**שקי חול חקר נתונים**

לפני הגעתכם לקורס תתבקשו להתנסות במחקר ועבודה עם נתונים. זהו קובץ הסבר והכוונה לביצוע שאלון **google forms** שתענו עליו מיד לאחר מכן. את משימה זו יש להגיש לפני תחילת הקורס. המשימה הינה בגדר חובה.

## משתנים

* משתנה - כל **מאפיין** או **תכונה** שניתן למדוד או לספור. הוא נקרא משתנה בגלל שהערך הנמדד שלו עשוי להשתנות. דוגמה למשתנה: גובה, ציון, מחיר, שם, עיר מגורים
* תצפית - **מדידה** של משתנה מסוים. כלומר, הערך של המשתנה הנחקר שמעניין אותי למדוד. דוגמה לתצפית: המשתנה שלי הוא גובה, ואחת מהתצפית בו היא 1.67.

תצפית

משתנה

|  |  |  |
| --- | --- | --- |
| **שם** | **עיר מגורים** | **גובה** |
| נעם | הרצליה | 1.58 |
| רועי | רעננה | 1.63 |
| דניאל | אור יהודה | 1.73 |
| יאיר | חיפה | 1.81 |

נוכל לחלק את המשתנים לשני סוגים:

**משתנה נומרי/כמותי**

**משתנה קטגוריאלי/איכותי**

* משתנה קטגוריאלי (categorical variable) –

נקרא גם משתנה איכותי, והוא משתנה שערכיו (או בעצם תצפיותיו) הם **טקסט**. מידע שהוא קטגורי יכיל מספר אופציות מוגבל.דוגמה למשתנה קטגוריאלי: מין, מצב משפחתי, דרגה בצבא וכו'.

* משתנה נומרי/כמותי (numeric variable) –

משתנה כמותי הוא בעצם משתנה **מספרי**. משתנה זה אינו בהכרח יכיל מספר אופציות מוגבל. דוגמה למשתנה נומרי: משקל, משכורת, לחץ דם וכו'.

## מדדים

בעולם הסטטיסטי ניתן לערוך חישובים רבים שייצגו לנו "התנהגות של משתנה".  
ניתן לייצג את אותה התנהגות על ידי מדדים שונים הנקראים **מדדי מרכז ומדדי פיזור.**

**מדדי מרכז**

* ממוצע - הממוצע הוא סכום הערכים חלקי כמות הערכים (סכום התצפיות חלקי כמות התצפיות). הממוצע מייצג את הרמה הכללית של המשתנים.

כלומר בסוף החישוב אנחנו מקבלים מספר שמייצג **בערך** את המספר שערכו סובבים המספרים.

עבור סט המספרים: 2,4,5,5,6,10 הערך הממוצע הינו 5.33

* חציון - הערך האמצעי בתוך סט ערכים. כלומר מחצית התצפיות קטנה ממנו/שווים לו והמחצית השנייה גדולה ממנו/שווים לו. כשמחשבים חציון הערכים מסודרים מהקטן לגדול.

עבור סט המספרים: 1,2,3 הערך החציוני הינו 2.

עבור מקרה בו סט הערכים בעל כמות ערכים זוגית, כמו סט הערכים הבא: 4,2,1,3

הערך החציוני יחושב באופן הבא:

ניקח את שני המספרים האמצעיים לאחר שמיינו אותם מהקטן לגדול, והמספרים הם 2,3. נחשב עבורם ממוצע. לסיום נקבל את הערך החציוני של הסט שהוא 2.5.

* שכיח - התצפית שמופיעה הכי הרבה פעמים במידע. כלומר הערך הנפוץ ביותר בתוך סט ערכים.

עבור סט המספרים: 1,1,1,2,2,2,2,3 השכיח הינו המספר 2

**מדדי פיזור**

* סטיית תקן - מתאר את פיזור הערכים סביב הממוצע. תוצאת סטיית התקן תיתן לנו ערך שמייצג את המרחק הצפוי של תצפית אקראית מהממוצע.

32

-10

19

71

48

35

ככול שתוצאת סטיית התקן תהיה קטנה יותר, אנחנו מבינים שההפרשים בין התצפיות לממוצע קטנים יותר, כלומר ערכי התצפיות קרובות לערכו של הממוצע, ובעצם מרוכזות יותר.

ככול שסטיית התקן גדולה יותר, ההפרשים בין התצפיות גדולים יותר, כלומר ערכי התצפיות רחוקות מערכו של הממוצע, ובעצם מפוזרות יותר.