר את שכיחות הקולות של 8 המפלגות הגדולות ביותר רק בישובים עם הדירוג חברתי כלכלי המתאים.
   ב. עבור כל מפלגה, צרו גרף עמודות בו מתואר שכיחות הקולות של המפלגה בישובים ב-10 האשכולות, מסודרים. כלומר, יש לעשות גרף עם 8 subplots מסודרים בתבנית של 2X8 כאשר בכל subplot יש את שכיחות הקולות של מפלגה מסויימת ב-10 האשכולות מסודרים לפי הסדר
   (**הערה:** בסעיף זה אנו מציגים למעשה את אותה אינפורמציה אשר בסעיף א. אבל בדרך אחרת).

הערות:

- חשבו על עיצוב הגרפים. תנו כותרת לצירים, שימו לב לאורך הצירים.
- השתמשו בצבעים, עובי נקודה, וכו׳ כדי להדגיש נקודות חשובות.
- מותר לכם להיות יצירתיים; נסו לכלול יותר ישובים בניתוח המשולב של הבחירות והנתונים הסוציואקונומיים. בשאלה 4 אפשר גם לחשב ולהציג על ה-bars עבור כל מפלגה את הממוצע +/- סטיית התקן של שכיחות ההצבעה למפלגה על פני ישובים באותו אשכול.