תרגול כיתה 4 – מבוא להסתברות וסטטיסטיקה

בדיקת השערות למדגמים מזווגים , מבחני חי-בריבוע, רגרסיה

רווח סמך ובדיקת השערות ל*הפרש תוחלות* של שתי אוכלוסיות *תלויות* (מדגמים מזווגים)

תהיינה התצפיות במדגם הראשון $\left(X_1,...,X_n\right)$, ובמדגם השני $\left(Y_1,...Y_n\right)$.

נגדיר את סדרת ההפרשים: $d_i = X_i - Y_i \quad (i=1,...,n)$. בעזרת סדרת ההפרשים הבעיה הופכת להסקה/בניית רו"ס על אוכלוסייה אחת, עם שונות לא ידועה, כאשר ההסקה היא על הפרש התוחלות באוכלוסיות $\mu_d = \mu_X - \mu_Y$.

השונות המדגמית של סדרת ההפרשים היא- $S_d^2 = \frac{\sum_{i=1}^{n}\left(d_i - \bar{d}\right)^2}{n-1}$

ממוצע סדרת ההפרשים- $\bar{d} = \frac{1}{n}\sum_{i=1}^{n} d_i$

רווח סמך ברמת סמך (=רמת בטחון) $1-\alpha$ עבור $\left(\mu_1 - \mu_2\right)$ נתון ע"י:

$$\bar{d} - t_{(n-1);1-\alpha/2} \cdot \frac{S_d}{\sqrt{n}} \le \left(\mu_1 - \mu_2\right) \le \bar{d} + t_{(n-1);1-\alpha/2} \cdot \frac{S_d}{\sqrt{n}}$$

או בכתיב אחר: $\boxed{\mu_d \in \bar{d} \pm t_{(n-1);1-\alpha/2} \cdot \frac{S_d}{\sqrt{n}}}$

סטטיסטי המבחן: $t_{\bar{d}} = \frac{\bar{d} - \mu_d}{S_d / \sqrt{n}} \sim t_{(n-1)}$

שאלה 1

במפעל "חרוצי התעשיה" ביקשו לבדוק את ההשערה שתפוקת העובדים הממוצעת גדולה יותר בשעות הבוקר מאשר בשעות אחה"צ. לשם כך נלקח מדגם מקרי אשר כלל 5 מעובדי המפעל והתקבלו התוצאות הבאות:

מס' עובד	5	4	3	2	1
תפוקה בבוקר	5.5	5.8	6	5.3	5.1
תפוקה אחה"צ	4.5	6.3	5	4.8	4.1

א. בנו רו"ס להפרש תפוקת העובדים הממוצעת בין הבוקר ואחה"צ, ברמת סמך של 95%.

ב. בצעו בדיקה של ההשערה בשאלה, בר"מ 1%.

פתרון:
המדגמים תלויים (מזווגים) משום שלאותו עובד נמדדה התפוקה בבוקר ואחה"צ.
X – תפוקת העובד בבוקר, Y – תפוקת העובד אחה"צ.
ההפרש: $d_i = X_i - Y_i \quad (i = 1, \ldots, 5)$
סדרת ההפרשים: $d_1 = 1,\ d_2 = 0.5,\ d_3 = 1,\ d_4 = -0.5,\ d_5 = 1$
ממוצע וסטיית תקן של סדרת ההפרשים: $\bar{d} = 0.6,\ S_d = 0.65$

א). רווח הסמך:
לצורך בניית רווח סמך ברמת סמך (=בטחון) 95%, נמצא תחילה את אלפא-
$$1-\alpha = 0.95 \Rightarrow \alpha = 0.5 \Rightarrow 1-\alpha/2 = 0.975$$
כעת נמצא בטבלת t את הערך המתאים: $t_{(n-1);1-\alpha/2} = t_{4;0.975} = 2.78$
מכאן שרווח הסמך-
$$\mu_d = \bar{d} \pm t_{(n-1);1-\alpha/2} \cdot \frac{S_d}{\sqrt{n}} = 0.6 \pm 2.78 \frac{0.65}{\sqrt{5}} = 0.6 \pm 0.808 \quad \Rightarrow \boxed{-0.208 \le \mu_d \le 1.408}$$

ב). בדיקת ההשערה:
$$\begin{cases} H_0: & \mu_d \le 0 \\ H_1: & \mu_d > 0 \end{cases}$$
השערת האפס = הפרש התפוקות בין הבוקר ואחה"צ <u>קטן או שווה</u> לאפס. ההשערה האלטרנטיבית = הפרש התפוקות בין הבוקר ואחה"צ <u>גדול</u> מאפס (כלומר, ההערה האלטרנטיבית היא שתפוקת העובדים הממוצעת גדולה יותר בשעות הבוקר מאשר בשעות אחה"צ – מה שמבקשים לבדוק).

תחת השערה האפס, ההנחה שאין הפרש בין התפוקות, כלומר $\mu_0 = 0$

סטטיסטי המבחן: $t_{\bar{d}} = \dfrac{\bar{d} - \mu_d}{S_d / \sqrt{n}} \sim t_{(n-1)}$

בעזרת הסטטיסטי נמצא את הערך הקריטי של המבחן (K):
$$K_d = \mu_0 + t_{(n-1);1-\alpha} \cdot \frac{S_d}{\sqrt{n}} = 0 + t_{(4);0.99} \cdot \frac{0.65}{\sqrt{5}} = 3.75 \cdot \frac{0.65}{\sqrt{5}} = 1.09$$

השוואת הערך הקריטי:
$$\bar{d} = 0.6 < K_d = 1.09$$
המסקנה: לא דוחים את H_0 בר"מ 1%.
כלומר, תפוקת העובדים הממוצעת <u>איננה</u> גדולה יותר בשעות הבוקר מאשר בשעות אחה"צ.

מבחני חי-בריבוע מבחן – מבחן טיב התאמה

עבור מדגם מקרי בגודל n מהאוכלוסייה ההשערות הן:

H_0: המשתנה מתאים להתפלגות.

H_1: המשתנה לא מתאים להתפלגות.

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

איזור הדחיה: $\chi^2 > \chi^2_{(k-1);1-\alpha}$

שאלה 2

קוביית משחק הוטלה 600 פעם במטרה לבדוק האם היא הוגנת. התקבלו התוצאות הבאות:

תוצאה	1	2	3	4	5	6
שכיחות	82	92	112	88	108	117

האם הקובייה אכן הוגנת בר"מ 10%?

פתרון:

השערות המבחן:

H_0: המשתנה מתאים להתפלגות.

H_1: המשתנה לא מתאים להתפלגות.

ההנחה שהמספרים בקובייה מתפלגים אחיד U(1,6). לכן-

$$E_i = p_i \cdot n = \tfrac{1}{6} \cdot 600 = 100 \qquad (i = 1,...,6)$$

הסטטיסטי:

$$\chi^2 = \sum_{i=1}^{k} \frac{(O_i - E_i)^2}{E_i}$$

$$= \frac{(82-100)^2}{100} + \frac{(92-100)^2}{100} + \frac{(112-100)^2}{100} + \frac{(88-100)^2}{100}$$

$$+ \frac{(108-100)^2}{100} + \frac{(117-100)^2}{100} = 9.94$$

$$9.94 = \chi^2 > \chi^2_{(k-1);1-\alpha} = \chi^2_{5;0.90} = 9.236$$

<= לכן דוחים H_0 בר"מ 10%, כלומר הקובייה לא הוגנת.

<u>מבחני חי-בריבוע מבחן – לאי תלות (בדיקת קשר בין משתנים)</u>

מבחן χ^2 לאי תלות הוא מבחן סטטיסטי הבודק האם אין תלות בין שני משתנים של אותה אוכלוסייה. נניח שלגבי שני המשתנים: האחד בעל r קטגוריות והאחר בעל c קטגוריות.

עבור מדגם מקרי בגודל n מהאוכלוסייה ההשערות הן:

H_0: אין תלות בין שני המשתנים

H_1: יש תלות בין שני המשתנים

איזור הדחייה הוא: $\chi^2 > \chi^2_{(r-1)(c-1):\,1-\alpha}$.

<u>שאלה 3</u>

נתונה הטבלה הבאה :

	מדעי החברה	מדעי הטבע	מדעי הרוח	סה"כ
טיפוס א'	56	47	50	153
טיפוס ב'	8	14	5	27
סה"כ	64	61	55	180

מגדירים את המשתנים: X – טיפוס אישיות, Y – תחום לימוד.
בדוק האם יש תלות בין טיפוס אישיות X , אליו משתייך נבדק מסוים לבין תחום הלימוד שלו Y, ברמת מובהקות 0.05.

<u>פתרון</u>:

ההשערות:

H_0: <u>אין</u> תלות בין העדפה לימודית לטיפוס האישיות

H_1: <u>יש</u> תלות בין העדפה לימודית לטיפוס האישיות

$O_{i,j}$ השכיחות אשר התקבלה במדגם עבור קטגוריה i של המשתנה האחד, ועבור קטגוריה j של המשתנה האחר. (זוהי הטבלה הנתונה בשאלה).

$E_{i,j}$ השכיחות הצפויה בהנחה כי H_0 נכונה עבור קטגוריה i של המשתנה האחד, ועבור קטגוריה j של המשתנה האחר.

נחשב את $E_{i,j} = \frac{A \cdot B}{n}$, כאשר A השכיחות של קטגוריה i, B השכיחות של קטגוריה j.

הטבלה הצפויה המתקבלת היא $E_{i,j}$:

סה"כ	מדעי הרוח	מדעי הטבע	מדעי החברה	
153	$\frac{153\cdot 55}{180}=46.75$	$\frac{153\cdot 61}{180}=51.85$	$\frac{153\cdot 64}{180}=54.4$	טיפוס א'
27	$\frac{27\cdot 55}{180}=8.25$	$\frac{27\cdot 61}{180}=9.15$	$\frac{27\cdot 64}{180}=9.6$	טיפוס ב'
180	55	61	64	סה"כ

הסטטיסטי-

$$\chi^2=\sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}}=$$

$$=\frac{(56-54.4)^2}{54.4}+\frac{(47-51.85)^2}{51.85}+\frac{(50-46.75)^2}{46.75}+\frac{(8-9.6)^2}{9.6}+\frac{(14-9.15)^2}{9.15}+$$

$$\frac{(5-8.25)^2}{8.25}=0.047+0.45+0.226+0.266+2.57+1.28=4.7$$

התפלגותו-

$$\chi^2=\sum_{i=1}^{2}\sum_{j=1}^{3}\frac{(O_{i,j}-E_{i,j})^2}{E_{i,j}}\sim\chi^2_{(2-1)(3-1)}$$

איזור הדחייה הוא $\chi^2>\chi^2_{(2-1)(3-1):\,1-\alpha}$.

אצלנו $\chi^2_{(2-1)(3-1):\,0.95}=5.99$ ומאחר ש- $4.7<5.99$ <= לא דוחים את H_0.

כלומר: ניתן להסיק כי באוכלוסייה אין תלות בין העדפה לימודית לטיפוס האישיות.

<u>רגרסיה ומתאם</u>

<u>מקדם המתאם (r)</u>

נסמן-

$$S_{XX}=\sum_{i=1}^{n}\left(x_i-\overline{x}\right)^2=\sum_{i=1}^{n}x_i^2-n\cdot\overline{x}^2$$

$$S_{YY}=\sum_{i=1}^{n}\left(y_i-\overline{y}\right)^2=\sum_{i=1}^{n}y_i^2-n\cdot\overline{y}^2$$

$$S_{XY}=\sum_{i=1}^{n}\left(x_i-\overline{x}\right)\left(y_i-\overline{y}\right)=\sum_{i=1}^{n}x_iy_i-n\cdot\overline{x}\cdot\overline{y}$$

נוסחת מקדם המתאם:

$$r = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}} = \frac{\sum_{i=1}^{n}(x_i-\overline{x})(y_i-\overline{y})}{\sqrt{\sum_{i=1}^{n}(x_i-\overline{x})^2}\sqrt{\sum_{i=1}^{n}(y_i-\overline{y})^2}} = \frac{\sum_{i=1}^{n}x_iy_i - n\cdot\overline{x}\cdot\overline{y}}{\sqrt{\left(\sum_{i=1}^{n}x_i^2 - n\cdot\overline{x}^2\right)\left(\sum_{i=1}^{n}y_i^2 - n\cdot\overline{y}^2\right)}}$$

<u>קו הרגרסיה:</u>

קו הרגרסיה של Y כתלות ב- X: $\hat{Y} = b_{Y|X}X + a_{Y|X}$

כאשר: $a_{Y|X} = \overline{Y} - b_{Y|X}\overline{X}$; $b_{Y|X} = \frac{S_{XY}}{S_{XX}} = r\frac{\sqrt{S_{YY}}}{\sqrt{S_{XX}}}$

קו הרגרסיה של X כתלות ב- Y: $\hat{X} = b_{X|Y}X + a_{X|Y}$

כאשר: $a_{X|Y} = \overline{X} - b_{X|Y}\overline{Y}$; $b_{X|Y} = \frac{S_{XY}}{S_{YY}} = r\frac{\sqrt{S_{XX}}}{\sqrt{S_{YY}}}$

<u>שאלה 4</u>

נתונים ציונים של 5 תלמידים במתמטיקה (x) ובפיזיקה (y) בסולם ציונים [0-10]:

y	x
7	6
7	7
6	8
8	9
7	10

א. חשב את מקדם המתאם בין הציונים במתמטיקה לציונים בפיזיקה.
ב. מצא את משוואת הרגרסיה של הציון בפיזיקה על סמך הציון במתמטיקה.
ג. מצא את משוואת הרגרסיה של הציון במתמטיקה על סמך הציון בפיזיקה.
ד. תלמיד קיבל ציון 8.5 במתמטיקה, איזה ציון תנבא לו בפיזיקה?.
ה. סעיף אינו קשור לסעיפים הקודמים. נתון קו רגרסיה $\hat{Y} = -2X + 5$.
   i. אם ידוע שהממוצע של X הוא 1.5, מהו הממוצע של Y?
   ii. אם ידוע שסטיית התקן של Y גדולה פי 3 מסטיית התקן של X, מהו מקדם המתאם בין X ל-Y?

פתרון:

א). נתון: $n = 5$

<u>חישובי עזר:</u>

הממוצעים: $\bar{X} = 8,\ \bar{Y} = 7$

$$S_{XX} = \left(6^2 + 7^2 + 8^2 + 9^2 + 10^2\right) - 5 \cdot 8^2 = 10$$

$$S_{YY} = \left(7^2 + 7^2 + 6^2 + 8^2 + 7^7\right) - 5 \cdot 7^2 = 2$$

$$S_{XY} = \left(6 \cdot 7 + 7 \cdot 7 + 8 \cdot 6 + 9 \cdot 8 + 10 \cdot 7\right) - 5 \cdot 7 \cdot 8 = 1$$

נציב בנוסחת מקדם המתאם:

$$r = \frac{S_{XY}}{\sqrt{S_{XX}}\sqrt{S_{YY}}} = \frac{1}{\sqrt{2} \cdot \sqrt{10}} = +0.224$$

כלומר, קיים מתאם לינארי חיובי (חלש) בין הציונים במתמטיקה לציונים בפיזיקה.

ב). קו הרגרסיה המבוקש $\hat{Y} = aX + b$

$$a = \bar{Y} - b\bar{X} = 7 - 0.1 \cdot 8 = 6.2 \qquad b = \frac{S_{XY}}{S_{XX}} = \frac{1}{10} = 0.1$$

והקו המבוקש: $\hat{Y} = 0.1X + 6.2$

ג). קו הרגרסיה המבוקש $\hat{X} = aY + b$

$$a = \bar{X} - b\bar{Y} = 8 - 0.5 \cdot 7 = 4.5 \qquad b = \frac{S_{XY}}{S_{YY}} = \frac{1}{2} = 0.5$$

והקו המבוקש: $\hat{X} = 0.5Y + 4.5$

ד). $\hat{Y}\left(8.5\right) = 0.1X + 6.2 = 0.1 \cdot 8.5 + 6.2 = 7.05$

ה). קו הרגרסיה הנתון: $\hat{Y} = -2X + 5$

(i). מקו הרגרסיה נובע: $a = 5, b = -2$. נציב בנוסחה $a_{Y|X} = \bar{Y} - b_{Y|X}\bar{X}$

$$5 = \bar{Y} - (-2) \cdot 1.5 = \bar{Y} + 3 \qquad \Rightarrow \boxed{\bar{Y} = 2}$$

(ii). נתון $\sqrt{S_{YY}} = 3\sqrt{S_{XX}}$, מכאן-

$$-2 = r \cdot \frac{\sqrt{S_{YY}}}{\sqrt{S_{XX}}} = r \cdot \frac{3\sqrt{S_{XX}}}{\sqrt{S_{XX}}} = 3r \qquad \Rightarrow \boxed{r = -2/3}$$