מבוא לסטטיסטיקה, מועד א 2026

זה לא פיתרון רשמי של צוות הקורס. בהצלחה :) יאן, 0535-290202

שאלה 1

נתון וקטור שבו הנתונים הם $x_1, x_2, ..., x_n$. בנתונים אלה, הטווח הוא R וסטיית התקן היא s.

נגדיר וקטור חדש כך שלכל i:

$$y_i = 3x_i - 5$$

נכון או לא נכון: בוקטור החדש הטווח שווה ל-$(3R - 5)$ וסטיית התקן שווה ל-$3s$. הסבירו.

הטענה היא **לא נכונה (חלקית)**.

לפני כל הפירוק המתמטי (שיכול להיראות מאיים אבל הוא באמת לא מסובך), טווח מושפע מטרנספורמציה לינארית רק בערך ההכפלה. בואו נראה:

1. חלק הטווח: הטווח הוא מקסימום פחות מינימום, לכן

$$\text{Range } (Y) = \max(y) - \min(y)$$

ומכיוון ש-$3x - y = 5$:

$$\max(y) = 3\max(x) - 5, \qquad \min(y) = 3\min(x) - 5$$

ולכן:

$$\text{Range } (Y) = [3\max(x) - 5] - [3\min(x) - 5]$$

$$= 3[\max(x) - \min(x)] = 3R$$

ה--5 מתבטל בחיסור, ולכן הטווח **לא** נהיה $3R - 5$, אלא $3R$.

2. חלק סטיית התקן: טרנס' לינארית משפיעה על סטיית התקן בהכפלה בקבוע בערך מוחלט b בלבד. לכן עבור $3X - Y = 5$:

$$S_Y = |3|\, S_X = 3S$$

הערה: בדף נוסחאות מופיע לכם בדף נוסחאות מופיע לכם $Var[X]\, b^2 \cdot Var[Y] =$. זה בדיוק כמו $S[X]\, |b| \cdot S[Y] =$

קוד R לאימות

```r
x <- c(10, 20, 30)
r_x <- diff(range(x))
s_x <- sd(x)

y <- 3 * x - 5
r_y <- diff(range(y))
s_y <- sd(y)

tibble::tibble(
  quantity = c("Range", "SD"),
  original = c(r_x, s_x),
```

```
  transformed = c(r_y, s_y),
  ratio = c(r_y / r_x, s_y / s_x)
)
```

שאלה 2

בניסוי מקרי, ההסתברות להצלחה היא 0.7. מאורע A = הצלחה אחת בדיוק מתוך 5 ניסיונות ומאורע B = הצלחה בניסיון החמישי.

נכון/לא נכון: יש תלות בין המאורעות A ו-B. הסבירו.

נכון: יש תלות בין A ו-B.

נסמן $p = 0.7$, $q = 0.3$.

1. הסתברות של A: בדיוק הצלחה אחת מתוך 5:

$$P(A) = \binom{5}{1} p^1 q^4 = 5 \cdot 0.7 \cdot 0.3^4 = 0.02835$$

2. הסתברות של B: הצלחה בניסיון החמישי:

$$P(B) = 0.7$$

3. חיתוך $A \cap B$: אם יש בדיוק הצלחה אחת ובנוסף ההצלחה היא בניסיון החמישי, אז 4 הניסיונות הראשונים חייבים להיות כישלון:

$$P(A \cap B) = q^4 p = 0.3^4 \cdot 0.7 = 0.00567$$

4. בדיקת אי-תלות:

$$P(A)P(B) = 0.02835 \cdot 0.7 = 0.019845$$

קיבלנו:

$$P(A \cap B) = 0.00567 \neq 0.019845 = P(A)P(B)$$

ולכן המאורעות **תלויים**.

נסתכל על הבעיה מכיוון אחר: האם ההסתברות האפריורית שונה מההסתברות הפוסטריורית?

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)} = \frac{0.00567}{0.7} = 0.0081$$

וזה שונה מ-$P(A) = 0.02835$. קיצר כן יש תלות ביניהם.

קוד R לאימות

```
p <- 0.7
q <- 1 - p

pA <- dbinom(1, size = 5, prob = p)
pB <- p
```

```r
pAnB <- q^4 * p

c(
  pA = pA,
  pB = pB,
  pAnB = pAnB,
  pA_times_pB = pA * pB,
  pA_given_B = pAnB / pB
)
```

שאלה 3

נכון/לא נכון: מכיוון ש-$P(A \cap B) \leq P(A)$, אז תמיד $P(A \mid B) < P(A)$. הסבירו.

לא נכון בכלל

נכון ואף טריוויאלי ש:

$$P(A \cap B) \leq P(A)$$

אבל שימו לב למכנה שמשתנה בהסתברות מותנית:

$$P(A \mid B) = \frac{P(A \cap B)}{P(B)}$$

כאשר מחלקים ב-$P(B)$, הערך יכול לגדול או להישאר זהה, בהתאם למידע ש-B נותן על A.

דוגמה נגדית פשוטה (קובייה הוגנת):

- $A = \{6\}$
- $B = \{2, 4, 6\}$ (מספר זוגי) אז:

$$P(A) = \frac{1}{6}$$

ו-

$$P(A \mid B) = \frac{P(\{6\})}{P(\{2, 4, 6\})} = \frac{1/6}{3/6} = \frac{1}{3}$$

ולכן:

$$P(A \mid B) = \frac{1}{3} > \frac{1}{6} = P(A)$$

כלומר, לא תמיד ההסתברות המותנית קטנה יותר.

קוד R לאימות

```r
set.seed(2026)
rolls <- sample(1:6, size = 1e5, replace = TRUE)

event_a <- rolls == 6
event_b <- rolls %% 2 == 0
```

```r
pA <- mean(event_a)
pAgivenB <- mean(event_a & event_b) / mean(event_b)

c(pA = pA, pA_given_B = pAgivenB)
```

שאלה 4

נכון/לא נכון: מתאם ספירמן לא דורש "תיקון" של סטיית תקן כי הוא מבוסס על דירוגים. הסבירו.

לא נכון

מתאם ספירמן מוגדר כמתאם פירסון על הדירוגים:

$$\rho_s = \text{cor}\ (R_X, R_Y)$$

כאשר R_X ו-R_Y הם הדירוגים.

כלומר, עדיין יש תקנון לפי סטיות תקן, אבל של **הדירוגים**:

$$\rho_s = \frac{\sum_{i=1}^{n}\left(R_{X,i} - \bar{R_X}\right)\left(R_{Y,i} - \bar{R_Y}\right)}{n\, s_{R_X}\, s_{R_Y}}$$

לכן לא נכון לומר שאין "חלוקה בסטיית תקן". יש חלוקה, פשוט אחרי המרה לדירוגים במקום ערכים גולמיים.

קוד R לאימות

```r
set.seed(2026)
x <- sample(100, 50, replace = TRUE)
set.seed(2025)
y <- sample(100, 50, replace = TRUE)

c(
  cor_of_ranks = cor(rank(x), rank(y)),
  spearman = cor(x, y, method = "spearman")
)
```

שאלה 5

בסקר מחירים נמצא:

$$\bar{x} = 32, \qquad \min = 28, \qquad \max = 36$$

נכון/לא נכון: הטווח הבין-רבעוני IQR קטן או שווה ל-8. הסבירו.

נכון

באופן אינטואיטיבי אנחנו מבינים שהטווח הבין רבעוני לא יכול להיות גדול מטווח המקסימום. בנוסחא פשוטה:

ראשית, הטווח הכולל הוא:

$$\text{Range} = \max - \min = 36 - 28 = 8$$

לפי הגדרה:

$$IQR = Q_3 - Q_1$$

ושני הרבעונים Q_1, Q_3 נמצאים בין המינימום למקסימום. לכן:

$$Q_3 - Q_1 \leq \max - \min$$

ומכאן:

$$IQR \leq 8$$

32 זו סתם הסחת דעת, מידע לא רלוונטי בשאלה הזו

קוד R לאימות

```
prices <- c(28, 28, 36, 36)

c(
  mean = mean(prices),
  range = diff(range(prices)),
  iqr = IQR(prices, type = 6)
)
```

שאלה 6

בגן יש 12 ילדים שעומדים במעגל. שלישייה מסוימת חייבת לעמוד יחד, וגם זוג מסוים חייב לעמוד יחד.

נכון/לא נכון: מספר הסידורים האפשריים הוא $483,840$. הסבירו.

נכון טוב כמובן שלמספר עצמו מגיעים בסוף. מה שחשוב זה לזהות את הקטע של הבלוקים הקבועים. אם לא היו הגבלות, כל ילד היה בלוק משלו, כיוון שיש הגבלות בחלוקה, נייצר בלוקים:

1. מאחדים את השלישייה לבלוק אחד.
2. מאחדים את הזוג לבלוק אחד.
3. שאר הילדים הם 7 יחידים.

לכן מספר ה"יחידות" במעגל הוא:

$$1 + 1 + 7 = 9$$

ניזכר שבמעגל מחסירים סידור אחד (של הראשון והאחרון):

$$(9 - 1)! = 8!$$

כעת מכפילים בסידורים פנימיים של הבלוקים:

- $3!$ עבור השלישייה
- $2!$ עבור הזוג

סה"כ:

$$8! \cdot 3! \cdot 2! = 40320 \cdot 6 \cdot 2 = 483840$$

קוד R לאימות

```
factorial(8) * factorial(3) * factorial(2)
```

שאלה 7

במעבדה לחקר הוויסות העצמי, מספר ניתוקי האינטרנט ביום עבודה מלא (8 שעות) מתפלג פואסונית עם תוחלת 4 ניתוקים ביום מלא:

סעיף א'

מה ההסתברות שביום שישי (יום עבודה חלקי של 6 שעות) יתרחשו בין 2 ל-4 ניתוקים (כולל)?

המלצה חמה, ברגע שאתם מזהים את סוג ההתפלגות והפרמטרים שלה, לכתוב אותה מסודר:

$$X_{8h} \sim \text{Poisson } (\lambda = 4)$$

בפואסון הפרמטר λ פרופורציוני לאורך זמן. בשאלה הוא מוגדר לפי יום עבודה של 8 שעות ולכן נצטרך להמיר אותו ליום של 6 שעות. ביום מלא: $\lambda = 4$. וביום שישי: 👇

$$\lambda_{6h} = 4 \cdot \frac{6}{8} = 3$$

כעת:

$$X_{6h} \sim \text{Poisson } (3)$$

בואו נכתוב מסודר מה בעצם רוצים מאיתנו

$$P(2 \leq X_{6h} \leq 4) = P(X = 2) + P(X = 3) + P(X = 4)$$

נחשב כל איבר לפי נוסחת פואסון:

$$P(X = k) = \frac{e^{-\lambda}\lambda^k}{k!}$$

ולכן:

$$P(X = 2) = \frac{e^{-3}3^2}{2!} = \frac{9e^{-3}}{2}$$

$$P(X = 3) = \frac{e^{-3}3^3}{3!} = \frac{27e^{-3}}{6} = \frac{9e^{-3}}{2}$$

$$P(X = 4) = \frac{e^{-3}3^4}{4!} = \frac{81e^{-3}}{24} = \frac{27e^{-3}}{8}$$

סכימה:

$$P(2 \leq X \leq 4) = \frac{9e^{-3}}{2} + \frac{9e^{-3}}{2} + \frac{27e^{-3}}{8} = \frac{99e^{-3}}{8} \approx 0.6161$$

כלומר ההסתברות היא בערך 0.616 (כ-61.6%).

קוד R לאימות

```
lambda8 <- 4
lambda6 <- lambda8 * 6 / 8

sum(dpois(2:4, lambda = lambda6))
```

סעיף ב'

ביום עבודה מלא (8 שעות), כבר עברו 20 דקות בלי ניתוק. מה ההסתברות שיעברו עוד 20 דקות עד הניתוק הבא?

נזהה את המעבר להתפלגות מעריכית שבה אנחנו מחשבים את הזמן שעובר עד הצלחה ראשונה. יחידת הזמן שוב משתנית ולכן נחלק את קצב האירועים ביום שלם של 8 שעות ב8 כדי לקבל קצב אירועים בשעה אחת.

קצב הניתוקים לשעה:

$$\lambda_{\text{hour}} = \frac{4}{8} = 0.5$$

הנתון "עברו כבר 20 דקות" אומר $T > 20 \text{ min}$.

$$P(T > 40 \text{ min } \mid T > 20 \text{ min })$$

אבל בגלל תכונת חוסר הזיכרון של התפלגות מעריכית: אגב השאלה בחלק א עם אי תלות.. האירוע האפריורי שווה לפוסטריורי! נוסחא מפונפנת שבסהכ מדגימה שזה שעבר זמן מסוים לא משנה את ההסתברות למאורע בזמן אחר

$$P(T > s + t \mid T > s) = P(T > t)$$

ולכן כאן:

$$P(T > 40 \text{ min } \mid T > 20 \text{ min }) = P(T > 20 \text{ min })$$

נעבור לשעות: 20 דקות = $\frac{1}{3}$ שעה.

נציב ולא נחשוש

$$P(T > t) = e^{-\lambda t}$$

$$P(T > 1/3) = e^{-0.5 \cdot (1/3)} = e^{-1/6} \approx 0.8465$$

כלומר ההסתברות היא בערך 0.8465 (כ-84.65%).

*נכון באותה מידה היה להפוך את הלמבדה לפי יחידות של דקה, ראו את קטע הקוד לדוגמא

קוד R לאימות

```
lambda_per_hour <- 4 / 8

# 20 דקות = 1/3 שעה
```

```r
1 - pexp(1 / 3, rate = lambda_per_hour)

# או לפי דקה
1 - pexp(20, rate = 4 / (8 * 60))
```

סעיף ג'

כל ניתוק מזוהה על ידי צוות המעבדה בהסתברות 0.7, באופן בלתי תלוי בניתוקים אחרים. באחד מ**ימי העבודה המלאים** (8 שעות) היו 6 ניתוקים. מה ההסתברות שזאת שעות המעבדה יזהה **לפחות** 2 מתוך הניתוקים?

נשים לב שאמרו לנו לפחות, ולכן יותר קל לחשב הסתברות משלימה של שני המאורעות 0 ו-1

נסמן Y = מספר הניתוקים שזוהו מתוך 6.

כל ניתוק הוא הצלחה/כישלון עם הסתברות הצלחה $p = 0.7$, לכן:

$$Y \sim \text{Binomial}\ (n = 6, p = 0.7)$$

צריך למצוא את:

$$P(Y \geq 2)$$

נלך על המשלים..

$$P(Y \geq 2) = 1 - P(Y = 0) - P(Y = 1)$$

נחשב:

$$P(Y = 0) = \binom{6}{0}(0.7)^0(0.3)^6 = 0.3^6 = 0.000729$$

$$P(Y = 1) = \binom{6}{1}(0.7)^1(0.3)^5 = 6 \cdot 0.7 \cdot 0.3^5 = 0.010206$$

$$P(Y \geq 2) = 1 - (0.000729 + 0.010206) = 0.989065 \approx 0.9891$$

כלומר ההסתברות לזהות לפחות שני ניתוקים היא כ-98.9%.

קוד R לאימות

```r
pbinom(1, size = 6, prob = 0.7, lower.tail = FALSE)
```

סעיף ד'

בגלל שזיהוי ניתוקי חשמלי בזמן שומר על ציוד המעבדה, החוקרת הראשית החליטה לנתמל חברי מעבדה שמחזיקים ניתוקים. על כל זיהוי מוצלח של ניתוק חברי המעבדה שזיהו אותו יקבלו 25 שקלים. יותר מחבר מעבדה אחד יכול לזכות בגמול על זיהוי ניתוק, בנוסף, אין תלות בין יכולת הזיהוי של חברי המעבדה השונים. **יוסי** טוב בזיהוי ניתוקים, ההסתברות שלו לזהות ניתוק היא 0.8. **דין** פחות טוב בזיהוי ניתוקים, ההסתברות שלו לזהות ניתוק היא 0.5. מה **התוחלת ושונות הגמול של יוסי ושל דין** (של כל אחד בנפרד) ביום בו התרחשו 6 ניתוקים? חשבו והסבירו את אופן החישוב ואת משמעות התוצאות שהתקבלו.

נסמן לכל אדם:

$$X \sim \text{Binomial } (n = 6, p)$$

כאשר X הוא מספר הזיהויים המוצלחים שלו.

הגמול הכספי:

$$Reward = 25X$$

מהדף נוסחאות, תוחלת ושונות התפלגות בינומית

$$E[X] = n \cdot p, \qquad \text{Var } (X) = n \cdot p \cdot q$$

הגמול הוא סהכ טרנס' לינארית (כמו בשאלה של מעדני יולו ממועד א 2020) $R = 25X$:

$$E[R] = 25E[X] = 25np$$

$$\text{Var } (R) = 25^2 \text{ Var } (X) = 25^2 \cdot n \cdot p \cdot q$$

יוסי ($p = 0.8$)

$$E[R_Y] = 25 \cdot 6 \cdot 0.8 = 120$$

$$\text{Var } (R_Y) = 25^2 \cdot 6 \cdot 0.8 \cdot 0.2 = 625 \cdot 0.96 = 600$$

דין ($p = 0.5$)

$$E[R_D] = 25 \cdot 6 \cdot 0.5 = 75$$

$$\text{Var } (R_D) = 25^2 \cdot 6 \cdot 0.5 \cdot 0.5 = 625 \cdot 1.5 = 937.5$$

משמעות:

- ליוסי תוחלת גמול גבוהה יותר (כלומר ירוויח יותר).
- לדין שונות גבוהה יותר, כלומר הגמול שלו תנודתי יותר.

קוד R לאימות

```
n <- 6
money <- 25

# יוסי
p_yossi <- 0.8
e_yossi <- money * (n * p_yossi)
v_yossi <- (money^2) * (n * p_yossi * (1 - p_yossi))

# דין
p_dan <- 0.5
e_dan <- money * (n * p_dan)
v_dan <- (money^2) * (n * p_dan * (1 - p_dan))

tibble::tibble(
  statistic = c("Expected", "Variance"),
  Yossi = c(e_yossi, v_yossi),
```

```
  Dan = c(e_dan, v_dan)
)
```

סעיף ה'

נתון הקוד הבא:

```
time_of_disconect <- seq(from = 0, to = 4, length = 1000)
rate <- 4
probabilty <- pexp(time_of_disconect, rate)

plot(time_of_disconect, probabilty)
abline(v = qexp(0.7, rate))
```

1. הסבירו סכמטית מה מצויר בגרף ומה משמעות הקו האנכי.
2. הסבירו מה הקוד עושה.

שירטוט של התפלגות מעריכית עם קו אנכי בנקודה שעד אליה יש 70% שהאירוע התרחש (לפי הקצב שהוגדר)

- ציר x: זמן עד ניתוק (ביחידות הזמן של המודל).
- ציר y: הסתברות מצטברת $P(T \leq t)$.
- הגרף עולה באופן מונוטוני מאפס ל1

הקו האנכי בשורה האחרונה הוא ב-

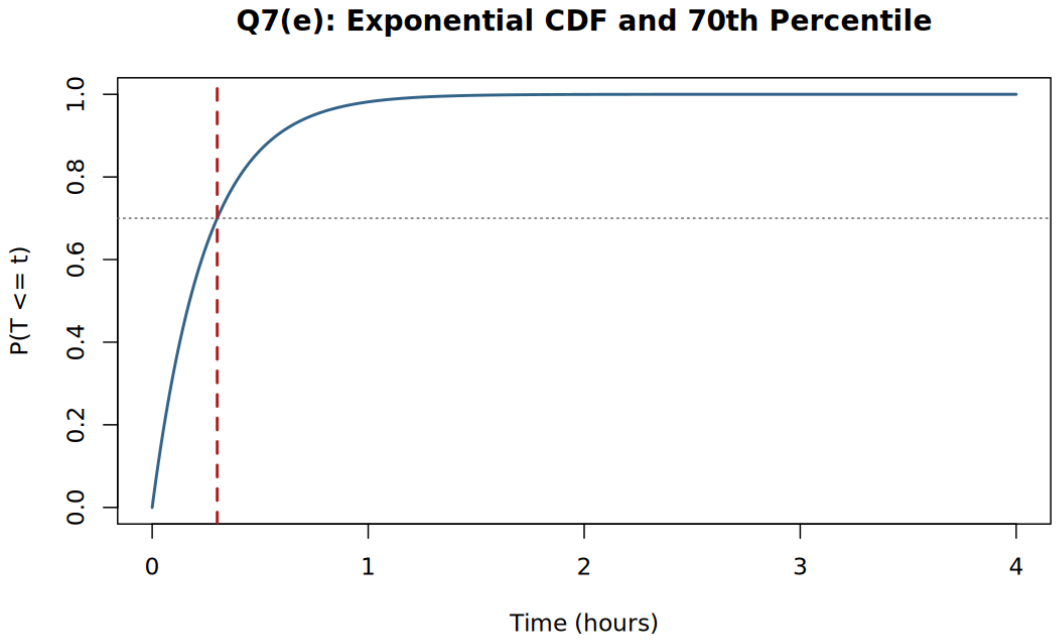
$$t_{0.7} = qexp(0.7, 4)$$

כלומר האחוזון ה-70 של זמן ההמתנה, כך שמתקיים במילים: עד זמן זה יש הסתברות של 70% שכבר קרה ניתוק.

הסבר שורה-שורה של הקוד:

1. `seq(from = 0, to = 4, length = 1000)` יוצר 1000 נקודות זמן בין 0 ל-4.
2. `rate <- 4` מגדיר את פרמטר הקצב λ.
3. `pexp(...)` מחשב לכל נקודת זמן את $P(T \leq t)$.
4. `plot(...)` מצייר את פונקציית ההתפלגות המצטברת.
5. `abline(v = qexp(0.7, rate))` מוסיף קו אנכי במיקום האחוזון ה-70.

$$qexp(0.7, 4) \approx 0.301$$



תרשים 1: זה לא הגרף שייצא ממה שהם כתבו...

שאלה 8

מירב כותבת את עבודת הגמר שלה על התפתחות מוסר בקרב ילדים בגילאי בית הספר היסודי. היא בנתה ניסוי, שבו השתתפו תלמידים מכיתות א, ו-ב. בתחילת הניסוי, היא בדקה כמה סוכריות כל תלמיד/ה לוקח/ת. לאחר מכן, היא מדדה את מספר הסוכריות שתלמידים חילקו לחבריהם לכיתה. מירב שיערה שיהיה קשר בין מספר הסוכריות שהתלמידים לקחו לבין מספר הסוכריות שהם יחלקו.

להלן התוצאות:

טבלה לסעיפים א-ג

מספר ילד/ה	מגדר	כיתה	סוכריות שלקח/ה	סוכריות שחילק/ה
1	ילד	א	4	1
2	ילד	א	5	2
3	ילדה	ב	5	5
4	ילדה	א	9	4
5	ילדה	ב	10	5
6	ילד	ב	3	1

סעיף א'

התוצאות של נבדקת מספר 3 הפתיעו את מירב. מה ניתן להגיד על המיקום היחסי של מספר הסוכריות שנבדקת מספר 3 לקחה ועל המיקום היחסי של מספר הסוכריות שהיא חילקה? חשבו והשוו בעזרת ציון Z. (10 נקודות)

נחשב ידנית את הממוצע וסטיית התקן לכל משתנה, ואז את ציון ה-Z.

$$z = \frac{x - \bar{x}}{s}$$

שלב 1: עבור "סוכריות שלקח/ה"

הנתונים: $4, 5, 5, 9, 10, 3$, ולכן:

$$\bar{x} = \frac{4+5+5+9+10+3}{6} = \frac{36}{6} = 6$$

סטיית תקן:

$$s_x = \sqrt{\frac{\sum (x_i - \bar{x})^2}{n}} = \sqrt{\frac{(4-6)^2 + (5-6)^2 + (5-6)^2 + (9-6)^2 + (10-6)^2 + (3-6)^2}{6}}$$

$$= \sqrt{\frac{4+1+1+9+16+9}{6}} = \sqrt{\frac{40}{6}} \approx 2.582$$

הילדה השלישית לקחה 5 סוכריות

$$x_3 = 5$$

עבור "לקחה":

$$z_{\text{took}} = \frac{5-6}{2.582} \approx -0.387$$

שלב 2: עבור "סוכריות שחילק/ה"

הנתונים: $1, 2, 5, 4, 5, 1$, ולכן:

$$\bar{y} = \frac{1+2+5+4+5+1}{6} = \frac{18}{6} = 3$$

סטיית תקן:

$$s_y = \sqrt{\frac{\sum (y_i - \bar{y})^2}{n}} = \sqrt{\frac{(1-3)^2 + (2-3)^2 + (5-3)^2 + (4-3)^2 + (5-3)^2 + (1-3)^2}{6}}$$

$$= \sqrt{\frac{4+1+4+1+4+4}{6}} = \sqrt{\frac{18}{6}} = \sqrt{3} \approx 1.732$$

אותה ילדה גם חילקה 5 סוכריות.

$$y_3 = 5$$

עבור "חילקה":

$$z_{\text{gave}} = \frac{5-3}{1.732} \approx 1.155$$

תשובה מילולית לילדה השלישית (במיקום 3) יש ציון תקן שלילי במשתנה "לקחה" (מתחת לממוצע) וציון תקן חיובי במשתנה "חילקה" (מעל הממוצע). היא לקחה יחסית מעט, אבל חילקה יחסית הרבה. ילדה טובה!

קוד R לאימות

```
library(tidyverse)

# נתונים לסעיפים א-ג
dat6 <- tibble(
  id = 1:6,
  gender = c("boy", "boy", "girl", "girl", "girl", "boy"),
  class = c("A", "A", "B", "A", "B", "B"),
  took = c(4, 5, 5, 9, 10, 3),
  gave = c(1, 2, 5, 4, 5, 1)
)

z_took_3 <- as.numeric(scale(dat6$took)[3])
z_gave_3 <- as.numeric(scale(dat6$gave)[3])

tibble(
  measure = c("z_took_3", "z_gave_3"),
  value = c(z_took_3, z_gave_3)
)
```

סעיף ב'

חשבו את מדד הקשר המתאים לתיאור הקשר בין מספר סוכריות שלקחו ומספר הסוכריות שחילקו. נמקו את בחירה במדד הקשר והסבירו את התוצאה שהתקבלה. (10 נקודות)

שני המשתנים כמותיים, לכן משתמשים במתאם פירסון:

$$r = \frac{\sum_{i=1}^{n}\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)}{n\, s_x s_y}$$

מסעיף א' כבר קיבלנו:

$$\bar{x} = 6, \quad \bar{y} = 3, \quad s_x = \sqrt{\frac{40}{6}} \approx 2.582, \quad s_y = \sqrt{\frac{18}{6}} \approx 1.732, \quad n = 6$$

נחשב את המכפלה $\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right)$ לכל נבדק:

i	x_i	y_i	$x_i - \bar{x}$	$y_i - \bar{y}$	מכפלה
1	4	1	-2	-2	4
2	5	2	-1	-1	1
3	5	5	-1	2	-2
4	9	4	3	1	3
5	10	5	4	2	8
6	3	1	-3	-2	6

ולכן:

$$\sum\left(x_i - \bar{x}\right)\left(y_i - \bar{y}\right) = 4 + 1 - 2 + 3 + 8 + 6 = 20$$

נציב בנוסחה:

$$r = \frac{20}{6 \cdot 2.582 \cdot 1.732} = \frac{20}{26.833} \approx 0.745$$

תשובה מילולית המתאם חיובי ובגודל בינוני-חזק. כלומר, ככל שבאופן יחסי ילד/ה לוקח/ת יותר סוכריות, כך יש נטייה לחלק יותר סוכריות.

קוד R לאימות

```
r_pearson <- cor(dat6$took, dat6$gave, method = "pearson")
r_pearson
```

סעיף ג'

מירב גילתה שהיא בטעות פספסה ילד אחד. אלה הנתונים של הילד שלא הוכנו בטבלה הקודמת:

מספר ילד/ה	מגדר	כיתה	סוכריות שלקח/ה	סוכריות שחילק/ה
7	ילד	ב	6	6

כיצד הוספה של נתוני הילד הנוסף (שמופיעים בטבלה המצורפת) למדגם תשפיע על סטיית התקן של מספר הסוכריות שלקחו הילדים? הסבירו את התשובה. (5 נקודות)

נסמן את נתוני ה"לקח/ה" המקוריים:

$$4, 5, 5, 9, 10, 3$$

כבר חישבנו:

$$\bar{x} = 6, \qquad \sum\left(x_i - \bar{x}\right)^2 = 40$$

לכן לפני ההוספה:

$$s_{\text{before}} = \sqrt{\frac{40}{6}} \approx 2.582$$

כעת מוסיפים את הילד הנוסף עם ערך $x_7 = 6$:

$$4, 5, 5, 9, 10, 3, 6$$

הממוצע נשאר 6, כי הוספנו בדיוק את הממוצע הקודם.
גם סכום ריבועי הסטיות נשאר 40, כי התרומה של הערך החדש היא:

$$(6-6)^2 = 0$$

אבל עכשיו $n = 7$, ולכן מחלקים ב-$n = 7$:

$$s_{\text{after}} = \sqrt{\frac{40}{7}} \approx 2.390$$

תשובה מילולית סטיית התקן **קטנה**. המונה לא השתנה אבל המכנה גדל מ6 ל7.

קוד R לאימות

```
sd_before <- sd(dat6$took)
sd_after <- sd(c(dat6$took, 6))

tibble(
  stage = c("before adding child 7", "after adding child 7"),
  sd_took = c(sd_before, sd_after)
)
```

סעיף ד'

מירב המשיכה לאסוף נתונים עד שהגיעה לטבלה הסופית (בסעיף זה התייחסו רק לטבלה הבאה, לא לנתונים הקודמים). האם יש תלות בין מגדר הילד/ה והכיתה בה הם לומדים? חשבו את מדד הקשר המתאים והסבירו את משמעות ערכו. (5 נקודות)

מספר ילד/ה	מגדר	כיתה	סוכריות שלקח/ה	סוכריות שחילק/ה
1	ילד	א	4	1
2	ילד	א	5	2
3	ילדה	ב	5	5
4	ילדה	א	9	4
5	ילדה	ב	10	5
6	ילד	ב	3	1
7	ילד	ב	6	6
8	ילדה	ב	10	6
9	ילד	א	9	5
10	ילד	א	5	2
11	ילד	א	6	5
12	ילדה	ב	10	9

שני המשתנים שמיים (קטגוריאלים), ולכן מדד מתאים הוא $Cramer's\ V$:

$$V = \sqrt{\frac{\chi^2}{n(k-1)}}, \qquad k = \min(r, c)$$

1) טבלת שכיחויות נצפית (O)

מגדר כיתה	א (1)	ב (2)	סה"כ (R_i)
ילד (1)	5	2	7
ילדה (2)	1	4	5
סה"כ (C_j)	6	6	12

2) טבלת שכיחויות צפויה (E) תחת אי-תלות

$$E_{ij} = \frac{R_i \cdot C_j}{n}$$

נחשב:

$$E_{11} = \frac{7 \cdot 6}{12} = 3.5, \quad E_{12} = \frac{7 \cdot 6}{12} = 3.5$$

$$E_{21} = \frac{5 \cdot 6}{12} = 2.5, \quad E_{22} = \frac{5 \cdot 6}{12} = 2.5$$

מגדר כיתה	א	ב
ילד	3.5	3.5
ילדה	2.5	2.5

3) חישוב χ^2

$$\chi^2 = \sum \frac{(O-E)^2}{E}$$

$$= \frac{(5-3.5)^2}{3.5} + \frac{(2-3.5)^2}{3.5} + \frac{(1-2.5)^2}{2.5} + \frac{(4-2.5)^2}{2.5}$$

$$= 0.6429 + 0.6429 + 0.9 + 0.9 = 3.0858 \approx 3.086$$

4) $Cramer'sV$

כאן $c = r = 2$, 2, לכן:

$$k = \min(2, 2) = 2, \qquad n = 12$$

$$V = \sqrt{\frac{3.0858}{12(2-1)}} = \sqrt{0.25715} \approx 0.507$$

תשובה מילולית יש תלות בעוצמה בינונית בין מגדר לכיתה. במדגם הזה יש ייצוג גבוה יותר של בנים בכיתה א' ושל בנות בכיתה ב', יחסית למה שהיינו מצפים תחת אי-תלות.

קוד R לאימות

```
dat12 <- tibble(
  id = 1:12,
   gender = c("boy", "boy", "girl", "girl", "girl", "boy", "boy", "girl",
"boy", "boy", "boy", "girl"),
  class = c("A", "A", "B", "A", "B", "B", "B", "B", "A", "A", "A", "B"),
  took = c(4, 5, 5, 9, 10, 3, 6, 10, 9, 5, 6, 10),
  gave = c(1, 2, 5, 4, 5, 1, 6, 6, 5, 2, 5, 9)
)

tab_gender_class <- table(dat12$gender, dat12$class)
tab_gender_class

chisq_out <- suppressWarnings(chisq.test(tab_gender_class, correct = FALSE))
chi2 <- as.numeric(chisq_out$statistic)
n <- sum(tab_gender_class)
k <- min(nrow(tab_gender_class), ncol(tab_gender_class))
cramer_v <- sqrt(chi2 / (n * (k - 1)))
cramer_v
```

בהצלחה בהמשך התואר :) יאן 0535290202