סטטיסטיקה

*משתנה

מה זה משתנה? משתנה הוא משהו שמשתנה בין אדם לאדם או בין נבדק לנבדק. בסטטיסטיקה אנחנו בודקים לרוב אנשים ואנחנו בודקים מה השוני בין אדם לאדם, ומתוך השוני הזה אנחנו מנסים להסיק כל מיני מסקנות שנלמד עליהם בהמשך.
דוגמאות:
- אם אנחנו בודקים צבע עיניים של אנשים - המשתנה יהיה צבע עיניים - ערך שמשתנה בין אדם לאדם.
- אם נרצה לבדוק את גובה המשכורת של קבוצת אנשים, המשתנה יהיה "שכר".
- אם נרצה לבדוק גובה של קבוצת אנשים, המשתנה יהיה "גובה", כי הור משתנה (כשמו כן הוא) בין אדם לאדם.
- אם נרצה לדעת היכן לומדים אנשים שונים, המשתנה יהיה "מוסד הלימוד".
- משתנה יכול להיות גם מידת אופטימיות, מידת שביעות רצון.

אז למעשה, משתנה יהיה הדבר שאותו בדקנו, חקרנו.

*מיון המשתנים

ניתן למיין את המשתנים בשני אופנים שונים:

תיאורטי (נומינלי)	תצפיתני (אופרציונלי)
משתנה תיאורטי הוא התחום שהחוקר רוצה לבדוק.	הדרך בה אנו משתמשים כדי להשיב על השאלה או התחום אותו אנחנו חוקרים.

לדוגמה:
תיאורטי: החוקר רוצה לבדוק אהבה לבעלי חיים.
תצפיתני: כאן השיטה הכי קלה היא לשאול מלא מלא שאלות, כמו: האם יש לך בע"ח בבית? (כן/לא) התעללת פעם בע"ח? (כן/לא), כמה אתה אוהב בע"ח (מ-1 עד 5).

דוגמה נוספת:
תיאורטי: למה תלמידים מתבגרים שונאים אסיפות הורים? כאן צריך גם להגדיר למשל, מהי שינאה.
תצפיתני: יש כאן המון שאלות שצריך לשאול באמצעות מחקר, כדי שניתן יהיה להשיב על השאלה.

דוגמה נוספת:
תיאורטי: סקרנות. זו הגדרה עמומה - מונח מילוני. במחקר שלנו נמקד את המונח סקרנות למשהו תיאורטי שאנחנו רוצים לבדוק. למשל, **"הצורך שלנו לרדוף אחרי ידע ולהשיגו" -** זה **משתנה תיאורטי**.
תצפיתני: איך במחקר ספציפי הולכים לבדוק את הסקרנות. במקרה שלנו נוכל לשאול למשל **"כמה ספרים קרא הנבדק בשנה?"** - זה **משתנה תצפיתי**.

במבחן ניתקל בשאלה כמו:

א. מהו המשתנה הבלתי תלוי התיאורטי?
מהו המשתנה הבלתי תלוי התצפיתי ומהם ערכיו?
ב. מהו המשתנה התלוי התיאורטי?
מהו המשתנה התלוי התצפיתי ומהם ערכיו?

בואו ניקח שאלה לדוגמה כפי שהיא עשויה להיות במבחן...

חוקר א' טען שככל שאנחנו מתאמצים יותר על מנת להשיג משהו, כך אנחנו נהנים ממנו יותר.
כדי לבדוק את טענתו, הוא פנה לבי תספר יסודי ודגם מקרית 60 ילדים בין הגילים 10-13.
בכל יום זומנו למעבדה 6 ילדים באופן אינדיווידואלי. נעשתה הקצאה של תנאי הניסוי. בהגיעו למעבדה, הסביר החוקר לכל ילד שהוא נבחר להשתתף במשלק "חפש את המטמון". מהרגע שהחוקר עוזב את החדר, יש לו 15 דקות למציאת חפיסת שוקולד במוסתרת במקום כלשהו בחדר.
לאחר שימצא את החפיסה, עליו לצאת מהחדר ולדווח זאת מיד לחוקר אחר שימתין בחוץ.
לנחקרי קבוצה א' חפיסת השוקולד היתה מונחת על הרצפה.
לנחקרי קבוצה ב' חפיסת השוקולד היתה מוחבאת על מדף ספרים, מאחורי הספרים.
לנחקרי קבוצה ג' חפיסת השוקולד היתה מוחבאת בתוך פתח ברצפה מתחת לשטיח.
כל הילדים הצליחו למצוא את חפיסת השוקולד בתוך 15 הדקות שהוקצו להם.
לאחר שכל ילד יצא מהמעבדה, עוזר המחקר, שלא ידע לאיזו קבוצה שייך כל ילד, ביקש ממנו לפתוח את החפיסה, לטעום מהשוקולד ולתת ציון בין 1 ל-5, כאשר 1 מציין "לא טעים בכלל" ו-5 מציין "טעים מאוד".
תוצאות המחקר העלו כי ילדי קבוצה ג' נתנו בממוצע את הדירוג הגבוה ביותר, ואילו ילדי קבוצה א' נתנו בממוצע את הדירוג הנמוך ביותר.

רמז: המשתנים התיאורטים תמיד יופיעו במשפט הראשון. זוהי הטענה שהחוקר מעלה.

בואו נענה על השאלות:

1. ראשית, נחפש מהי הטענה או השאלה של החוקר. ממנה נוכל להבין מהם המשתנים התיאורטים.
בדוגמה שלנו: **חוקר א' טען שככל שאנחנו מתאמצים יותר על מנת להשיג משהו, כך אנחנו נהנים ממנו יותר.** לכן **המשתנה התלוי התיאורטי** יהיה **מידת ההנאה** ו**המשתנה הבלתי תלוי התיאורטי** יהיה **מידת המאמץ**.
2. כעת נחפש את המשתנים התצפיתים. בדוגמה שלנו החוקר ביצע מחקר ובו בדק **כיצד מיקום חפיסת השוקולד משפיע על רמת הטעימות שלו**. ולכן, **המשתנה התלוי התצפיתי** יהיה **מידת טעימות השוקולד** ו**המשתנה הבלתי תלוי התצפיתי** יהיה **מיקום החפיסה**.
3. כעת נמצא את הערכים אותם קבע החוקר. בדוגמה שלנו בחר החוקר ל**משתנה התלוי - קביעת רמת הטעימות של השוקולד** ערכים של **1 "לא טעים בכלל" עד 5 "טעים מאוד"**. ולערכים של **המשתנה הבלתי תלוי - מיקום השוקולד**: **"על הרצפה", "מאחורי הספרים" ו"מתחת לשטיח"**.

והתשובה שלנו תראה כך:

סוג המשתנה	משתנה תיאורטי	משתנה תצפיתי	ערכים
תלוי	מידת ההנאה	רמת טעימות	1 לא טעים בכלל - 5 טעים מאוד
בלתי תלוי	מידת המאמץ	מיקום החפיסה	על הרצפה, מאחורי הספרים, מתחת לשטיח

*משתנה כמותי ומשתנה איכותי

משתנה איכותי

משתנה כמותי

רציף

בדיד

המשתנה שבדקנו יכול להתחלק לשני סוגים שונים. מה ההבדל ביניהם?

משתנה איכותי הוא משתנה שלא נוכל למדוד אותו מבחינה מספרית.
דוגמאות למשתנים איכותיים:
- צבע חולצה (שחור, כחול, אדום, ירוק)
- מקצוע (נגר, מורה, עו"ד, רופא)

משתנה איכותי לא ניתן למדידה במספרים (צבע, מקצוע).

משתנה כמותי הוא משתנה שניתן למדוד אותו במספרים.
דוגמאות:
- מספר ילדים במשפחה
- כמות המטבעות שיש למישהו בארנק
- כמות הפעמים שבדקתי אימייל ביום

משתנה כמותי הוא משתנה שניתן למדוד במספרים (גובה, כמות מטבעות).

משתנה כמותי מתחלק לשני סוגים: משתנה רציף ומשתנה בדיד
מה ההבדל ביניהם?

משתנה בדיד הוא משתנה שלא נוכל למדוד אותו בחצאים או רבעים למשל או בנקודה עשרונית.
דוגמאות:
- מספר הלחיצות על העכבר לא יכול להמדד ב-3.5 או שלחצנו 3 פעמים על העכבר או שלחצנו עליו 4 פעמים.
- כמות הפעמים שניגשנו לבחינה. או שנגשנו פעמיים לבחינה או שנגשנו 3 פעמים לבחינה. אבל לא נוכל לגשת לבחינה 2.9 פעמים.
- כמות ילדים במשפחה. אין לנו חצי ילד במשפחה. או שיש לנו ילד אחד או שאין לנו ילדים.
- כן, גם אפס ניתן למדידה. כשאין ילדים במשפחה, או כשלא ניגשנו לבחינה, או כשלא לחצנו על העכבר, המדידה תהיה - אפס (אפס ילדים, אפס לחיצות על העכבר).

משתנה כמותי בדיד ניתן למדידה רק במספרים שלמים (0, 4, 22, וכו').

משתנה רציף הוא משתנה שנמצא על <u>רצף</u> מסוים וניתן למדוד אותו בנקודות עשרוניות.
דוגמאות:
- גובה של אדם הוא 184.6 ס"מ.
- משכורת של אדם היא 10,567.20 ש"ח (עשרת אלפים, חמש מאות שבעים שקלים ועשרים אגורות).

משתנה כמותי רציף ניתן למדידה על הרצף שבין המספרים השלמים (10.5, חצי, רבע, 345.80)

*סולמות מדידה

ישנם ארבעה סוגים של סולמות מדידה

- **סולם שמי** (נומינלי קטגוריאלי) (nominal scale)
- **סולם סדר** (אורדינלי) (ordinal scale)

שני הסולמות האלו תמיד מתקשרים למשתנה **איכותי**

- **סולם מנה** (יחס) (rational sacle)
- **סולם רווח** (אינטרוולי) (interval scale)

שני הסולמות האלו תמיד מתקשרים למשתנה **כמותי**

ועכשיו נסביר מה תפקידו של כולם סולם מדידה...

סולם שמי
משמש להבדיל בין ערכים שונים. כמו: מקצוע, עיר, צבע עיניים. המאפיין של הסולם השמי הוא שלא ניתן לסדר את האלמנטים לפי סדר מסוים, אלא רק להבדיל ביניהם. אי אפשר לומר שמהנדס יותר טוב מרופא, או שכחול יותר יפה מלבן.
דוגמאות:
- מקצוע: נגר, ספר זמר, רופא.
- עיר: תל-אביב, חולון, ערד, חיפה.
- צבע עיניים: ירוק, חום, כחול, חום דבש.

בסולם שמי ניתן רק להבדיל בין אלמנטים.
בסולם סדר ניתן לדרג אלמנטים.

סולם סדר
משמש לדרג אלמנטים באמצעות מלל.
דוגמאות:
- גובה: נמוך, בינוני, גבוה
- דרגה צבאית: רב"ט, סמל, רס"ן, אלוף.
- גיל: תינוק, צעיר, מבוגר, זקן.

סולם מנה
ניתן לא רק לדרג את האלמנטים, אלא גם למדוד אותם. הסולם הזה מדבר כבר במספרים. למשל, מי יותר גבוה? מי יותר נמוך ו-פי כמה?
דוגמאות:
- גיל: 20, 40, 70, 90 (אדם בן 40 גדול פי 2 מאדם בן 20).
- שכר: 1,000 ש"ח, 2,000 ש"ח, 40,000 ש"ח, 52,500 ש"ח (שכר של 40,000 ש"ח גדול פי 20 משכר של 2,000 ש"ח).
- מס' חדרים בבית: 1, 2, 3, 4 (בבית עם 4 חדרים יש פי שניים יותר חדרים מבית עם שני חדרים).

סולם רווח (לא ממש שימושי)
מתקשר למדידת טמפרטורה ואיקיו. הוא נקרא סולם רווח מכיוון שניתן לומר על האלמנטים בכמה משהו גדול ממשהו, אבל לא ניתן להגיד פי כמה. הגדרת האפס משמעה - העדר, חוסר, אין.
זאת אומרת, אפשר לומר שמחר יהיו 20 מעלות ומחרתיים יהיו 23 מעלות, אבל לא ניתן לומר שמחרתיים יהיה פי כבה וכבה יותר חם.
הסיבה היא שאין נקודת אפס מוחלטת, אבל זה לא ממש חשוב.
טמפרטורה ואיקיו נחשבים סולם רווח.

חשוב לזכור (סטודנטים רבים מתבלבלים כאן):

"שביעות רצון" הוא משתנה מסוג **סדר**.
מכיון ששביעות רצון ניתן למדוד כך:

מועטה	סבירה	בינונית	רבה	רבה מאוד

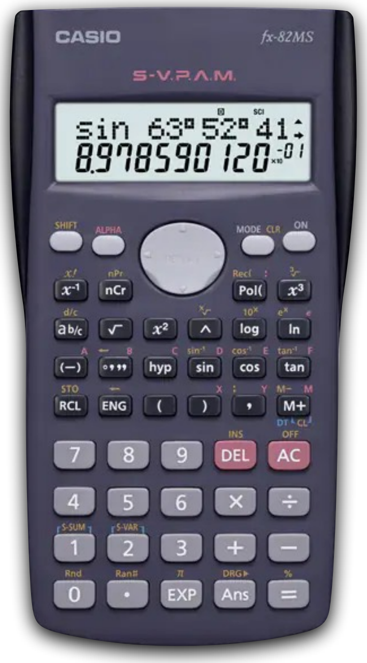
אך חשוב לשים לב ששביעות רצון למשל ניתן למדוד גם במספרים. אבל זה לא יהפוך אותו לסולם מדידה מסוג **מנה**, מכיון שהמספרים רק מחליפים מילים.

מועטה	1	2	3	4	5	רבה

תרגול: מהו סולם המדידה המתאים למשתנים הבאים?

א. דרגות בצבא
ב. אחוז האלכוהול במשקה
ג. גובה של אדם
ד. צבע עיניים
ה. טמפרטורה במעלות צלזיוס
ו. מספרי טלפון
ז. מרחק אווירי בין ירושלים לערים אחרות
ח. שביעות הרצון משירות האינטרנט

אגב, זה המחשבון המומלץ לבחינה



*משתנה תלוי ומשתנה בלתי תלוי

מעתה נגדיר משתנה באות X
את השכיחות (הערך הכי נפוץ) נגדיר באות f

שכיח הוא הערך הכי נפוץ.

דוגמה: צבע עיניים של אנשים
x - המשתנה: צבע עיניים
f - השכיח: כמה אנשים עם צבע עיניים דומה

הצבע השכיח הוא חום,
כי חום יש להכי הרבה
אנשים (5).

השכיח = צבע עיניים חום

x	f
שחור	3
חום	5
ירוק	2

דוגמה נוספת: ציון בבחינה
x - המשתנה: ציון בבחינה
f - השכיח: כמה אנשים קיבלו את אותו ציון?

הציונים הם: **80 , 70 , 60 , 70 , 100**

הציון "70" מופיע
פעמיים. שאר הציונים
מופיעים פעם אחת.

השכיח = ציון 70.

f	x
1	60
2	70
1	80
1	100

הכי הרבה
אנשים קיבלו
ציון 70

דוגמה נוספת: כמות המיילים שקבוצת אנשים מקבלים.
x - המשתנה: כמות מיילים
f - השכיח: כמות המיילים הכי נפוצה.

הציונים הם: **80 , 70 , 60 , 70 , 100**

f	x
3	2 מיילים
5	4 מיילים
1	5 מיילים
6	10 מיילים

הכי הרבה אנשים (6)
קיבלו 10 מיילים.

השכיח הוא 10.

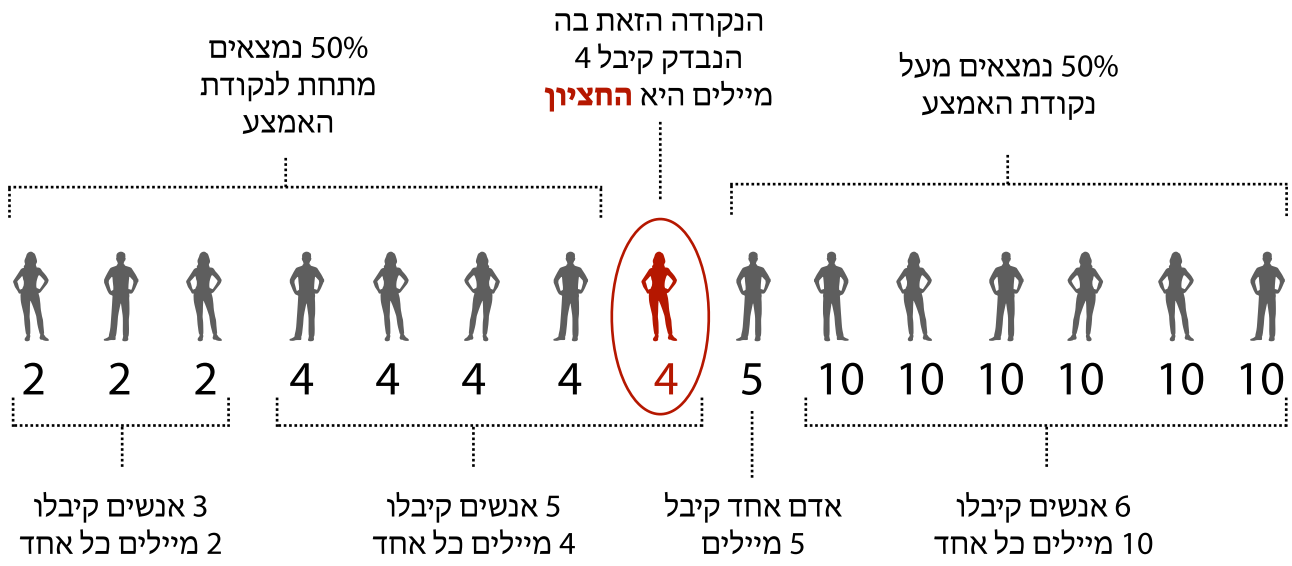
הכי הרבה אנשים

*חציון (Md)

חציון הוא ערך ש 50% מהנתונים נמצאים מעליו ו-50% נמצאים מתחתיו.
כדי להבין מהו חציון, ניקח את הדוגמה האחרונה ונפרוס אותה לרוחב.
אז, אם שלושה אנשים קיבלו 2 מיילים כל אחד, וחמישה אנשים קיבלו 4 מיילים כל אחד, ואדם אחד קיבל 5 מיילים, ושישה אנשים קיבלו 10 מיילים כל אחד, הפריסה תראה כך:

2 2 2 4 4 4 4 4 5 10 10 10 10 10 10

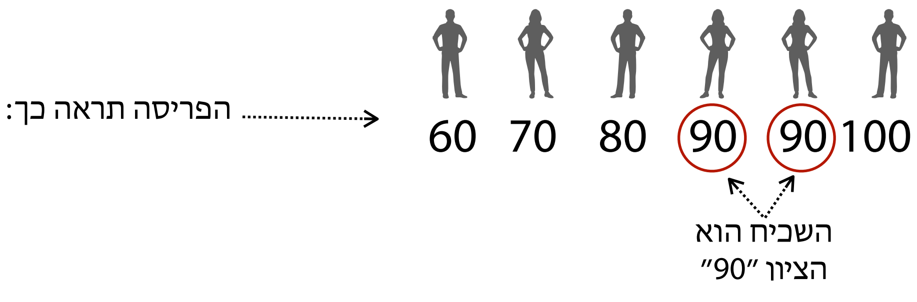
לפנינו בסך הכל 15 נבדקים. המספר מייצג את כמות המיילים שכל נבדק קיבל.
אם רוצים להבין את הפריסה טוב יותר, אז זה נראה כך:



החציון, כשמו כן הוא - <u>חוצה</u> את הנתונים לשניים. כמות האנשים שמעליו זהה לכמות האנשים שמתחתיו.

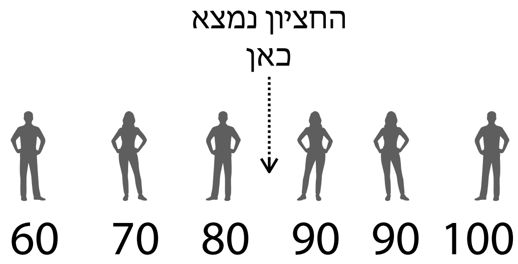
אבל רגע... בדוגמה האחרונה היה מספר אי זוגי של נבדקים. מה קורה כשיש מספר זוגי של נבדקים? איך אפשר למצוא את האמצע - את **החציון**?

דוגמה: 6 נבדקים קיבלו ציונים בבחינה. הציונים התחלקו כך: 60 , 70 , 80 , 90 , 90 , 100.



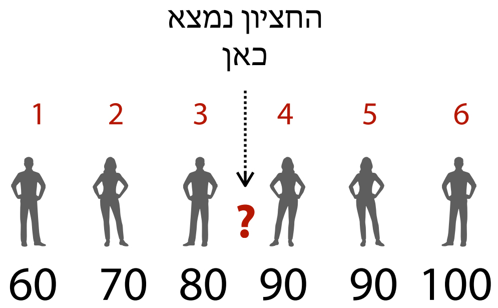
למה? כי השכיח הוא הנתון שחוזר הכי הרבה פעמים - והציון 90 מופיע פעמיים. כל השאר מופיעים פעם אחת.

ועכשיו נחפש את **החציון**:



אבל איך נמצא אותו בדיוק?

ראשית, נמצא כמה נבדקים יש לנו. במקרה זה יש לנו 6 נבדקים. לצורך הענין נמספר אותם. אנו יודעים שהחציון נמצא בין נבדק מס׳ 3 לנבדק מס׳ 4.



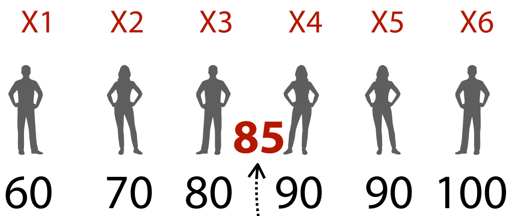
כדי למצוא את החציון אנחנו עושים ממוצע ציונים של שני הנבדקים שנמצאים באמצע, כלומר ממוצע ציונים של נבדק מס׳ 3 ונבדק מס׳ 4. במקרה הזה זה די קל. הממוצע בין 80 ל-90 הוא 85. החציון הוא הציון "85".

אבל מכיוון שתמיד נתמודד עם מסה עצומה של נבדקים, אנו זקוקים לנוסחה.

נגדיר את מספר הנבדקים הכולל באות **n**.

כל משתנה נגדיר באות **x** ולכל נבדק נצמיד מספר. זה יראה כך **X1**

כעת הפריסה שלנו תראה כך:



$$\frac{x_3+x_4}{2}=\frac{80+90}{2}=85$$

אבל כבר אמרנו שנתמודד עם מסה גדולה של אנשים, ולכן נזדקק לנוסחה קצת יותר משוכללת. הנוסחה תראה כך.

n מציינת את כמות הנבדקים הכוללת

הציון של הנבדק שנמצא באמצע

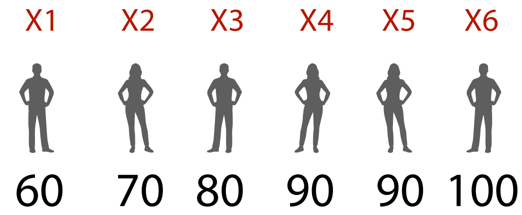
הציון של הנבדק שנמצא אחרי הנבדק שבאמצע

חציון

$$Md = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} = ?$$

את הציונים של שניהם נחלק לשניים כדי למצוא את הממוצע

ואם ניקח את הדוגמה האחרונה, הנוסחה תראה כך:



כמות הנבדקים חלקי שתיים + הנבדק הבא

כמות הנבדקים חלק שתיים

בדוגמה שלנו מדובר ב-6 נבדקים

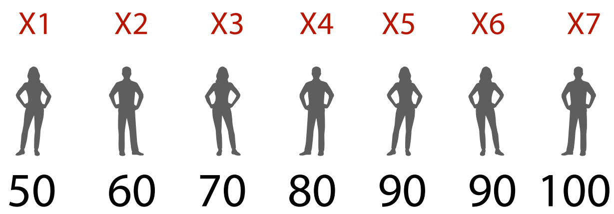
מצאנו מי הם שני הנבדקים שבאמצע

הציונים של נבדק 3 ונבדק 4

החציון

$$Md = \frac{X_{\frac{n}{2}} + X_{\frac{n}{2}+1}}{2} = \frac{X_{\frac{6}{2}} + X_{\frac{6}{2}+1}}{2} = \frac{X_3 + X_4}{2} = \frac{80+90}{2} = \mathbf{85}$$

כך מוצאים חציון כאשר כמות האנשים זוגית. אבל מה קורה כשכמות האנשים אינה זוגית - כמו בדוגמה הבאה?



במצב בו יש כמות נבדקים זוגית, נשתמש בנוסחה הרבה יותר פשוטה והיא תראה כך:

7 נבדקים + 1

כמות הנבדקים + הנבדק הבא, חלקי שתיים

8 נבדקים בסה"כ לחלק לשניים

מצאנו שנבדק מס' 4 הוא זה שנמצא באמצע

$$Md = \frac{X_{n+1}}{2} = \frac{X_{7+1}}{2} = \frac{X_8}{2} = X_4 = \mathbf{80}$$

החציון הוא הציון של נבדק מס' 4

חשוב לזכור! אנו צריכים טבלה כדי שנוכל להתמודד עם מסה גדולה של נבדקים

ועכשיו תרגול:

נבדקו מחירי שבע דירות בבאר-שבע:

מחירי הדירות במיליוני שקלים הם: 0.7, 1, 1.1, 1.3, 1.5, 1.5, 1.9

א. מהו החציון?

ב. מהו השכיח?

ג. מהו הממוצע?

	X7	X6	X5	X4	X3	X2	X1
מיקום הבתים	X7	X6	X5	X4	X3	X2	X1
מחיר הבתים במליוני ש"ח	1.9	1.5	1.5	1.3	1.1	1.0	0.7

א. **החציון** Md:

מכיוון שיש לנו מס' נבדקים אי זוגי, נשתמש בנוסחה הפשוטה...

8 בתים בסה"כ לחלק לשניים

7 בתים + 1

מצאנו שבית מס' 4 הוא זה שנמצא באמצע

$$Md = \frac{X_{7+1}}{2} = \frac{X_8}{2} = X_4 = \mathbf{1.3}$$

החציון הוא המחיר של בית מס' 4

ב. **השכיח** Mo:

השכיח הוא המחיר שחוזר

על עצמו הכי הרבה פעמים
במקרה שלנו אלו בתים
ששווים **1.5** מליון ש"ח...

ג. **הממוצע:**
ממוצע המחירים הוא סך כל מחירי הבתים, לחלק מספר הבתים הכולל, שהוא **7**.

$$\frac{0.7 + 1.0 + 1.1 + 1.3 + 1.5 + 1.5 + 1.9}{7} = \frac{9}{7} = \mathbf{1.285}$$

*טווח (R)

הטווח הוא ההפרש בין המדד הנמוך/קטן ביותר למדד הגבוה/רב יותר

לדוגמה:

X כמות פוסטים	f אנשים
0	2
3	8
7	4
10	7
15	8
30	10

בדוגמה הזאת בדקנו כמה פוסטים אנשים קוראים ברשתות החברתיות.
ה-**X** הוא כמות הפוסטים.
ה-**f** הוא מספר האנשים שקוראים כמות פוסטים מסוימת.

כדי למצוא את הטווח אנחנו צריכים לשאול כמה הכי פחות פוסטים אדם קורא וכמה הכי הרבה ולמצוא את ההפרש ביניהם - זהו **הטווח**.
אז בדוגמה שלנו:
0 - היא כמות הפוסטים **המועטה** ביותר.
30 - היא כמות הפוסטים **הרבה** ביותר.
ההפרש בין 0 ל 30 = 30.
לכן:
הטווח הוא 30.

והנוסחה לחישוב טווח תיראה כך:

$$\mathbf{R} = \underbrace{X_{max}}_{\text{פחות כמות פוסטים מקסימלית}} - \underbrace{X_{min}}_{\text{כמות פוסטים מינימלית}} = 30 - 0 = \mathbf{30}$$

חישוב אמצע טווח (**mr**) נראה כך:

$$\mathbf{mr} = \frac{X_{max} + X_{min}}{2} = \frac{70 + 100}{2} = \mathbf{8.5}$$

מכיוון שהדוגמה הזאת כבר יותר מתאימה לטבלה עם מידע רב, בואו ננצל אותה כדי למצוא את מדדי המרכז, שהם: **הממוצע**, **החציון והשכיח**.

כזכור **f** קטנה מייצגת את השכיחות. במקרה שלנו, כמות האנשים שקראו כמות פוסטים מסוימת. לדוגמה 8 אנשים קראו כל אחד 3 פוסטים.
אבל כשמדובר במסה של נתונים, לא נוכל לפרוס אותה כמו בדוגמאות הקודמות. ולכן, נשתמש במקרה הזה ב-**F** גדולה (קפיטל).

F גדולה מייצגת את סך האנשים המצטבר שקראו את סך כל הפוסטים.

X כמות פוסטים	f אנשים	F כמות אנשים מצטברת	
0	2	**2**	עד כאן קראו 2 אנשים פוסטים.
3	8	**10**	עד כאן קראו 10 אנשים פוסטים.
7	4	**14**	עד כאן קראו 14 אנשים פוסטים.
10	7	**21**	עד כאן קראו 21 אנשים פוסטים.
15	8	**29**	עד כאן קראו 29 אנשים פוסטים.
30	10	**39**	עד כאן קראו 39 אנשים פוסטים.

במקרה שלנו **F** הוא **39**. והוא יאפשר לנו למצוא באמצעות הנוסחאות שלנו את מדדי המרכז.

בואו נמצא את **החציון** באמצעות הנוסחה.
מדובר במספר אי זוגי של אנשים (**39**) ולכן נשתמש בנוסחה לחישוב חציון שמתאימה למספר אי זוגי של נבדקים, והיא תראה כך:

39 נבדקים 1+

40 נבדקים לחלק לשניים

האדם ה-20 הוא זה שנמצא באמצע

$$Md = \frac{X_{39+1}}{2} = \frac{X_{40}}{2} = X_{20} = \mathbf{10}$$

כמות הפוסטים שהאדם ה-**20** קורא היא **10**

X כמות פוסטים	f אנשים	F כמות אנשים מצטברת
0	2	**2**
3	8	**10**
7	4	**14**
10	7	**21**
15	8	**29**
30	10	**39**

האדם ה-20 קורא 10 פוסטים

האדם ה-20 נמצא כאן

החציון הוא 10 פוסטים

*שכיחות יחסית מצטברת

זוהי דרך נוספת למצוא את החציון. אנו עושים זאת באמצעות חישוב כמות האנשים באחוזים

כמות פוסטים (x)	אנשים (f)	כמות אנשים מצטברת (F)	שכיחות יחסית מצטברת	
0	2	2	**5.12%**	כלל האנשים עד כאן באחוזים
3	8	10	**25.6%**	
7	4	14	**35.89%**	
10	7	21	**53.8%**	**החציון** נמצא היכן שהוא כאן - באיזור ה-50%
15	8	29	**74.3%**	
30	10	39	**100%**	כלל האנשים עד כאן באחוזים

החציון הוא 10 פוסטים

כדי למצוא את **הממוצע** בטבלה נחשב כמה פוסטים נקראים על ידי כמה אנשים בסך הכל:

למשל, 3 אנשים
קוראים 8 פוסטים
כל אחד, וכן הלאה...

$$\bar{X} = \frac{0 \times 2 + 3 \times 8 + 7 \times 4 + 10 \times 7 + 15 \times 8 + 30 \times 10}{39} =$$

$$= \frac{0 + 24 + 28 + 70 + 120 + 300}{39} = \frac{542}{39} = 13.897$$

מספר הפוסטים שנקראים **בממוצע**.

נותר לנו למצוא את **השכיח**. כלומר, מספר הפוסטים שנקרא ע"י הכי הרבה אנשים...

אז מהו **השכיח**?
השכיח הוא ה-**x** שה-**F** שלו הוא הכי גבוה. לכן, **30 פוסטים**, כי הוא נקרא ע״י **10** אנשים, שהיא כמות האנשים הכי גדולה שקוראת את אותה כמות פוסטים.

x כמות פוסטים	f אנשים	F כמות אנשים מצטברת	שכיחות יחסית מצטברת
0	2	2	**5.12%**
3	8	10	**25.6%**
7	4	14	**35.89%**
10	7	21	**53.8%**
15	8	29	**74.3%**
30	10	39	**100%**

כמות הפוסטים שנקראת ע״י הכי הרבה אנשים

הכי הרבה אנשים שקוראים פוסטים

*מאון (או אחוזון) (Q)

מושג נוסף הוא המאון (מהמילה מאה).
אפשר למצוא כל מאון שנרצה על גבי הטבלה, כמו: מאון 30, מאון 25, מאון 61, מאון 50 (מאון 50 הוא במקרה גם החציון שלנו).
אז למה אנחנו צריכים את המאון הזה? כי לא תמיד נחפש את החציון - את נקודת האמצע. לפעמים נרצה למצוא נתונים בנקודה אחרת על ה-100%.

למשל, אם ישאלו אותנו מהו המאון ה-30 בדוגמה שלנו? למעשה שאלו אותנו מהו הערך שעד אליו יש לנו 30% אנשים שקראו מספר פוסטים זהה. והתשובה...

זו הסיבה שחשוב לנו לחשב שכיחות יחסית מצטברת

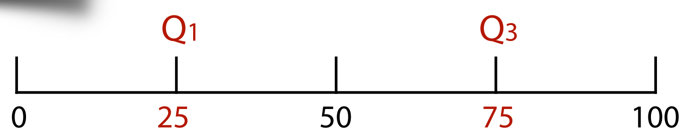
מאון 30 נמצא כאן, בתוך ה-35.89%. ולכן, **מאון 30 הוא 7 פוסטים**. (מתייחס לפוסטים הנקראים ולא לכמות האנשים.

0	2	2	**5.12%**
3	8	10	**25.6%**
7	4	14	**35.89%**
10	7	21	**53.8%**
15	8	29	**74.3%**

*טווח בינרבעוני (IQR)

זו הסיבה שחשוב לנו לדעת למצוא מאון (אחוזון)

ישנן שתי נקודות בטבלה או על הסקלה שנקראות "רבעונים":
רבעון 25 שנקרא גם **רבעון תחתון - Q_1**
ו**רבעון 75** שנקרא גם **רבעון עליון - Q_3**



טווח בינרבעוני (בין-רבעוני) הוא הטווח שבין הרבעון התחתון לרבעון העליון.
כדי למצוא את הטווח, נצטרך קודם למצוא מהו הרבעון התחתון ומהו הרבעון העליון.

ושוב, אם נביט בדוגמה שלנו, נגלה שהרבעון התחתון (ה-25%) נמצא בתוך ה 25.6%, והרבעון העליון (ה-75%) נמצא בתוך ה 100%. כל שנותר לנו הוא למצוא את ה-X (המשתנה שהוא כמות פוסטים). כלומר, כמה פוסטים נקראים ברבעון התחתון וכמה פוסטים נקראים ברבעון העליון ולחשב את ההפרש ביניהם…

X כמות פוסטים	f אנשים	F כמות אנשים מצטברת	שכיחות יחסית מצטברת
0	2	2	**5.12%**
3	8	10	**25.6%**
7	4	14	**35.89%**
10	7	21	**53.8%**
15	8	29	**74.3%**
30	10	39	**100%**

ברבעון התחתון נקראים 3 פוסטים

הרבעון התחתון נמצא כאן

ברבעון העליון נקראים 30 פוסטים

הרבעון העליון נמצא כאן

והנוסחה תהיה:

פחות מס' הפוסטים שנקראים ונמצאים ברבעון התחתון

מס' הפוסטים שנקראים ונמצאים ברבעון העליון

$$Q_3 - Q_1 = 30 - 3 = \mathbf{27}$$

הטווח הבינרבעוני הוא 27

*מדדי פיזור

כאמור, **מדדי מרכז** (או "מדדי מיקום מרכזי") מוגדרים כ**חציון**, **ממוצע** ו**שכיח**.
בשונה ממדדי מרכז, ישנם מדדים נוספים שנקראים מדדי פיזור. על חלקם כבר למדנו, כמו: **טווח** ו**טווח בינרבעוני**.
שני מדדי הפיזור הנוספים הם שונות וסטיית תקן.

*שונות (S^2)

מה זה שונות? זה לא כל כך חשוב. מה שחשוב זה למה אנחנו צריכים למצוא שונות. התשובה היא שהיא עוזרת לנו למצוא **סטיית תקן**. אז לפני שנגיע לדבר על סטיית תקן, בואו נבין איך מוצאים שונות.

הנוסחה למציאת שונות דומה לנוסחה למציאת ממוצע,
רק שכאן ה-X הוא בריבוע. והנוסחה תראה כך:

ערך משותף למס' אנשים

סכום

כפול כמות האנשים עם אותו ערך

$$S^2 = \frac{\Sigma X^2 \times f}{n} - \bar{X}^2$$

פחות הממוצע בריבוע

לחלק למספר האנשים הכולל

וכדי להבין איך הנוסחה עובדת, נשתמש בדוגמה הבאה:
100 אנשים קיבלו ציונים לפי ההתפלגות הבאה:

F	f	X
20	20	70
50	30	80
80	30	90
90	10	92
100	10	100

שימו לב: שבהתפלגות שבדוגמה ישנם שני שכיחים (70 ו-90).
30 אנשים קיבלו ציון 70, ו-30 אנשים נוספים קיבלו ציון 90. כן, זה אפשרי שיהיה יותר משכיח אחד.

כדי לחשב שונות, אנחנו צריכים קודם כל למצוא ממוצע:

ממוצע

$$\bar{X} = \frac{70 \times 20 + 80 \times 30 + 90 \times 30 + 92 \times 10 + 100 \times 10}{100} = 84.2$$

כעת, כשיש לנו את הממוצע, נוכל לחשב שונות:

שונות

$$S^2 = \frac{70^2 \times 20 + 80^2 \times 30 + 90^2 \times 30 + 92^2 \times 10 + 100^2 \times 10}{100} - 84.2^2 = 86.76$$

*סטיית תקן

סטיית תקן היא מדד לתיאור פיזור של נתונים מספריים סביב הממוצע שלהם.

סטיית תקן $S = \sqrt{S^2}$ שונות בריבוע

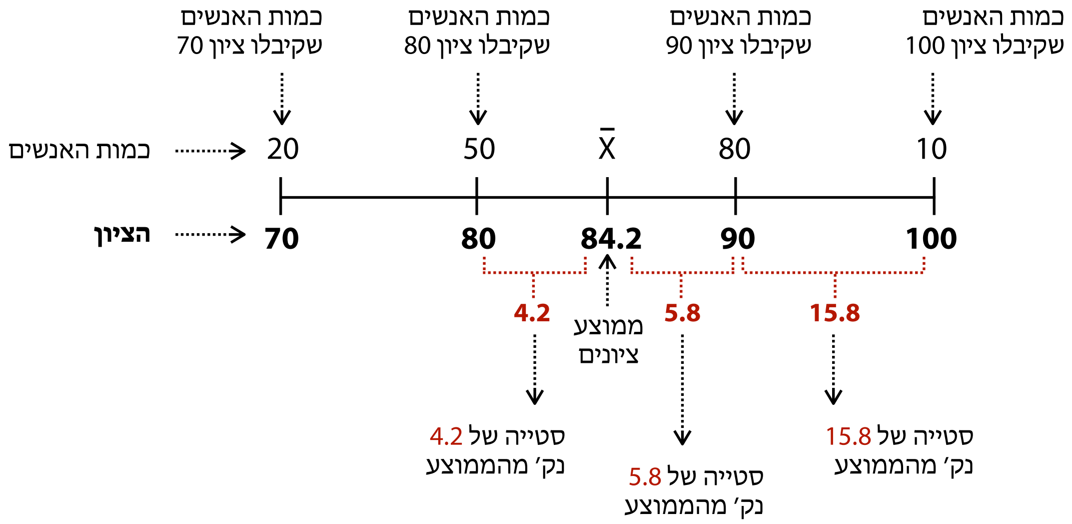
לסיכום, כדי למצוא **סטיית תקן**, אנחנו צריכים למצוא **שונות**, להוציא ממנה שורש וקיבלנו סטיית תקן.

כן, למצוא **סטיית תקן (S)** זה פשוט לחשב שורש של **שונות**.

בואו נמצא את סטיית התקן שבדוגמה האחרונה באמצעות הנוסחה:

$$S = \sqrt{S^2} = \sqrt{86.76} = \mathbf{9.31}$$

אבל למה אנחנו צריכים למצוא סטיית תקן?
פשוט מאוד. סטיית תקן משכללת לנו את כל המרחקים סביב הממוצע. במילים אחרות, סטיית תקן מראה לנו בכמה כל אדם סטה מהממוצע. היא מחברת את כל המרחקים של האנשים מאותו ממוצע והיא מחלקת בכמות האנשים הכללית.



*פלט סטטיסטיקה תיאורית

פלט של שנות ותק של עובדים במפעל:

X - שנות ותק של האנשים

f - כמות האנשים שיש להם כך וכך שנות ותק

לא צריך לעניין אותנו

כמות האנשים עם אותן שנות ותק, מתוך כלל האנשים

אחוז מצטבר

		Frequency	Percent	Valid Percent	Cumulative Percent
Valid	1	3	3.6	3.6	3.6
	2	2	2.4	2.4	6.0
	5	8	9.5	9.5	15.5
	7	7	8.3	8.3	23.8
	8	4	4.8	4.8	28.6
	9	8	9.5	9.5	38.1
	11	2	2.4	2.4	40.5
	12	8	9.5	9.5	52.0
	14	10	11.9	11.9	61.9
	15	1	1.2	1.2	63.1
	16	8	9.5	9.5	72.6
	17	18	21.4	21.4	94.0
	19	2	2.4	2.4	96.4
	20	1	1.2	1.2	97.6
	21	1	1.2	1.2	98.8
	22	1	1.2	1.2	100.0
	Total	84	100.0	100.0	

שנת ותק אחת

למשל, 8 אנשים שיש להם 5 שנות ותק

11 שנות ותק

עד 22 שנות ותק

n - סך כל האנשים

תמיד נגיע ל-100%

alid Percent	C
3.6	
2.4	
9.5	
8.3	
4.8	

למשל, 15.5 זה האחוז שהצטבר מחישוב של 3.6 + 2.4 + 9.5.

בפלט אין לנו **f** מצטבר, אלא יש לנו אחוז מצטבר (Cumuliative Percent).

החציון: כדי למצוא את החציון בפלט, קודם נחפש בעמודת **Cumuliative Percent** את הערך בו נמצא ה-50% בפעם הראשונה. במקרה שלנו מדובר ב-51.3%. ולכן החציון הוא 12.

11	2	2.4	2.4	40.5
12	8	9.5	9.5	52.0
14	10	11.9	11.9	61.9

השכיח:
השכיח, כאמור, הוא הערך שהשכיחות שלו הי הכי גבוהה.
לשם כך נחפש בטבלת ה- **Frequency** היכן יש הכי הרבה עובדים במפעל שיש להם הכי הרבה שנות ותק.

נמצא ש-**18** עובדים הוא המספר הכי גבוה ולעובדים אלה יש **17** שנות ותק.
כלומר, השכיח הוא **17**.

		Frequency	Per
Valid	1	3	
	2	2	
	5	8	
	7	7	
	8	4	
	9	8	
	11	2	
	12	8	
	14	10	
	15	1	
	16	8	
	17 ←	18	
	19	2	
	20	1	
	21	1	
	22	1	
	Total	84	

ממוצע:
ממוצע נצטרך לחשב.

$$\frac{1 \times 3 + 2 \times 2 + 5 \times 8 + 7 \times 7 \ldots}{84} =$$

שימו לב שבמקרה הזה החציון הוא 12 והשכיח הוא 17.

ב**התפלגות א סימטרית חיובית** (זנב ימני), הממוצע גדול מהחציון והחציון גדול מהשכיח.

ב**התפלגות א סימטרית שלילית** (זנב שמאלי), השכיח גדול מהחציון והחציון גדול מהממוצע.

כלומר, במקרה שלנו מדובר בהתפלגות שלילית.

			Frequency	Per
Valid	1	1 x 3 +	3	
	2	2 x 2 +	2	
	5	5 x 8 +	8	
	7	7 x 7 +	7	
	8	וכן הלאה	4	
	9	...	8	
	11		2	
	12		8	
	14		10	
	15		1	
	16		8	
	17		18	
	19		2	
	20		1	
	21		1	
	22		1	
	Total		84	

מאון:

בואו נמצא למשל את **מאון-90** (העשירון העליון). לשם כך, נחפש בעמודת **Cumuliative Percent** את המקום בו בפעם הראשונה אנו מזהים 90%. בפעם הראשונה אנו מזהים 90% בשורה בה מופיע 94% ולכן המאון ה-90 הוא **17**.

15	1	1.2	1.2	63.1
16	8	9.5	9.5	72.6
17 ←	18	21.4	21.4	94.0
19	2	2.4	2.4	96.4

רבעון תחתון (Q1):

הרבעון התחתון הוא המאון ה-25. ה-X שמתחתיו יש לנו 25% מהעובדים.
בפעם הראשונה אנו מזהים 25% ומעלה בשורה בה רשום 28.6%, וה-X שלה הוא **8**.

5	8	9.5	9.5	15.5
7	7	8.3	8.3	23.8
8	4	4.8	4.8	28.6
9	8	9.5	9.5	38.1
11	2	2.4	2.4	40.5

רבעון עליון (Q3):

הרבעון העליון הוא המאון ה-75. ה-X שמתחתיו יש לנו 75% מהעובדים.
בפעם הראשונה אנו מזהים 75% ומעלה בשורה בה רשום 94.0%, וה-X שלה הוא **17**.

15	1	1.2	1.2	63.1
16	8	9.5	9.5	72.6
17	18	21.4	21.4	94.0
19	2	2.4	2.4	96.4
20	1	1.2	1.2	97.6

הטווח הבין רבעוני:

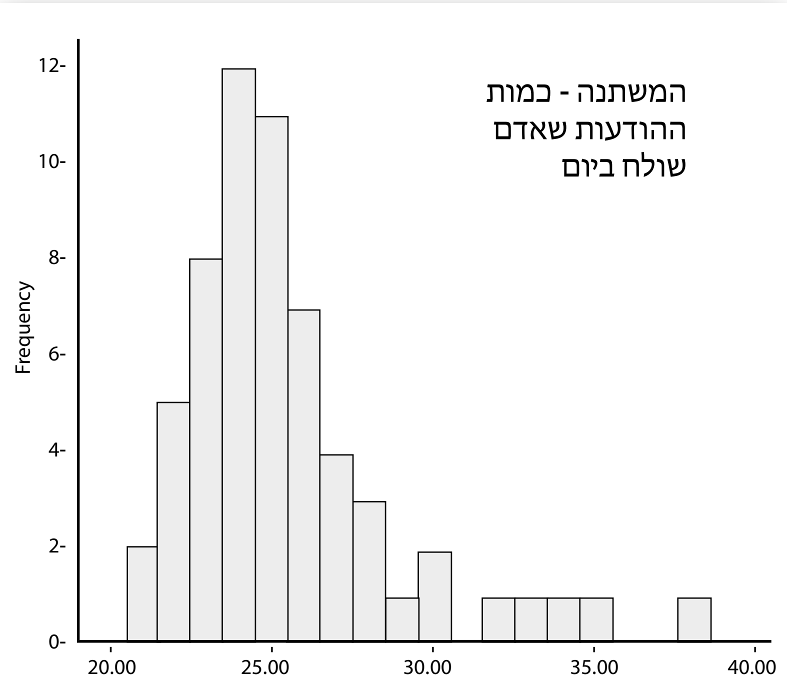
זהו הטווח שבין הרבעון העליון לתחתון. כלומר, ההפרש בין 17 שנות ותק ל-8 שנות ותק.

$$Q_3 - Q_1 = 17 - 8 = 9$$

בטבלה שלפנינו אנו רואים שהתוכנה כבר ביצעה לנו את כל החישובים שאנחנו צריכים:

N	Valid	84	סך כל האנשים שנבדקו.
	Missing	0	כל האנשים שבדקנו נכחו ולא אבד שום טופס.
Mean		12.00	הממוצע.
Meidan		13.00	החציון.
Mode		17	השכיח.
Std. Deviation		5.193	סטיית התקן.
Variance		26.964	השונות (סטיית התקן בריבוע).
Range			הטווח.
Minimum		1	ה-x הנמוך ביותר (שנת ותק אחת).
Maximum		22	ה-x הגבוה ביותר (20 שנות ותק).
Sum		1008	סכום כל f כפול ה-x שלו (סך כל העובדים כפול כל שנות הוותק שלהם).
Percentiles	20	7.00	מאונים: מאון 20
	25	8.00	מאון 25 (Q1)
	40	11.00	מאון 40
	50	13.00	מאון 50 (שהוא גם החציון).
	60	14.00	מאון 60
	75	17.00	מאון 75 (Q3)
	80	17.00	מאון 80

*פלט מול היסטוגרמה



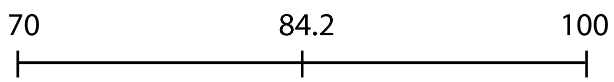
N	Valid	60
Missing		
Mean		25.5667
Meidan		25.0000
Mode		24.00
Std. Deviation		3.39674
Variance		11.538
Range		17.00
Minimum		21.00
Maximum		38.00
Sum		1534.00
Percentiles	25	23.2500
	50	25.0000
	70	26.0000
	75	26.7500

*השפעת הוספת נתונים חדשים על מדדי מרכז ופיזור

- **הממוצע**: ברגע שמוסיפים אנשים חדשים לנתונים, הממוצע משתנה. אם הם יביאו ערכים נמוכים מהממוצע (כמו ציון במבחן), הממוצע ירד. אם הם יביאו ציונים גבוהים מהממוצע, הממוצע יעלה.
- **השכיח**: ישתנה בהתאם
- **החציון**: בחציון נמצאים 50% מהאנשים מעליו ו-50% אנשים מתחתיו. כלומר, גם החציון ישתנה. אם האנשים הביאו ערכים נמוכים מהממוצע, החציון ירד (יזוז שמאלה), כדי לאזן את כמות האנשים שמצדדיו. וההפך.
- סטיית התקן: אם נוסיף נתונים חדשים שסטיית התקן שלהם קטנה יותר מסטיית התקן, אז סטיית התקן תקטן, וההפך. לא משנה היכן הם ממוקמים (מפוזרים).

חוק: כשמוסיפים נתונים חדשים מתחת לחציון - החציון יקטן, וההפך! אם נוסיף את אותה כמות אנשים משני הצדדים, החציון לא ישתנה.

X	f	F
70	20	20
80	30	50
90	30	80
92	10	90
100	10	100



עדיין לא סיימתי את החלק הזה

*חישוב ממוצע וסטיית תקן במשתנה רציף

מחלקות

בטבלה עם מחלקות קיימים ערכים של X בין לבין. למשל, 1 עד 5 נבחנים שקיבלו ציון 70.
לדוגמה: אנשים שמעשנים סיגריות.
יש לשים לב שלכל מחלקה (**X**) יש גבול עליון וגבול תחתון. הם נקראים **L** (Limit).
הגבול התחתון מסומן ב $\mathbf{L_0}$ (L-אפס) כמות הסיגריות הנמוכה באותה שורה. הגבול העליון מסומן ב $\mathbf{L_1}$ (L-אחד), שהוא כמות הסיגריות הגבוהה באותה שורה. בדוגמה שלנו, במחלקה הראשונה בדוגמה הערכים הם 0 ו-10.
בדוגמה שבזאת, נצטרך להוסיף שלוש עמודות נוספות, והן יקראו:

- **אמצע מחלקה**: כשיש לנו שני ערכים באותה מחלקה ואנחנו רוצים למצוא את האמצע שלהם - נחבר ביניהם ונחלק לשתיים. כלומר, גבול עליון ועוד גבול תחתון חלקי שתיים.
- **רוחב מחלקה**: גבול עליון פחות גבול תחתון.
- **צפיפות מחלקה**: זוהי עמודה של שכיחות לעמודת רוחב אחת. אנו לוקחים את צפיפות האנשים במחלקה ומחלקים בנתון מצאנו בעמודת הרוחב.

מח׳	סיגריות	אנשים	מצטבר	אמצע מחלקה	רוחב מחלקה	צפיפות מחלקה
	X	f	F	Xi		
1	0-10	25	25	0+10/2=5	10-0=10	25/10=2.5
2	10-20	70	95	10+20/3=10	20-10=10	70/10=7
3	20-40	65	160	20+40/2=30	40-20=20	65/20=3.75
4	40-80	40	200	40+80/2=60	80-40=40	40/40=1
			n=200			

שכיח (בטבלה של מחלקות): הוא אמצע המחלקה הצפופה ביותר.
במקרה שלנו, המחלקה בה הצפיפות הגדולה ביותר היא מחלקה מס׳ 2. והאמצע שלה הוא 10. לכן השכיח הוא 10.

$$Mo = 15$$

ממוצע (בטבלה של מחלקות): כמו במשתנה בדיד, משתמשים באותה הנוסחה, רק בשימוש בעמודות החדשות.

הערך המשותף של ה-X באותה מחלקה

סכום

$$\bar{X} = \frac{\Sigma X_i \times f(X)}{n}$$

הערך המשותף של ה-X באותה מחלקה

לחלק למספר האנשים הכולל

$$\bar{X} = \frac{5x25+15x70+30x65+60x40}{200} = 27$$

סטיית התקן (בטבלה של מחלקות): כדי למצוא את סטיית התקן, כאמור, נצטרך קודם למצוא שונות.

ערך משותף למס' אנשים

סכום

כפול כמות האנשים עם אותו ערך

$$S^2 = \frac{\sum Xi^2 \times f}{n} - \bar{X}^2$$

לחלק למספר האנשים הכולל

פחות הממוצע בריבוע שמצאנו קודם

$$S^2 = \frac{5^2 \times 25 + 15^2 \times 70 + 30^2 \times 65 + 60^2 \times 40}{200} - 27.626^2 = 331$$

וכדי למצוא סטיית תקן נצטרך להוציא שורש מהשונות:

סטיית תקן $\rightarrow S = \sqrt{S^2} \leftarrow$ שונות בריבוע

$S = \sqrt{331} = 18.19 \leftarrow$ סיגריות

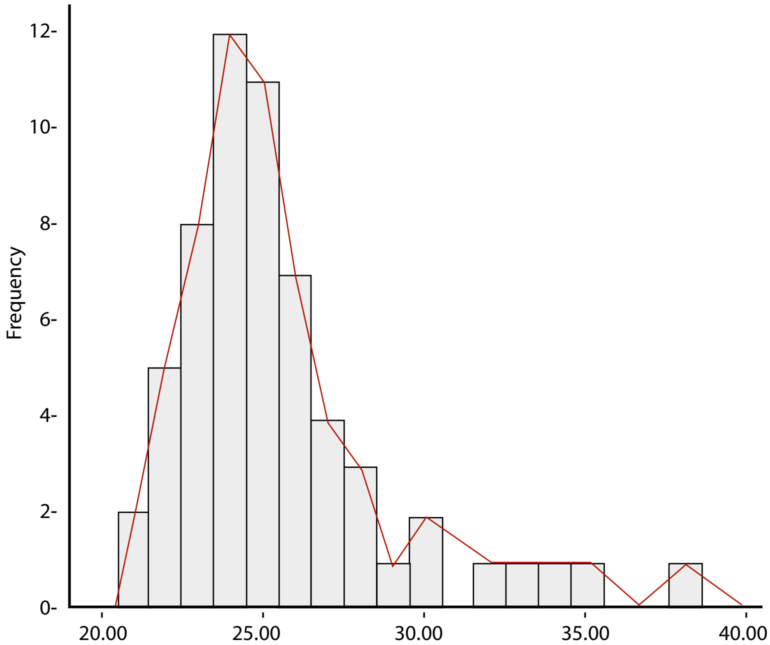
טווח (בטבלה עם מחלקות):
ניקח את ה-X הכי גבוה במחלקות ונחסיר ממנו את ה-X הכי נמוך.

$$(\text{Range})\ R = X\text{max} - X\text{min} = 80 - 0 = 80$$

אמצע טווח (בטבלה עם מחלקות):
האמצע של שני הקצוות. כלומר, נחבר בין ה-X הכי נמוך ועוד ה-X הכי גבוה ונחלק לשניים.

$$(\text{Middle range})\ Mr = \frac{X\text{max} - X\text{min}}{2} = \frac{0 + 80}{2} = 40$$

*סוגי התפלגויות תצוגה גרפית



עדיין עובד על החלק הזה

סימטרי, א-סימטרי חיובי, א-סימטרי שלילי (לסכם)

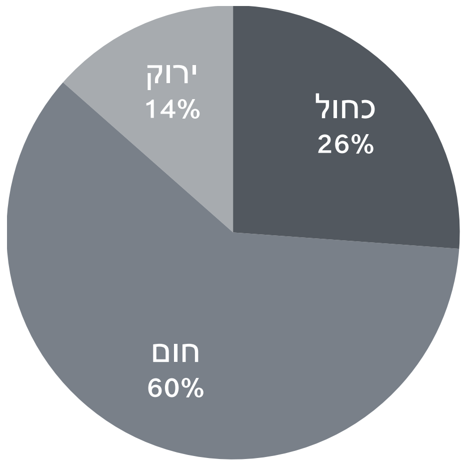
כשהמשתנה הנחקר הוא **בסולם סדר**, ניתן לחשב חציון ושכיח ולהמחיש את ההתפלגות בעזרת **גרף מקלות**.

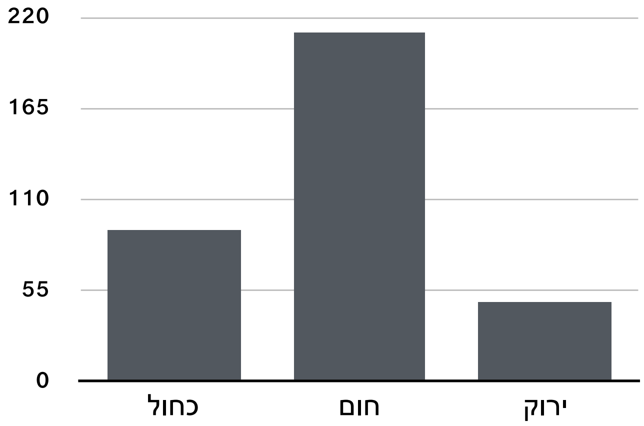
חוק! כשהמשתנה הנחקר הוא **בסולם שמי**, ניתן לחשב רק שכיח ולתאר את ההתפלגות בגרף פאי בלבד.

הערות בשבילי
ממוצע לא מייצג את המציאות האמיתית
חציון כן מייצג.
למשל, שכר במשק. יש כאלו שמקבלים 2000 ש"ח, הרוב מקבלים 6000 ש"ח ויש מעטים שמקבלים מאות אלפים. המעטים שמקבלים מאות אלפים מושכים את הממוצע כלפי מעלה לכיוון ה-8000 ש"ח.
מה שעושה החציון הוא להציב את הממוצע על הנקודה שבה יש כמות גדולה של אנשים, אלו שמרוויחים 6000 ש"ח. החציון יהיה רחוק מהממוצע וקרוב לכמות האנשים הגדולה ביותר.

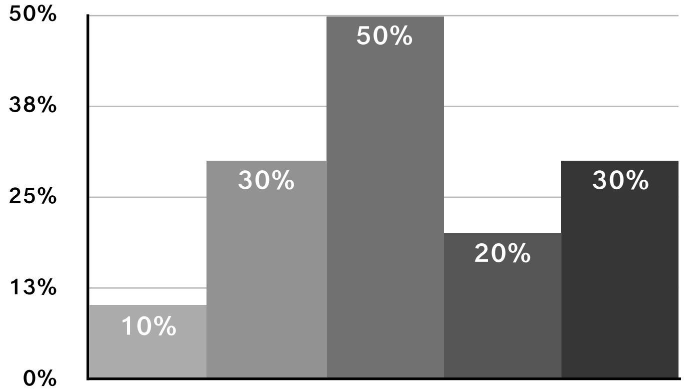
*סוגי תרשימים

פאי ו**עמודות**: משתמשים בהם לרוב במשתנים איכותיים. היא מתאימה למשתנים שאין להם הרבה ערכים.





היסטוגרם: משמש לתיאור שכיחות של משתנה רציף.



פוליגון: