

מעבדה בסטיסטיקה 52568 - 2020-21, מטלה 1. להגשה והצגה ב-25.10

במטלה זו נקרא את קובץ ההצבעה בבחירות לכנסת ה-23 (מרץ 2020) ונציג את דפוסי ההצבעה בישובים

הדרכה: את כל חלקי המשימה ניתן לבצע על ידי שינויים קלים של מחברת הפיתון ReadElectionData.ipynb שהודגמה בכיתה. תחילה, הורידו את קובץ זה ל-colab משלכם, וכן טענו את קובץ הנתונים: 'votes per city 2020.csv'.

אם תרצו לעבוד במחשב שלכם ולא בסביבת colab, עליכם לשים את הקבצים בתיקיית עבודה לבחירתכם ולהתקין את המודולים os, matplotlib, pandas, numpy כדי לעבוד עם הקובץ. תוכלו לקרוא את הקובץ לתוך data-frame באמצעות פקודת read_csv של pandas כפי שהראינו בכיתה ומתואר במחברת הפיתון.

1. מצאו את העמודות המציינות את מספר הקולות הפסולים ואת מספר כלל המצביעים בכל ישוב וציירו היסטוגרמה של **אחוז** הקולות הפסולים בישוב על פני כל הישובים. השתמשו ב-100 תאים שווים מרחק בין 0 לבין אחוז הקולות הפסולים המקסימלי.
2. בנוסף, מצאו את שלושת הישובים עם **אחוז** הקולות הפסולים המקסימלי וציינו את שמם ואת האחוז. כתבו פונקציה המציגה עבור שני ישובים bar-plot של ההצבעה ל-8 המפלגות שעברו את אחוז החסימה (מעל 3.25% הצבעה כלל ארצית) שניהם אחד ליד השני בצבעים שונים כפי שהודגם בכיתה. בחרו 3 זוגות של ערים גדולות (מעל 50,000 מצביעים) והציגו את ההצבעה ב-3 זוגות הערים בעזרת הפונקציה שכתבתם. תארו את ההבדלים והדמיון בין הערים בהצבעה למפלגות השונות.
3. עבור שתי התפלגויות קטגוריות על n ערכים עם ההסתברויות:

$$p=[p_1,\dots,p_n], q=[q_1,\dots,q_n]$$

נגדיר את המרחק הריבועי להיות:

$$d(p,q)=(p_1-q_1)^2+\dots+(p_n-q_n)^2$$

(זוהי דוגמה לפונקציית מרחק בין התפלגויות).

- א. מצאו את הישוב **המייצג ביותר** את אחוז ההצבעה בכלל ישראל - כלומר הישוב שהתפלגות ההצבעה בו (על פני כל המפלגות) היא **הקרובה** ביותר להתפלגות ההצבעה הארצית עבור מדד זה (ממזער את המרחק הריבועי מהתפלגות ההצבעה הארצית) והציגו bar-plot של הישוב ביחד עם התפלגות ההצבעה הארצית. הסבירו את תשובתכם.
- ב. חזרו על א' עבור הישוב שהתפלגות ההצבעה בו היא **הרחוקה** ביותר מהתפלגות ההצבעה הארצית.

הערות:

- אם אינכם משתמשים במחברת python יהיה עליכם לשמור גרפים לקובץ כדי להכין את הסיכום. כדי לשמור גרף לקובץ השתמשו בפקודה savefig במודול matplotlib.
- חשבו על עיצוב הגרפים. תנו כותרת לצירים וכותרת כללית לגרף. תנו כותרות לעמודות במידת הצורך. הוסיפו legend המתאר כל קו במידת הצורך. שימו לב לגבולות הצירים. השתמשו בצבעים וכו' כדי להדגיש הבדלים.
- מותר להיות יצירתיים; נסו לחפש ישובים הדומים זה לזה בהצבעה על פני כל המפלגות, לאפיין ישובים עם אחוז קולות פסולים גבוה וכו'