תרגול מס׳ 9
מבחני χ^2

מבחני חי בריבוע הם קבוצה שימושית של מבחנים המאפשרים לבצע בדיקת השערות לגבי:

1. **טיב ההתאמה של נתונים להתפלגות תיאורטית (Goodness Of Fit)**
2. **אי-תלות**

המבחנים משתמשים בסטטיסטי המתפלג בקירוב חי בריבוע, ומכאן שמם.

<u>**מבחן חי בריבוע לטיב התאמה**</u>

בודק האם לפי נתוני המדגם סביר שהאוכלוסיה מתפלגת בהתפלגות מסוימת.

השערת האפס: האוכלוסייה מתפלגת בהתפלגות תיאורטית כלשהי (נורמלית, פואסונית וכו׳).

השערת האלטרנטיבה: ההתפלגות האמיתית של האוכלוסייה אחרת.

שימו לב שבמקרה זה, האינפורמציה החדשה (=ההתפלגות) ״תאושר״ ע״י קבלת H_0!

הרעיון הכללי: מחלקים את הערכיים האפשריים שיכולים להתקבל עפ״י ההתפלגות התיאורטית לקבוצות, ובודקים אם הפיזור לקבוצות עפ״י תוצאות המדגם מספיק קרוב לפיזור הצפוי.

<u>**שלבי העבודה**</u>

(1) מוודאים שהפרמטרים של ההתפלגות התיאורטית ידועים. אם לא, יש לאמוד אותם.

(2) יש לנסח בבירור את ההשערות הנבדקות.

(3) מחלקים את נתוני המדגם לקבוצות **(מס׳ הקבוצות הסופי יסומן ב-$\boldsymbol{k}$)**:

- נסמן: E_i – **תוחלת מספר הפריטים הצפויים** בקבוצה i, כלומר $\boldsymbol{E_i = n \cdot p_i}$.
- אם נתונה חלוקה לקבוצות, מוודאים שמתקיים $E_i \geq 5$ בכל קבוצה. אם יש קבוצה שלא מקיימת תנאי זה, מאחדים אותה עם קבוצה סמוכה.
- אם לא נתונה חלוקה לקבוצות, יוצרים אותה בעצמנו כך שכל הערכים האפשריים של ההתפלגות יכוסו, ובנוסף יתקיים בכל קבוצה i: $E_i \geq 5$.

(4) מס׳ הפריטים בכל קבוצה כפי שהתקבלו **בפועל במדגם** יסומן ב- $\boldsymbol{O_i}$.

(5) חישוב סטטיסטי המבחן: $\chi^2_{emp} = \sum_{i=1}^{k} \frac{(E_i - O_i)^2}{E_i}$.

כאשר מתקיים התנאי $E_i \geq 5$, הסטטיסטי מתפלג בקירוב χ^2 עם $(k-p-1)$ דרגות חופש, כאשר p הוא מס׳ הפרמטרים שנאמדו (אם לא נאמדו כלל פרמטרים, $p = 0$).

(6) כלל ההכרעה בר״מ α: דחה את H_0 אם מתקיים $\chi^2_{emp} > \chi^{2\,(k-p-1)}_{1-\alpha}$

<u>מבחן חי בריבוע לאי-תלות</u>

בודק אם קיימת תלות בין שני משתנים **קטגוריאליים**.

המשמעות של אי-תלות: ההתפלגות של אחד המשתנים זהה תחת כל קטגוריה של המשתנה השני.

נתונים שני משתנים קטגוריאליים: X (בעל r קטגוריות) ו-Y (בעל c קטגוריות).

O_{ij} – מספר התצפיות במדגם שמאופיינות ע"י קטגוריה i במשתנה X וקטגוריה j במשתנה Y.

בונים לוח שכיחויות, ובשולי הלוח מציגים את השכיחויות השוליות של כל משתנה.

X, Y	גבוהים	נמוכים	סה"כ
מהנדסים	מהנדסים גבוהים	מהנדסים נמוכים	סך כל המהנדסים במדגם
פסיכולוגים	פסיכולוגים גבוהים	פסיכולוגים נמוכים	סך כל הפסיכולוגים במדגם
סה"כ	סך כל הגבוהים במדגם	סך כל הנמוכים במדגם	סך כל האנשים שנדגמו

באופן כללי:

X, Y	Y_1	Y_2	…	Y_j	…	Y_c	סה"כ
X_1	O_{11}	O_{12}	…	O_{1j}	…	O_{1c}	$f_{1\bullet} \equiv \sum_{j=1}^{j=c} O_{1j}$
X_2	O_{21}	O_{22}	…	…	…	…	$f_{2\bullet}$
…	…	…	…	…	…	…	…
X_i	O_{i1}	…	…	O_{ij}	…	O_{ic}	…
…	…	…	…	…	…	…	…
X_r	O_{r1}	…	…	O_{rj}	…	O_{rc}	$f_{r\bullet}$
סה"כ	$f_{\bullet 1} = \sum_{i=1}^{r} O_{i1}$	$f_{\bullet 2}$	…	…	…	$f_{\bullet c}$	n

מערכת ההשערות:

H_0: $P(X_i \cap Y_j) = P(X_i) \cdot P(Y_j)$ $\Leftrightarrow$ לא קיימת תלות בין משתנה השורות למשתנה העמודות

H_1: המשתנים תלויים = אחרת

סטטיסטי המבחן לאי תלות: $\chi^2_{emp} = \sum_{i=1}^{r}\sum_{j=1}^{c}\frac{(O_{ij}-E_{ij})^2}{E_{ij}}$ **כאשר:** $E_{ij} = \frac{f_{i\bullet} \times f_{\bullet j}}{n}$

ע"מ להשתמש במבחן נדרש שבכל תא יתקיים $E_{ij} \geq 5$.

כלל ההחלטה ברמת מובהקות α: דחה את H_0 אם $\chi^2_{emp} > \chi^2_{1-\alpha,[(r-1)\cdot(c-1)]}$

<u>שאלה 1</u>

בכדי לקבל רשיון נהיגה אדם ניגש לטסטים עד שהוא עובר אחד בהצלחה. הדעה הרווחת היא שההסתברות לעבור טסט בהצלחה היא 0.6 (ללא תלות בין הטסטים השונים של אותו פרט). להלן טבלה המתארת את מספר הטסטים של מדגם מקרי של נהגים בעלי רשיון:

מס׳ טסטים	1	2	3	4	5 ויותר
מספר נהגים	58	18	13	9	2

האם על פי נתוני המדגם ניתן לקבל את הדעה הרווחת ברמת מובהקות 0.01? נמק.

<u>שאלה 1</u>

נסמן X מספר הטסטים **עד** להצלחה.

H_0: X~G (0.6), H_1: else

n=100

<u>הנתונים מחולקים לקבוצות:</u>

מס׳ טסטים	1	2	3	4	5 ויותר
O_i	58	18	13	9	2
$E_i = n \cdot p_i$	$100 \cdot 0.6 = 60$	$100 \cdot 0.4 \cdot 0.6 = 24$	$100 \cdot 0.4^2 \cdot 0.6 = 9.6$	$100 \cdot 0.4^3 \cdot 0.6 = 3.84$	$100 \cdot 0.4^4 = 2.56$

E_i בשתי הקבוצות האחרונות קטן מ-5. נאחד אותן לקבוצה אחת:

מס׳ טסטים	1	2	3	4 ויותר
O_i מעודכן	58	18	13	11
E_i מעודכן	$100 \cdot 0.6 = 60$	$100 \cdot 0.4 \cdot 0.6 = 24$	$100 \cdot 0.4^2 \cdot 0.6 = 9.6$	$100 \cdot 0.4^3 \cdot 0.6 + 100 \cdot 0.4^4 = 6.4$
$\frac{(E_i - O_i)^2}{E_i}$	0.066	1.5	1.204	3.306

$$\chi^2_{emp} = \sum \frac{(E_i - O_i)^2}{E_i} = 6.607$$

ערך טבלאי: $\chi^2_{0.99}(3) = 11.345$ הערך הטבלאי גדול מסטטיסטי המבחן ולכן לא נדחה את השערת האפס כלומר ניתן לקבל את הדעה הרווחת ברמת מובהקות 0.01.

<u>שאלה 2</u>

בכל שנה בודקת המועצה לצרכנות 500 מוצרים המהווים מדגם מייצג של אוכלוסיית המוצרים. המועצה עורכת בדיקות איכות ומחיר. על מנת לאפשר השוואה לאורך שנים מתבצעת הבדיקה ב-3 רמות איכות (גבוהה, בינונית ונמוכה) ו-2 רמות מחיר (גבוה, נמוך). לאחר ביצוע סקר השנה נמצאה התפלגות השכיחויות הבאה :

איכות/מחיר	גבוה	נמוך
גבוהה	110	80
בינונית	85	70
נמוכה	55	100

לאור התוצאות האם ניתן לומר ברמת מובהקות 0.05 כי השנה אין קשר בין איכות ומחיר? נמק. **(לא)**

<u>שאלה 2</u>

מדובר במבחן לאי תלות.

איכות/מחיר	גבוה	נמוך	סה״כ
גבוהה	110	80	**190**
בינונית	85	70	**155**
נמוכה	55	100	**155**
סה״כ	**250**	**250**	**500**

$$E_{11} = E_{12} = \frac{190 \cdot 250}{500} = 95, \qquad E_{21} = E_{22} = E_{31} = E_{32} = \frac{155 \cdot 250}{500} = 77.5$$

$$\chi^2 = \frac{(80-95)^2}{95} + \frac{(110-95)^2}{95} + \frac{(70-77.5)^2}{77.5}$$

$$+ \frac{(85-77.5)^2}{77.5} + \frac{(100-77.5)^2}{77.5} + \frac{(55-77.5)^2}{77.5} = 19.25$$

$$\chi^2 = 19.25 > \chi^2_{0.95}(1 \cdot 2) = 5.991$$

ולכן נדחה בר״מ 0.05 כלומר יש קשר בין איכות למחיר.

<u>שאלה 3</u>

חברת מוסיקה הוציאה אוסף של 4 תקליטורים המכילים את מיטב להיטיו של אלוויס פרסלי. החברה החליטה להפיץ את האוסף ע"י משלוח מכתבים לקונים פוטנציאלים. לשם כך פנתה לחברת שיווק שלה מאגר כתובות גדול מאוד ומתוכו נבחר מדגם מקרי של 6000 כתובות ונשלח אליהן מכתב ובו תיאור האוסף. בעקבות כך בוצעו 270 רכישות.

החברה החליטה לבדוק האם לאזור המגורים יש קשר למידת ההיענות של המשפחות להצעת הרכישה של האוסף. בבדיקה שערכה התברר שמבין 6000 המשפחות שבמדגם:

- 2400 התגוררו בערים גדולות והן ביצעו 130 הזמנות
- 1200 התגוררו בישובים כפריים והן ביצעו 50 הזמנות
- השאר התגוררו בערים קטנות ובינוניות

האם יש תלות בין אזור המגורים למידת ההיענות להצעת הרכישה ברמת מובהקות 0.05? נמק. **(כן)**

<u>שאלה 3</u>

נרכז את הנתונים בטבלה:

סוג יישוב / רכישה	1 ערים גדולות	2 כפרים	3 ערים קטנות/בינוניות	סה"כ
1 - כן	130	50	90	270
2 - לא	2270	1150	2310	5730
סה"כ	2400	1200	2400	6000

הנתונים שמופיעים בשאלה מסומנים באפור. את שאר הנתונים ניתן להשלים בעזרתם.

ההשערות:

H_0: אין תלות בין אזור המגורים למידת ההיענות

H_1: אחרת

	1 ערים גדולות	2 כפרים	3 ערים קטנות/בינוניות
1 - כן	$E_{11}=\frac{270\cdot 2400}{6000}=108$	$E_{12}=\frac{270\cdot 1200}{6000}=54$	$E_{13}=\frac{270\cdot 2400}{6000}=108$
2 - לא	$E_{21}=\frac{2400\cdot 5730}{6000}=2292$	$E_{22}=\frac{1200\cdot 5730}{6000}=1146$	$E_{23}=\frac{2400\cdot 5730}{6000}=2292$

חישוב סטטיסטי המבחן:

$$\chi^2_{emp}=\frac{(130-108)^2}{108}+\frac{(50-54)^2}{54}+\frac{(90-108)^2}{108}+\frac{(2270-2292)^2}{2292}+\frac{(1150-1146)^2}{1146}$$
$$+\frac{(2310-2292)^2}{2292}=8.14>5.99=\chi^2_{0.95,\ (1\times 2)}$$

ולכן נדחה את H_0 בר"מ 0.05, כלומר קיימת תלות בין אזור המגורים לבין מידת ההיענות.