הסקה סטטיסטית שיעור 10

הרצאה

מצגת שיעור 11

בדיקת אי-תלות בטבלאות דו-מימדיות

דוגמה: דיכאון ומצב משפחתי

בדיקה של 159 מטופלים בעלי דיכאון, מסווגים לפי רמת דיכאן (קל, בינוני, חמור) ומצב משפחתי (רווק, נשוי, גרוש):

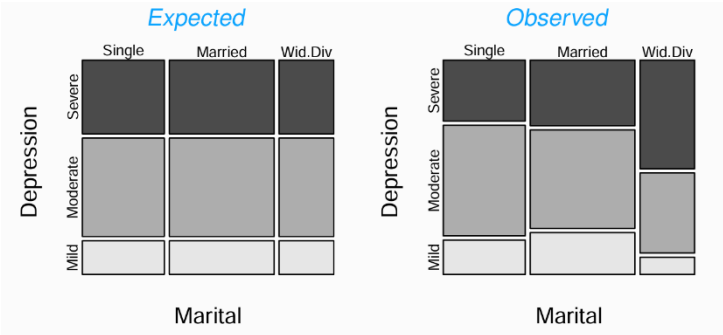
Depression	Marital Status			Total
	Single	*Married*	*Wid/Div*	
Severe	16	22	19	57
Moderate	29	33	14	76
Mild	9	14	3	26
Total	54	69	36	159

האם מצב משפחתי משפיע על דיכאון?

אפשר לחשב את ההתפלגות המותנית של רמת דיכאון בהינתן מצב משפחתי ע"י חלוקת המספר בכל תא, בסכום של העמודה:

Depression	**Marital Status**			
	Single	*Married*	*Wid/Div*	Overall
Severe	$\frac{16}{54} \approx 0.30$	$\frac{22}{69} \approx 0.32$	$\frac{19}{36} \approx 0.53$	$\frac{57}{159} \approx 0.36$
Moderate	$\frac{29}{54} \approx 0.54$	$\frac{33}{69} \approx 0.48$	$\frac{14}{36} \approx 0.39$	$\frac{76}{159} \approx 0.48$
Mild	$\frac{9}{54} \approx 0.17$	$\frac{14}{69} \approx 0.20$	$\frac{3}{36} \approx 0.08$	$\frac{26}{159} \approx 0.16$
Column Total	1	1	1	1

אם אין השפעה, נצפה שה- *observed* יהיה כמו ה- *expected*.



אם המשתנה בעמודות והמשתנה בשורות בלתי תלויים, אז ההתפלגות המותנית של העמודה בהינתן שורה:

$$P(\text{column var.} \mid \text{row var.}) = \frac{\text{cell count}}{\text{row total}}$$

אמורה להיות שווה להתפלגות השולית של המשתנה בעמודה:

$$P(\text{column var.}) = \frac{\text{column total}}{\text{overall total}}$$

כלומר, הספירה המצופה של התאים תחת הנחת אי-תלות היא:

$$\text{expected cell count} = \frac{\text{row total} \times \text{column total}}{\text{overall total}}$$

אז הספירה המצופה:

Depression	**Marital Status**			**Row Total**
	Single	*Married*	*Wid/Div*	
Severe	$\frac{57\times54}{159} = 19.37$	$\frac{57\times69}{159} = 24.74$	$\frac{57\times36}{159} = 12.91$	57
Moderate	$\frac{76\times54}{159} = 25.81$	$\frac{76\times69}{159} = 32.98$	$\frac{76\times36}{159} = 17.21$	76
Mild	$\frac{26\times54}{159} = 8.83$	$\frac{26\times69}{159} = 11.28$	$\frac{26\times36}{159} = 5.89$	26
Column Total	54	69	36	159

בדיקת אי-תלות

1. השערות:

השערת האפס H_0 היא שהמשתנים של השורות והעמודות בת"ל.

ההשערה האלטרנטיבית H_A היא שהמשתנים של השורות והעמודות לא בת"ל.

2. נבנה טבלה של ספירה מצופה לפי הנוסחה:

$$\text{expected cell count} = \frac{\text{row total} \times \text{column total}}{\text{overall total}}$$

3. אם השערת האפס נכונה, הספירה המצופה והאמיתית אמורות להיות קרובות. נמדוד את הקרבה שלהן ע"י סטטיסטי כי-בריבוע:

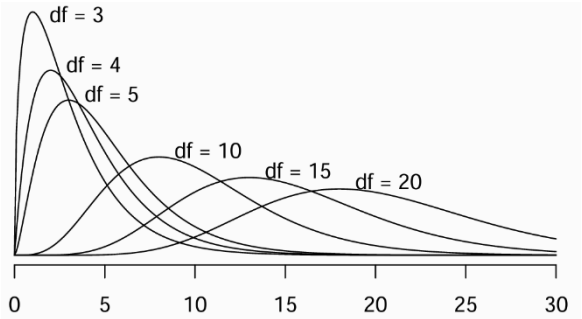
$$\chi^2 = \sum_{\text{all cells}} \frac{(O-E)^2}{E}$$

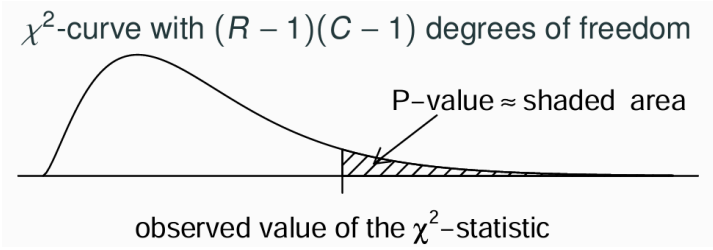
4. ככל שהסטטיסטי גדול, זאת עדות חזקה יותר נגד השערת האפס (בעד דחייה).

5. מה הגודל הרגיל של χ^2 תחת H_0?

לסטטיסטי χ^2 יש התפלגות בערך χ^2 עם $(R-1)(C-1)$ דרגות חופש.

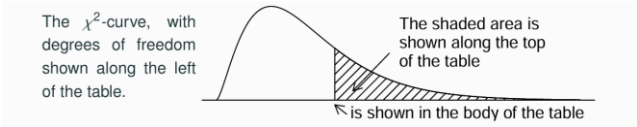
הערך *p-value* הוא בערך השטח מתחת לזנב הימני מעבר לסטטיסטי χ^2:



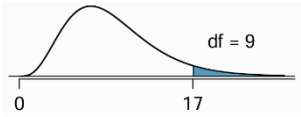


טבלה:

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82
	3	3.66	4.64	6.25	7.81	9.84	11.34	12.84	16.27
	4	4.88	5.99	7.78	9.49	11.67	13.28	14.86	18.47
	5	6.06	7.29	9.24	11.07	13.39	15.09	16.75	20.52
	6	7.23	8.56	10.64	12.59	15.03	16.81	18.55	22.46
	7	8.38	9.80	12.02	14.07	16.62	18.48	20.28	24.32
	8	9.52	11.03	13.36	15.51	18.17	20.09	21.95	26.12
	9	10.66	12.24	14.68	16.92	19.68	21.67	23.59	27.88
	10	11.78	13.44	15.99	18.31	21.16	23.21	25.19	29.59
	11	12.90	14.63	17.28	19.68	22.62	24.72	26.76	31.26
	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮



אז לדוגמה, נניח ש $\chi^2 = 10.3$, עם df = 6. הערך *p-value* יהיה בין העמודה של 8.56 ל-10.64. כלומר $0.1 < p < 0.2$.



בחזרה לדוגמה של הדיכאון:

Depression	Marital Status			Row Total
	Single	*Married*	*Wid/Div*	
Severe	16 (19.36)	22 (24.74)	19 (12.90)	57
Moderate	29 (25.81)	33 (32.98)	14 (17.21)	76
Mild	9 (8.83)	14 (11.28)	3 (5.89)	26
Column Total	54	69	36	159

בכל תא:

ספירה אמיתית
(ספירה מצופה)

אצלנו:

$$\chi^2 = 6.83, \qquad df = 4$$

כלומר, $0.1 < p < 0.2$, אז עם מובהקות 0.005, לא נדחה את השערת האפס.

אם H_0 נכונה, יש בערך 10% הסתברות לקבל $\chi^2 > 7.78$. אז $\chi^2 = 6.83$ לא ממש מפתיע.

מתי נשתמש במבחן כי-בריבוע?

- כאשר הדגימות הן דגימות רנדומיות פשוטות (*Simple Random Samples – SRS*).
- כל הספירות המצופות הן 5 או יותר.

דוגמה 1

האם שתיית אלכוהול בזמן הריון (נמדד במספר משקאות ביום) משפיעה על הסבירות לפגמים באיברי מין של התינוקות.

Alcohol Consumption	Observed Malformation Absent	Observed Malformation Present	Expected Malformation Absent	Expected Malformation Present
0	17,066	48	17,065.14	48.86
< 1	14,464	38	14,460.60	41.40
1-2	788	5	790.74	2.26
3-5	126	1	126.64	0.36
≥ 6	37	1	37.89	0.11

הטבלה הזאת לא מתאימה למבחן כי-בריבוע, כי לא כל הספירות המצופות הן לפחות 5.

דוגמה 2

827 תושבי קליפורניה שנרשמו לבחירות, נשאלו את השאלה הבאה: האם אתם תומכים או מתנגדים לשאיבת נפט וגז טבעי מול חופי קליפורניה? או שאתם לא בקיאים מספיק כדי לומר? התגובות חולקו לפי האם הנשאל סיים לימודים גבוהים (מכללה) או לא.

	College Grad Yes	College Grad No	Total
Support	154	132	286
Oppose	180	126	306
Do not know	104	131	235
Total	438	389	827

נבנה מבחן כי-בריבוע כדי לבדוק האם יש השפעה: השערת האפס – אין השפעה.

הספירה המצופה:

	College Grad Yes	College Grad No	Total
Support	$\frac{286\times438}{827} = 151.47$	$\frac{286\times389}{827} = 134.53$	286
Oppose	$\frac{306\times438}{827} = 162.07$	$\frac{306\times389}{827} = 143.93$	306
Do not know	$\frac{235\times438}{827} = 124.46$	$\frac{235\times389}{827} = 110.54$	235
Total	438	389	827

מתקיים:

$$\chi^2 \approx 11.46, \qquad df = 2$$

כלומר $0.001 < p < 0.005$. כלומר יש הבדל משמעותי.

דוגמה 3

בדיקה האם טיפול בהורמונים בתקופת עצירת וסת משפיע על הסיכוי לסרטן. השערת האפס – המשתנים בת"ל, אין השפעה.

	Cancer	No Cancer	Total
Hormone	107	8399	8506
Placebo	88	8014	8102
Total	195	16413	16608

Expected counts:

	cancer	no cancer	total
hormone	$\frac{8506\times195}{16608} = 99.87$	$\frac{8506\times16413}{16608} = 8406.13$	8506
placebo	$\frac{8102\times195}{16608} = 95.13$	$\frac{8102\times16413}{16608} = 8006.87$	8102
total	195	16413	16608

מתקיים: $\chi^2 \approx 1.0553$, עם $df = 1$. $p > 0.3$, כלומר לא נדחה את השערת האפס.

לחלופין, אפשר לבצע מבחן z לשני מדגמים (אם אנחנו יודעים את השונות והמדגם גדול) או מבחן t (אם השונות לא ידועה או המדגם קטן). בגלל שהמדגם גדול נשתמש במבחן z (ונשתמש ב- *standard error* במקום שונות). סטטיסטי המבחן הוא:

$$z = \frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\widehat{p}\,(1-\widehat{p}\,)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} = \frac{\frac{107}{8606} - \frac{88}{8102}}{\sqrt{\frac{195}{16608}\left(1-\frac{195}{16608}\right)\left(\frac{1}{8506}+\frac{1}{8102}\right)}} \approx 1.02728$$

כאשר כל p מייצג את היחס בין הספירה לסכום הכולל תחת השערת האפס.

הערך *p-value* הדו-צדדי הוא:

$$2P(Z > 1.03) = 2(1 - 0.8485) = 0.303$$

זה עובד גם באופן כללי: מבחן כי-בריבוע לטבלה 2×2 זהה למבחן z ליחסים דו-צדדי לשני מדגמים:

$$H_0 : p_1 = p_2 \quad \text{v.s.} \quad H_a : p_1 \neq p_2$$

	observed			expected	
	success	failure	total	success	failure
sample 1	X_1	$n_1 - X_1$	n_1	$n_1\hat{p}$	$n_1(1-\hat{p})$
sample 2	X_2	$n_2 - X_2$	n_2	$n_2\hat{p}$	$n_2(1-\hat{p})$
total	$X_1 + X_2$	$n_1+n_2-X_1-X_2$	$n_1 + n_2$	where $\hat{p} = \frac{X_1+X_2}{n_1+n_2}$	

מתקיים:

$$\chi^2\text{-statistic} = \sum \frac{(O-E)^2}{E} = \left(\frac{\widehat{p_1} - \widehat{p_2}}{\sqrt{\widehat{p}\,(1-\widehat{p}\,)\left(\frac{1}{n_1}+\frac{1}{n_2}\right)}} \right)^2 = (z\text{-statistic})^2$$

כאשר:

$$\widehat{p_1} = \frac{X_1}{n_1}, \qquad \widehat{p_2} = \frac{X_2}{n_2}$$

ושני המבחנים נותנים ערכי *p-values* זהים. יש הוכחה ארוכה ומייגעת (לא עשינו).

תרגול

תרגיל 1

תרגיל 2

נתון:

	אישה	גבר	
קדימה	44	3	47
אמצע	56	16	72
אחורה	22	10	32
	122	29	151

טבלת Expected:

	אישה	גבר
קדימה	37.9	9.02
אמצע	58.17	13.82
אחורה	25.8	6.14

$$\chi^2 = 8.39, \qquad df = 2$$

רמת מובהקות $\alpha = 0.05$. לפי הטבלה:

כלומר נדחה את השערת האפס.

Upper tail		0.3	0.2	0.1	0.05	0.02	0.01	0.005	0.001
df	1	1.07	1.64	2.71	3.84	5.41	6.63	7.88	10.83
	2	2.41	3.22	4.61	5.99	7.82	9.21	10.60	13.82