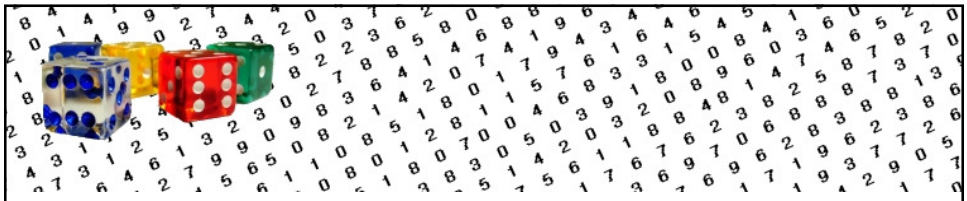


סקירת שיטות עיקריות לניתוח לוחות שכיחות

יום עיון למשתמשים בנתוני הלמ"ס – 5.1.2010

ד"ר יוסי לוי

http://www.sci-princess.com

הליידי טועמת תה

Fisher's Tea Tasting Experiment

	Guess Poured First		
Poured First	Milk	Tea	Total
Milk	3	1	4
Tea	1	3	4
Total	4	4	**8**

האם ליידי אנגליה יכולה לזהות מה נמזג קודם לספל: תה או חלב?

©יוסי לוי, 2012

מהו לוח שכיחות?

- התפלגות משותפת של שניים או יותר משתנים קטגוריים
- כל משתנה יכול להיות שמי או סדור

אמונה בחיים אחרי המוות

מין	כן	לא או לא בטוח/ה	סך הכל
נשים	$n_{11} = 509$	$n_{21} = 116$	$n_{1+} = 625$
גברים	$n_{21} = 398$	$n_{22} = 104$	$n_{2+} = 502$
סך הכל	$n_{+1} = 907$	$n_{+2} = 220$	$n = 1127$

אמונה בחיים אחרי המוות

מין	כן	לא או לא בטוח/ה	סך הכל
נשים	$p_{11} = \frac{509}{1127} = 0.452$	$p_{21} = \frac{116}{1127} = 0.103$	$p_{1+} = \frac{625}{1127} = 0.555$
גברים	$p_{21} = \frac{398}{1127} = 0.353$	$p_{22} = \frac{104}{1127} = 0.092$	$p_{2+} = \frac{502}{1127} = 0.445$
סך הכל	$p_{+1} = \frac{907}{1127} = 0.805$	$p_{+2} = \frac{220}{1127} = 0.195$	$p_{++} = \frac{1127}{1127} = 1.000$

התפלגויות שוליות

שאלת ~~מיליון~~ ביליון הדולר: האם יש קשר בין המשתנים?

אוטם שריר הלב (Myocardial Infarction)

קבוצה	כן	לא	סך הכל
פלסבו	189	10845	11034
אספירין	104	10933	11037

Source: Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study. *New Engl. J. Med.*, **318**: 262–264, 1988.

$p_{PL} = 189/11034 = 0.0171 \quad p_{AS} = 104/11037 = 0.0094$

רווח סמך (95%) להפרש הפרופורציות: $P_{PL} - P_{AS} = 0.0077 \pm 0.0003$

יחס הסיכון: $\text{Relative Risk} = \frac{0.0171}{0.0094} = 1.82$

רווח סמך (95%) ליחס הסיכון: (1.43, 2.30)

©יוסי לוי, 2012

שאלת ביליון הדולר:
האם יש קשר בין המשתנים?

אוטם שריר הלב (Myocardial Infarction)

קבוצה	כן	לא	סך הכל
פלסבו	189	10845	11034
אספירין	104	10933	11037

Source: Preliminary Report: Findings from the Aspirin Component of the Ongoing Physicians' Health Study. *New Engl. J. Med.*, **318**: 262–264, 1988.

$$\text{Odds Placebo} = \frac{189/11034}{10845/11034} = 0.0174$$

$$\text{Odds Aspirin} = \frac{104/11037}{10933/11037} = 0.0095$$

$$\text{Odds Ratio} = \frac{\text{Odds Placebo}}{\text{Odds Aspirin}} = \frac{0.0174}{0.0095} = \frac{189 \times 10933}{104 \times 10845} = 1.82$$

שאלת ביליון הדולר:
האם יש קשר בין המשתנים?

שכיחויות צפויות של אוטם שריר הלב בהנחת אי תלות בטיפול

קבוצה	כן	לא	סך הכל
פלסבו	$\frac{11034}{22071} \cdot 293 = 146.52$	$\frac{11034}{22071} \cdot 21778 = 10887.52$	11034
אספירין	$\frac{11037}{22071} \cdot 293 = 146.48$	$\frac{11037}{22071} \cdot 21778 = 10890.48$	11037
סך הכל	293	21778	22071

המרחק בין הערכים הנצפים והשכיחויות צפויות תחת הנחת אי תלות

קבוצה	כן	לא	סך הכל
פלסבו	$\frac{(104-146.52)^2}{146.52} = 12.339$	$\frac{(10845-10887.52)^2}{10887.52} = 0.166$	11034
אספירין	$\frac{(198-146.48)^2}{146.48} = 12.343$	$\frac{(10933-10890.48)^2}{10890.48} = 0.166$	11037
סך הכל	293	21778	22071

$$\chi^2 = 0.166 + 0.166 + 12.339 + 12.343 = 25.01$$

$$df = 1, \quad p - \text{value} < 0.001$$

©יוסי לוי, 2012

האם יש קשר בין מגדר והעדפה פוליטית?

	Democrat	Independent	Republican	Total
Female	762 703.7	327 319.7	468 533.7	1557
Male	484 542.3	239 246.4	477 411.3	1200
Total	1246	566	945	2757

$$E_{uj} = \frac{n_{i+} \times n_{+j}}{n} = \frac{1557 \times 1246}{2757}$$

$$\chi^2 = 30.1, \quad df = 2,$$
$$p - value < 0.001$$

יש קשר. מאין הוא נובע?

קשר בין מגדר והעדפה פוליטית – ניתוח שאריות

	Democrat	Independent	Republican
Female	762 ★ 2.20	327 0.41	468 ★ -2.84
Male	484 ★ -2.50	239 -0.47	477 ★ 3.24

$$\text{Residual}_{ij} = \frac{n_{ij} - E_{ij}}{\sqrt{E_{ij}}} = \frac{762 - 703.7}{\sqrt{703.7}}$$

- יש יותר נשים דמוקרטיות מהצפוי, ויותר גברים רפובליקנים
- יש פחות נשים רפובליקניות מהצפוי, ופחות גברים דמוקרטים

⇦ נשים יותר נוטות שמאלה, וגברים ימינה

קשר בין מגדר והעדפה פוליטית – Odds Ratio

	Democrat	Independent	Republican
Female	762	~~327~~	468
Male	484	~~239~~	477

$$\text{Odds Ratio} = \frac{762 \times 477}{468 \times 484} = \frac{762:468}{484:477} = \frac{1.63:1}{1:014:1} = 1.60$$

מבין המזדהים פוליטית, יחס הסיכויים להזדהות עם הדמוקרטים מול הזדהות עם הרפובליקנים גדול ב-60% אצל הנשים לעומת הגברים

עבריינות – מספר הרשעות ומקום המגורים

אידור מגורים

מספר הרשעות	שכונת מצוקה עירונית	שכונת יוקרה עירונית	עירוני אחר	לא עירוני	סך הכל
1	50	25	0	0	75
2 או 3	60	10	25	4	99
4 ויותר	75	10	15	1	101
סך הכל	185	45	40	5	275

ברור שיש קשר, אבל מה היא עצמתו?

$$\chi^2 = 42.7, \quad df = 6, \quad \text{p} - \text{value} < 0.001$$

©יוסי לוי, 2012

מדדי קשר מבוססי חי-בריבוע

$$\phi = \sqrt{\frac{\chi^2}{n}} = \sqrt{\frac{42.7}{275}} = 0.394$$

$$P = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{42.7}{42.7 + 275}} = 0.367$$

$$v = \sqrt{\frac{\chi^2 / n}{\min\{R-1, C-1\}}} = \sqrt{\frac{42.7/275}{\min\{3-1, 4-1\}}} = 0.279$$

עצמת הקשר חלשה

האם אזור המגורים מנבא את רמת העבריינות?

איזור מגורים

מספר הרשעות	שכונת מצוקה עירונית	שכונת יוקרה עירונית	עירוני אחר	לא עירוני	סך הכל
1	50	25	0	0	75
2 או 3	60	10	25	4	99
4 ויותר	75	10	15	1	101
סך הכל	185	45	40	5	275

ניבוי מספר הרשעות	4 ויותר	1	2 או 3	2 או 3	4 ויותר
מספר הניבויים המוטעים	185-75=110	45-25=20	40-25=15	5-4=1	275-101=174

©יוסי לוי, 2012

λ – מדד קשר לעצמת הניבוי

- ללא מידע על מקום המגורים היו 174 ניבויים מוטעים
- כאשר השתמשנו במידע על מקום המגורים היו 146 ניבויים מוטעים
- המידע על מקום המגורים "חסך" 174-146=28 ניבויים מוטעים, שהם כ-16%
- במלים אחרות – המידע על מקום המגורים הצליח להסביר כ-16% מהשונות ברמת העבריינות
- מקום המגורים הוא מנבא חלש לרמת העבריינות
- למעשה חישבנו כי λ=0.16

האם רמת העבריינות יכולה "לנבא" את איזור המגורים?

אידור מגורים

מספר הרשעות	שכונת מצוקה עירונית	שכונת יוקרה עירונית	עירוני אחר	לא עירוני	סך הכל
1	50	25	0	0	75
2 או 3	60	10	25	4	99
4 ויותר	75	10	15	1	101
סך הכל	185	45	40	5	275

למשתנה המנבא צריכות להיות יותר קטגוריות מאשר למשתנה המנובא

©יוסי לוי, 2012

אלכוהול, עישון ומריחואנה

Alcohol (A), Cigarette (C), and Marijuana (M) Use for High School Seniors

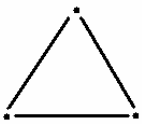
Alcohol Use	Cigarette Use	Marijuana Use: Yes	Marijuana Use: No
Yes	Yes	911	538
	No	44	456
No	Yes	3	43
	No	2	279

האם יש קשר בין השלושה?

מבני קשר אפשריים בין 3 משתנים

2 משתנים בלתי תלויים זה בזה בהינתן ערך המשתנה השלישי

קשר מלא בין כל 3 המשתנים



3 המשתנים אינם תלויים זה בזה

2 משתנים תלויים זה בזה, והשלישי אינו תלוי בהם

©יוסי לוי, 2012

אלכוהול, עישון ומריחואנה – מודל לוג-לינארי

Fitted Values for Loglinear Models

Alcohol Use	Cigarette Use	Marijuana Use	Loglinear Model (A, C, M)	(AC, M)	(AM, CM)	(AC, AM, CM)	(ACM)
Yes	Yes	Yes	540.0	611.2	909.24	910.4	911
		No	740.2	837.8	438.84	538.6	538
	No	Yes	282.1	210.9	45.76	44.6	44
		No	386.7	289.1	555.16	455.4	456
No	Yes	Yes	90.6	19.4	4.76	3.6	3
		No	124.2	26.6	142.16	42.4	43
	No	Yes	47.3	118.5	0.24	1.4	2
		No	64.9	162.5	179.84	279.6	279

כל שלושת המשתנים תלויים זה בזה

מקורות

לקריאה והרחבה בנושאים אלה:

- Alan Agresti (1996) An Introduction to Categorical Data Analysis ,Second Edition
- Alan Agresti (2002) Categorical Data Analysis ,Second Edition

כל הדוגמאות במצגת זו נלקחו מספרים אלה מאת אגרסטי, מלבד הדוגמה על הקשר בין רמת העבריינות ומקום המגורים, שנלקחה מ-"מבוא לסטטיסטיקה למדעי החברה", יחידה 5, האוניברסיטה הפתוחה.

©יוסי לוי, 2012