סטטיסטיקה לתלמידי הנדסת תעשיה וניהול

פתרון הבחינה - מועד א

חלק א – שאלות סגורות:

השתמשו במידע הבא כדי לענות על שאלות 3-1:
מרצה במכללה מתעניין במספר הימים הממוצע **שתלמידי הנדסה במכללה** נעדרים במהלך סמסטר.

1. מהי האוכלוסייה שבה מתעניין המרצה?
   א) כל הסטודנטים במכללה
   ב) כל הסטודנטים להנדסה בארץ
   ג) כל הסטודנטים שלו במכללה
   ד) כל הסטודנטים להנדסה במכללה
   ה) אף תשובה אינה נכונה

פתרון: התשובה הנכונה היא ד. השאלה מתייחסת בפירוש לתלמידי ההנדסה במכללה (ראו הדגשה שנוספה בגוף השאלה)

2. יהי M מספר הימים בסמסטר שבהם סטודנט להנדסה נעדר. במקרה זה, M הוא דוגמה ל:
   א) אומד
   ב) מִשְׁתַּנֶּה.
   ג) פרמטר
   ד) אף תשובה אינה נכונה

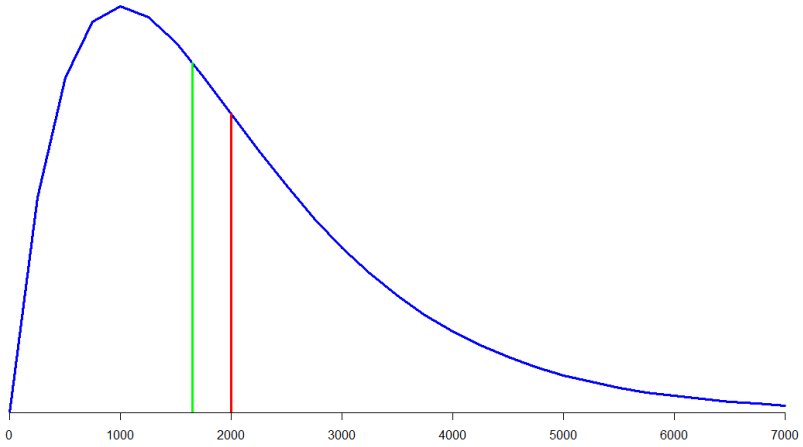
פתרון: התשובה הנכונה היא ב. M הוא נתון שעשוי לקבל ערכים שונים אצל סטודנטים שונים.

3. מדגם שלקח המרצה הניב ממוצע של 3.5 ימים. ערך זה הוא דוגמה ל:
   א) פָּרָמֶטֶר
   ב) משתנה
   ג) אומד (אומדן)
   ד) א+ב
   ה) ב+ג
   ו) אף תשובה אינה נכונה

פתרון: התשובה הנכונה היא ג. זהו ממוצע הערכים של M שהתקבלו במדגם על פני כל הסטודנטים.

4. התפלגות ההכנסה בחלק ממדינות העולם השלישי נחשבת אסימטרית ימנית. נניח שאנחנו בוחנים מדינת עולם שלישי, בה השכר הממוצע הוא 2,000 דולר לשנה עם סטיית תקן של 8,000 דולר. קבעו מה נכון:
   א. פחות מ-50% משתכרים יותר מ-2000 דולר
   ב. יותר מ-50% משתכרים יותר מ-2000 דולר
   ג. 50% משתכרים יותר מ-2000 דולר
   ד. לא ניתן לדעת האם אחוז המשתכרים יותר מ-2000 דולר נמוך מ-50%, שווה ל50% או גבוה מ-50%
   ה. אף תשובה אינה נכונה.

פתרון: התשובה הנכונה היא א.
בהתפלגות אסימטרית ימנית (התפלגות עם זנב ימני) הממוצע גדול מהחציון, מכיוון שיש ערכים גבוהים ש-"מושכים" את ערכו כלפי מעלה. מכאן שהחציון קטן מ-2000. ראו שרטוט: הממוצע מסומן בקו אדום, והחציון בקו ירוק:



אחוז האנשים ששכרם גבוה מהחציון מבוטא על ידי השטח מתחת לעקומה שמימין לקו הירוק והוא שווה ל-50%. אחוז האנשים ששכרם גבוה מהממוצע, מבוטא על ידי השטח שמתחת לעקומה מימין לקו האדום, ושטח זה קטן מהשטח שמימין לקו הירוק. לכן אחוז האנשים ששכרם גבוה הממוצע חייב להיות קטן מ-50%.

5. בעבר דיווח ארגון כי בני נוער בילו 4.5 שעות בשבוע, בממוצע, בסמארטפון. הארגון חושב שכרגע הממוצע גבוה יותר. ל- 15 בני נוער שנבחרו באופן אקראי מדדו כמה שעות בשבוע הם בילו בסמארטפון. ממוצע המדגם היה 4.75 שעות עם סטיית תקן מדגמית של 2.0 שעות.
השערת האפס והאלטרנטיבה הן:
   א) $H_0: \bar{x} = 4.5,\ H_1: \bar{x} > 4.5$
   ב) $H_0: \mu \geq 4.5,\ H_1: \mu < 4.5$
   ג) $H_0: \mu = 4.75,\ H_1: \mu > 4.75$
   ד) $H_0: \mu = 4.5,\ H_1: \mu > 4.5$
   ה) אף תשובה אינה נכונה

פתרון: התשובה הנכונה היא ד.
ראשית, ההשערות מנוסחות ביחס לתוחלת μ, ולכן תשובה א נפסלת מייד.
בעבר הממוצע היה 4.5, כלומר ההנחה הראשונית היא כי התוחלת שווה ל-4.5, וזה פוסל את תשובה ב, וגם את תשובה ג. כעת אנו משערים כי ערך זה עלה, כלומר כי עכשיו התוחלת גבוהה מ-4.5, וזו ההשערה האלטרנטיבית בתשובה ד.

6. כאשר נוצרת תרופה חדשה, על חברת התרופות להעמיד אותה לבדיקה לפני שתקבל את האישור הדרוש ממינהל המזון והתרופות (FDA) לשיווק התרופה. נניח שהשערת האפס היא "התרופה אינה בטוחה". מהי השגיאה מסוג 2?
   א) להסיק שהתרופה בטוחה כאשר למעשה היא אינה בטוחה.
   ב) לא להסיק שהתרופה בטוחה כשלמעשה היא בטוחה.
   ג) להסיק כי התרופה בטוחה כשלמעשה היא בטוחה.
   ד) לא להסיק שהתרופה אינה בטוחה כאשר, למעשה, היא אינה בטוחה.
   ה) אף תשובה אינה נכונה

פתרון: השערת האפס היא "התרופה אינה בטוחה", ומכאן שההשערה האלטרנטיבית היא "התרופה בטוחה". שגיאה/טעות מסוג שני מתרחשת רק כאשר ההשערה האלטרנטיבית נכונה, כלומר כאשר התרופה בטוחה.
במצב זה, אם לא דוחים את השערת האפס מסיקים כי התרופה אינה בטוחה, ובמילים אחרות, לא מסיקים כי התרופה בטוחה. אך יש להסיק כי התרופה בטוחה, וזאת מכיוון שאנו מניחים כי ההשערה האלטרנטיבית נכונה. לכן התשובה הנכונה היא ב.

7. מרצים חושבים שפחות מ -50% מתלמידי המכללה נכחו בהרצאות זום. לשם בדיקת ההשערה, לקחו מדגם של 84 סטודנטים בקורס מסוים, ובדקו כמה מהם נכחו בזום. התברר כי 36 מתוכם נכחו בזום. שגיאה מסוג I היא להסיק שאחוז תלמידי המכללה שנכחו הם:
   א) לפחות 50%, כאשר למעשה, הם פחות מ -50%
   ב) 50%, כאשר למעשה, הם 50%
   ג) פחות מ -50%, כאשר למעשה, הם לפחות 50%
   ד) פחות מ -50%, כאשר למעשה, הם פחות מ -50%
   ה) אף תשובה אינה נכונה

פתרון: זה מצב של בדיקת השערה על פרופורציה.
השערת האפס היא שלפחות 50% מהתלמידים נוכחים בהרצאות, והמרצים משערים את ההשערה האלטרנטיבית כי פחות מ-50% מהתלמידים נוכחים בהרצאות.
שגיאה/טעות מסוג ראשון מתרחשת כאשר דוחים את השערת האפס באופן מוטעה, כלומר כאשר השערת האפס נכונה, או במילים אחרות: מסיקים כי ההשערה האלטרנטיבית נכונה כאשר השערת האפס נכונה.
מכאן שטעות מסוג ראשון היא להסיק כי אחוז התלמידים שנכחו בהרצאות קטן מ-50% בעוד למעשה המצב הוא כי לפחות 50% נכחו בהרצאות. התשובה הנכונה היא ג.

8. משפחות מתבקשות להשתתף בסקר, בו הן נשאלות לגבי: (1) ההכנסה השנתית שלהן. (2) החיסכון השנתי שלהן. (3) אם הם ילידי הארץ. אם רוצים לראות באופן ויזואלי אם יש קשר בין ההכנסה השנתית והחיסכון השנתי של משפחה, מה יהיה הגרף המתאים ביותר?
   א. היסטוגרמה של ההכנסות לצד היסטוגרמה של החסכונות
   ב. תרשים פיזור (scatter plot, X-T plot)
   ג. side-by-side box plots (תרשימי קופסה נפרדים לילידי הארץ ולעולים)
   ד. תרשים עוגה של ההכנסות לצד תרשים עוגה של החסכונות
   ה. אף תשובה אינה נכונה

פתרון: יש לשים לב כי השאלה מתייחסת לשלושה אלמנטים: הכנסות, חסכונות, והמשפחה עצמה. התשובה הנכונה היא תרשים פיזור, בו באחד הצירים נמדדת ההכנסה של כל משפחה, ובציר האחר נמדד החיסכון של כל משפחה. עבור כל משפחה יש את שני הנתונים, ומסומנת עבורה נקודה בתרשים.

תשובות א, ו-ד אינן נכונות, מכיוון שכל אחת בדרכה משווה את התפלגות ההכנסות להתפלגות החסכונות, וגורם המשפחה לא נלקח בחשבון.
מצב דומה יש גם בשאלה ג, מאחר וגם שני תרשימי קופסה זה לצד זה משווים הם בין התפלגויות. מלבד זאת, ניתן היה לפסול מייד את תשובה ג משתי סיבות: היא התייחסה לסטטוס של המשפחות כילידי הארץ או עולים, והשאלה עצמה, לאחר ההקדמה, לא התייחסה לסטטוס זה. בנוסף, לא צויין כלל איזה משתנה או משתנים יתארו דיאגרמות הקופסה.

9. חוקרים מבקשים לבדוק השערה לפיה סטודנטים ישנים פחות מ-7 שעות שינה בלילה, בממוצע. סקר שנערך בקרב 150 סטודנטים הניב ממוצע של 6.24 שעות עם סטיית תקן של 1.93 שעות. מובהקות התוצאה (p-value) היא:
   a. $P(\bar{x} < 7)$
   b. $P(\mu < 6.24)$
   c. $P(\mathrm{Z} < -1.847)$
   d. $P(\mu < 7)$
   e. None of the above

פתרון:
ההשערות הן:

$$H_0: \mu = 7,\ \ H_1: \mu < 7$$

נדחה את השערת האפס כאשר ממוצע המדגם קטן מ-7 באופן מובהק, כלומר כאשר ממוצע המדגם קטן מ-7 פחות משהו, למשל כאשר $\bar{X} < 7 - z_\alpha \cdot \frac{\sigma}{\sqrt{n}}$, או, למי שלמד מבחן t, כאשר $\bar{X} < 7 - t_{\alpha,n-1} \cdot \frac{s}{\sqrt{n}}$. מבחינה מעשית, ההבדל בין ערך ה-z וערך ה-t זניח בגודל מדגם כזה, ולצורך חישוב ה-p-value אין כלל צורך לדעת מהו הערך שמתחתיו דוחים את השערת האפס.

על פי הגדרתו, ה-p-value הוא ההסתברות כי ממוצע המדגם יהיה קטן מ-6.24, וזאת תחת ההנחה כי השערת האפס נכונה.
תשובה a אינה נכונה מכיוון שזו ההסתברות כי ממוצע המדגם קטן מ-7.
בתשובה b אמנם מופיע המספר 6.24 אך ההסתברות אינה מתייחסת לממוצע המדגם אלא לתוחלת μ. למעשה, לביטוי זה אין משמעות מכיוון ש-μ אינו משתנה מקרי.
גם לביטוי בתשובה d אין משמעות, שוב מכיוון ש-μ אינו משתנה מקרי.

נבדוק אם תשובה c נכונה.
כאמור, ה-p-value הוא ההסתברות כי ממוצע המדגם יהיה קטן מ-6.24, וזאת תחת ההנחה כי השערת האפס נכונה.
נחשב:

$$P_{H_0}(\bar{x} < 6.24) =$$
$$P\left(\frac{\bar{x} - 7}{1.93/\sqrt{150}} < \frac{6.24 - 7}{1.93/\sqrt{150}}\right) =$$
$$P\left(\frac{\bar{x} - 7}{1.93/\sqrt{150}} < -4.82\right) =$$
$$P(Z < -4.82)$$

השוויון האחרון (המעבר בין השורה השלישית בחישוב לשורה הרביעית והאחרונה בחישוב) מתבסס על ההנחה כי גודל המדגם, 150, מספיק גדול. החישוב הזה מראה כי גם תשובה c אינה נכונה לכן התשובה הנכונה היא e, כלומר אף אחת מארבע התשובות הראשונות שהוצעו אינה נכונה.

ב-3 השאלות הבאות, ענו נכון / לא נכון:

10. עבור שני מדגמים משתי אוכלוסיות שונות, אם סטיית התקן של המדגם הראשון קטנה מסטיית התקן של המדגם השני, אז גם התחום הבין רבעוני של המדגם הראשון קטן מהתחום הבין רבעוני של המדגם השני

פתרון: הטענה אינה נכונה. להלן דוגמה:
מדגם ראשון: 1,2,4,5. סטיית התקן היא 1.83 (כאשר מחלקים ב-n-1), והתחום הבין רבעוני שווה ל-2.
מדגם שני: 1,2,3,200. כאן סטיית התקן שווה ל-99 (שוב, כאשר מחלקים ב-n-1), אבל התחום הבין רבעוני שווה ל-1.
סטיית התקן של המדגם הראשון קטנה מסטיית התקן של המדגם השני, אבל התחום הבין רבעוני של המדגם הראשון גדול מהתחום הבין רבעוני של המדגם השני.

11. חוקר מדד את אורכם של כלים פרה-היסטוריים שנמצאו באתר ארכיאולוגי בס"מ. לצורך פרסום התוצאות בארה"ב, היה עליו לתרגם את הנתונים לאינצ'ים, ולכן הכפיל כל נתון ב-2.54. לכן ממוצע האורכים והשונות שלהם הוכפלו גם הם ב-2.54.

פתרון: הטענה אינה נכונה. כאשר מכפילים נתונים בקבוע, הממוצע אכן מוכפל בקבוע, אבל השונות מוכפלת בריבוע של הקבוע, במקרה הזה ב-6.4516.

12. אם בודקים השערה ודוחים אותה ברמת מובהקות של 5%, אז דוחים אותה גם ברמת מובהקות של 2.5%.

פתרון: הטענה אינה נכונה. קל לראות את זה אם נעזרים ב-p-value. לדוגמה, אם ה-p-value שווה ל4%, אז הוא קטן מ-5% אבל גדול מ-2.5%. לכן דוחים את ההשערה ברמת מובהקות של 5%, אבל לא דוחים אותה גם ברמת מובהקות של 2.5%.
מעניין לציין כי הטענה ההפוכה נכונה: אם דוחים את ההשערה ברמת מובהקות של 2.5%, אז דוחים אותה גם ברמת מובהקות של 5%.

חלק ב' – שאלות פתוחות:

13. מכונה ממלאת בקבוקי חלב. הכמות הממוצעת של חלב בכל בקבוק אמורה להיות 950 מ"ל, עם סטיית תקן ידועה של 1.77 מ"ל. ידוע כי כמות החלב שהמכונה ממלאת בכל בקבוק מתפלגת נורמלית. כדי לבדוק אם המכונה פועלת כראוי (כלומר התוחלת היא אכן 950), ייבחרו 36 בקבוקים מלאים באופן אקראי והכמות הממוצעת תימדד.
    א. מהן ההשערות הנבדקות?
    ב. מהן המשמעויות של טעות מסוג ראשון וטעות מסוג שני כאן?
    ג. אם נבחן את ההשערה ברמת מובהקות 5%, מה צריך להיות ממוצע המדגם כדי שנדחה את השערת האפס (כלומר מהו איזור הדחיה)?
    ד. אם יש תקלה במכונה והתוחלת היא 945 מ"ל, מה העצמה של המבחן הסטטיסטי שמצאתם בסעיף ג לאתר תקלה זו? (הציגו את הדרך, אין צורך לקרוא ערכים מטבלה סטטיסטית)

פתרון:

א. השערת האפס היא כי המכונה פועלת כראוי, כלומר התוחלת היא 950 מ"ל (הממוצע שאמור להיות), וההשערת האלטרנטיבית היא שהמכונה אינה פועלת כראוי כלומר התוחלת שונה מ-950:

$$H_0: \mu = 950, \quad H_1: \mu \neq 950$$

ב. טעות מסוג ראשון היא דחיה מוטעית של השערת האפס, כלומר הסקה כי המכונה אינה פועלת כראוי כאשר היא כן פועלת כראוי.

טעות מסוג שני מתרחשת כאשר השערת האפס אינה נכונה, ולא דוחים אותה. המשמעות היא כי נסיק שהמכונה פועלת כראוי למרות שאינה פועלת כראוי.

ג. תשובה זו מתבססת על כך שסטיית התקן ידועה.
אנו נדחה את השערת האפס כאשר ממוצע המדגם שונה באופן מובהק מ-950, כלומר כאשר

$$\bar{X} > \mu_0 + z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 950 + 1.96 \cdot \frac{1.77}{\sqrt{36}} = 950.58$$

או כאשר

$$\bar{X} < \mu_0 - z_{\alpha/2} \cdot \frac{\sigma}{\sqrt{n}} = 950 - 1.96 \cdot \frac{1.77}{\sqrt{36}} = 949.42$$

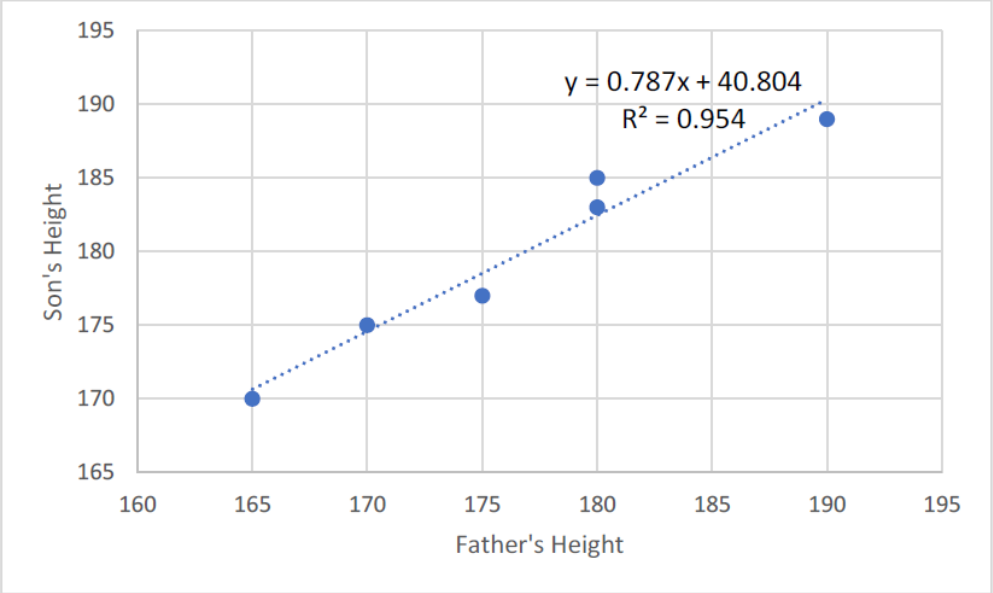
הבקשה כאן היא לחשב את העצמה של המבחן אם תחת ההשערה האלטרנטיבית התוחלת היא 945. העצמה היא ההסתברות לדחות את השערת האפס תחת ההנחה כי ההשערה האלטרנטיבית נכונה.
שימו לב כי מכיוון שזו השערה דו-צדדית, עלינו לחשב שתי הסתברויות!

$$P_{H_1}(\bar{X} < 949.42) = P\left(\frac{\bar{X}-945}{1.93/\sqrt{36}} < \frac{949.42-945}{1.93/\sqrt{36}}\right) = P(Z < 13.74)$$

$$P_{H_1}(\bar{X} > 950.58) = P\left(\frac{\bar{X}-945}{1.93/\sqrt{36}} > \frac{950.58-945}{1.93/\sqrt{36}}\right) = P(Z > 17.35)$$

לסיכום: העצמה שווה ל- $P(Z < 13.74) + P(Z > 17.35)$

14. נמדדו הגבהים של 6 זוגות של אבות ובנים – כל האבות ילידי שנת 1970, וכל הבנים ילידי שנת 2000. זאת על מנת למצוא מודל רגרסיה ליניארית לחיזוי גובה הבן בהתבסס על גובה אביו. מודל הרגרסיה שהתקבל הוא:



א. גובהו של אב הוא 188 ס"מ. מה היית מצפה שגובה בנו יהיה בהתבסס על המודל הזה?
ב. הפרש הגובה בין שני אבות הוא 10 ס"מ. מהו ההפרש בניבוי הגובה של שני הבנים שלהם?
ג. גובהו של אחד האבות הוא 180 ס"מ וגובה בנו הוא 185 ס"מ. מהו ערך השארית (השגיאה)?
ד. כעת רוצים לבדוק השערה שילידי שנת 2000 גבוהים מילידי שנת 1970. כיצד תשתמשו בנתונים כדי לבדוק השערה זו? כתבו בשורה אחת או שתיים מהן ההשערות הנבדקות, מהו המבחן הסטטיסטי המתאים ומדוע.

פתרון:

א. נציב את גובה האב במשוואת הרגרסיה ונקבל את החיזוי/ניבוי/הערך הצפוי לגובה הבן:

$$\text{expected son height} = 0.787 \cdot 188 + 40.804 = 188.76$$

ב. ההפרש הוא $10 \cdot b = 10 \cdot 0.787 = 7.87$

מי שקשה לו לראות זאת יכול לקחת דוגמה, למשל שני אבות שגובה אחד מהם הוא 170 ס"מ וגובה השני הוא 180 ס"מ, ולחשב את הגבהים הצפויים/נחזים של שני הבנים:

$$y_1 = 0.787 \cdot 180 + 40.804$$
$$y_2 = 0.787 \cdot 170 + 40.804$$

ואז ניתן לראות כי לא משנה מהם הגבהים של האבות אלא רק מה הפרש הגבהים:

$$\begin{aligned} y_1 - y_2 &= (0.787 \cdot 180 + 40.804) - (0.787 \cdot 170 + 40.804) \\ &= 0.787 \cdot 180 + 40.804 - 0.787 \cdot 180 - 40.804 \\ &= 0.787 \cdot 180 - 0.787 \cdot 170 = 0.787 \cdot (180 - 170) \\ &= 0.787 \cdot 10 \end{aligned}$$

ג. השארית היא ההפרש בין הערך האמיתי של y, כלומר של גובה הבן, ובין הניבוי של גובה הבן לפי גובה האב.
נתון כי גובה האב הוא 180 ס"מ. הניבוי לגובה הבן על פי גובה האב הוא

$$\hat{y} = 0.787 \cdot 180 + 40.804 = 182.5$$

גובה הבן הוא 185 לכן השארית שווה ל-

$$y - \hat{y} = 185 - 182.4 = -2.5$$

ד. נסמן את תוחלת הגובה של ילידי 1970 ב-μ_1 ואת תוחלת הגובה שלי ילידי שנת 2000 ב-μ_2. השערת האפס היא כי אין הבדל בין תוחלות הדבהים, בעוד שההשערה האלטרנטיבית היא כי תוחלת הגובה של ילידי שנת 2000 גדולה מתוחלת הגובה של ילידי שנת 1970:

$$H_0: \mu_1 = \mu_2 \quad H_1: \mu_1 < \mu_2$$

אבל: יש לשים לב כי הנתונים הם נתוני גובה של אבות ובנים ולא מדובר כאן בשתי אוכלוסיות בלתי תלויות אלא בזוגות של נתונים: גובה האב וגובה הבן. לכן יש להשתמש כאן במבחן למדגמים מזווגים.