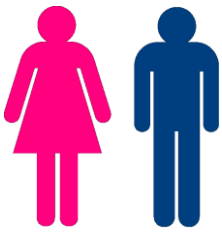
לוחות שכיחות: ניתוח נתונים בסולם מדידה שמי

תזכורת

- משתנים בסולם מדידה שמי הם משתנים שלערך שלהם אין משמעות מספרית
- מבחינה תיאורית יש את הכלים הבאים:
  - טבלת שכיחות
  - שכיח
  - דיאגרמת עמודות



© Yossi Levy 2021

שאלות מעניינות לגבי נתונים שמיים/קטגוריים

- מה ההתפלגות?
- האם יש קשרים בין שני משתנים או יותר?
- האם יש הבדלים בין אוכלוסיות?
- האם אפשר להסביר את ההתפלגות של משתנה קטגורי על סמך משתנים אחרים? (רגרסיה)

התפלגות של משתנה קטגורי – מבחן טיב ההתאמה

- במדגם כלשהו בגודל 100, התברר כי 58 מהנדגמים היו נשים ונדגמו רק 42 גברים.
- בהנחה כי 50% מהאוכלוסייה הם נשים, האם נתוני המדגם מתאימים להרכב האוכלוסייה?

ערך	Observed	Expected	$(O-E)^2$	$\frac{(O-E)^2}{E}$
נשים	58	50	$(58-50)^2=64$	$64/50=1.28$
גברים	42	50	$(50-42)^2=64$	$64/50=1.28$
סך הכל				2.56

נדחה את ההשערה כי התפלגות המינים במדגם שונה מההתפלגות באוכלוסייה אם הסכום של העמודה האחרונה גדול מאפס באופן מובהק

© Yossi Levy 2021

www.sci-princess.info

מבחן חי-בריבוע

- לסטטיסטי המוגדר על ידי הסכום של $\frac{(O-E)^2}{E}$ קוראים סטטיסטי חי בריבוע
- ההתפלגות של סטטיסטי זה נקראת התפלגות חי-בריבוע
- להתפלגות זו יש פרמטר הנקרא מספר דרגות החופש
- בדוגמה שראינו קודם – מספר דרגות החופש שווה ל-1

$\frac{(O-E)^2}{E}$	$(O-E)^2$	Expected	Observed	ערך
$64/50 = 1.28$	$(58-50)^2 = 64$	50	58	נשים
$64/50 = 1.28$	$(50-42)^2 = 64$	50	42	גברים
2.56				סך הכל

- אם רמת המובהקות שלנו היא α=0.05 לא נדחה את השערת האפס כי p-value = 0.1096

דוגמה: סוגי דם

- לפי ויקיפדיה, ל-32% מאוכלוסיית ישראל יש סוג דם +O, ל-34% יש סוג דם +A, ל-17% יש סוג דם +B, ול-7% יש סוג דם +AB. לשאר האוכלוסייה, 10% , יש סוגי דם עם rH שלילי.
- במבצע התרמת דם, הגיעו 200 תורמים. מתוכם 56 היו בעלי סוג דם +O, 63 עם +A, 38 עם +B, 15 עם +AB והשאר עם rH שלילי מסוגים שונים
- האם התפלגות סוגי הדם של התורמים משקפת את התפלגות סוגי הדם באוכלוסייה? (זוהי השערת האפס)

© Yossi Levy 2021
www.sci-princess.info

דוגמה: תורמי דם

Blood type	O+	A+	B+	AB+	rH -	Total
Observed	56	63	38	15	28	200
Population frequency	0.32	0.34	0.17	0.07	0.1	1
Expected	64	68	34	14	20	200
$O-E$	8	5	-4	-1	8	0
$(O-E)^2$	64	25	16	1	64	
$(O-E)^2/E$	1.00	0.37	0.47	0.07	3.2	5.11

$p-value = 0.2762$

לא דוחים את השערת האפס לפיה התפלגות סוגי הדם של התורמים משקפת את הפלגות סוגי הדם באוכלוסייה.

דוגמה: הבליץ על לונדון

האם לגרמנים הייתה יכולת לכוון את הטילים לפגיעה במטרה מסויימת?

© Yossi Levy 2021

הבליץ על לונדון

- העיר חולקה ל-576 איזורים ריבועיים, כל אחד מהם בגודל קמ"ר אחד
- נאספו נתונים על מספר הטילים שפגעו בכל איזור (בלילה אחד)
- אם לגרמנים אין יכולת כוונון, אז למספר הטילים שפוגעים באיזור מסויים יש התפלגות פואסון
- להתפלגות זו יש פרמטר יחיד המסומן באות λ וזוהי התוחלת של ההתפלגות

מספר הפגיעות	מספר האיזורים Observed
0	229
1	211
2	93
3	35
4	7
5*	1

הבליץ על לונדון – האם לגרמנים הייתה יכולת כוונון?

- נאמוד את λ על ידי המספר הממוצע של טילים שפעו בכל איזור
- נחשב את ההסתברות כי באיזור מסויים יפגעו k טילים על ידי פונקציית ההסתברות של מ"מ פואסוני
- נחשב את המספר הצפוי של איזורים בהם פגעו k טילים
- נשווה את ההתפלגות בה צפינו להתפלגות הצפויה בעזרת מבחן חי-בריבוע
- השערת האפס היא כי לגרמנים אין יכולת כוונון,
- במילים אחרות: השערת האפס היא כי התפלגות מספר הפגיעות היא התפלגות פואסון
- אם דוחים את השערת האפס כי ההתפלגות היא פואסונית נסיק כי לגרמנים יש יכולת כוונון

© Yossi Levy 2021

www.sci-princess.info

התפלגות פואסון

- להתפלגות פואסון יש פרמטר יחיד λ וזוהי התוחלת של ההתפלגות
- נאמוד את λ על ידי ממוצע המדגם:

מספר הפגיעות	מספר האיזורים Observed
0	229
1	211
2	93
3	35
4	7
5*	1

$$\hat{\lambda} = \frac{229 \cdot 0 + 211 \cdot 1 + 93 \cdot 2 + 35 \cdot 3 + 7 \cdot 4 + 1 \cdot 5}{576} = \frac{535}{576} = 0.929$$

- האמדן שלנו לתוחלת מספר הפגיעות באיזור מסויים, אם לגרמנים אין יכולת כוונון, הוא 0.929

חישובי ההסתברויות למספר הפגיעות באיזור מסויים

- פונקציית ההסתברות של התפלגות פואסון היא

$$P(X = k) = \frac{e^{-\lambda} \cdot \lambda^k}{k!}$$

$$P(X = 0) = \frac{e^{-0.929} \cdot 0.929^0}{0!} = 0.395$$

$$P(X = 1) = \frac{e^{-0.929} \cdot 0.929^1}{1!} = 0.367$$

$$P(X = 2) = \frac{e^{-0.929} \cdot 0.929^2}{2!} = 0.0.170$$

© Yossi Levy 2021

www.sci-princess.info

חישובים

$\chi^2 = 1.220$

$df = 4$

p-value $= 0.1252$

מספר הפגיעות	מספר האיזורים Observed	הסתברות $P(X=k)$	Expected	$(O-E)^2$	$(O-E)^2/E$
0	229	0.395	227.520	2.190	0.010
1	211	0.367	211.392	0.154	0.001
2	93	0.170	97.920	24.206	0.247
3	35	0.053	30.528	19.999	0.655
4	7	0.012	6.912	0.008	0.001
5+	1	0.003	1.728	0.530	0.307
סך הכל	576		576		1.220

- לא דוחים את השערת האפס
- **אין להסיק כי לגרמנים אין יכולת כוונון!**

הקשר בין מין והעדפה פוליטית

	Democrat	Independent	Republican	Total
Female	762	327	468	1557
Male	484	239	477	1200
Total	1246	566	945	2757

www.sci-princess.info

© Yossi Levy 2021

הקשר בין מין והעדפה פוליטית

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

Expected	Democrat	Independent	Republican
Female	4.95	0.16	8.37
Male	6.43	0.17	10.67

	Democrat	Independent	Republican
Female	$\frac{(762-703.0)^2}{703.0}$	$\frac{(327-319.8)^2}{319.8}$	$\frac{(468-534.9)^2}{534.9}$
Male	$\frac{(484-543.1)^2}{543.1}$	$\frac{(239-245.4)^2}{245.4}$	$\frac{(477-410.8)^2}{410.8}$

χ^2=4.95+0.16+8.37+6.43+0.17+10.67=30.75 $df = 2$ p-value < 0.0000

הקשר בין מין והעדפה פוליטית – מאין נובעים ההבדלים?

	Democrat	Independent	Republican
Female	$\frac{762-703.0}{\sqrt{703.0}}$	$\frac{327-319.8}{\sqrt{319.8}}$	$\frac{468-534.9}{\sqrt{534.9}}$
Male	$\frac{484-543.1}{\sqrt{543.1}}$	$\frac{239-245.4}{\sqrt{245.4}}$	$\frac{477-410.8}{\sqrt{410.8}}$

$(O-E)/\sqrt{E}$ נסתכל על

	Democrat	Independent	Republican
Female	2.23	0.40	-2.89
Male	-2.54	-0.41	3.27

גברים יותר נוטים להעדיף את הרפובליקנים ונשים את הדמוקרטים

www.sci-princess.info
© Yossi Levy 2021

מקדמי קשר – measures of association

- מקדמי קשר הם הכללה של מושג המתאם.
- מדד קשר, על פי הגדרתו, מודד את עצמת הקשר בין שניים או יותר משתנים
- קיימים עשרות, אם לא מאות, מדדי קשר שונים
- אנו נראה רק ארבעה מדדים:
  - שלושה מדדים המבוססים על המרחק של ההתפלגות המשותפת ממצב של אי תלות, כלומר מבוססים על סטטיסטי חי בריבוע
  - מדד אחד המבוסס על היכולת של משתנה אחד לנבא את הערך של המשתנה האחר

דוגמה: הקשר בין איזור מגורים ורמת הפשיעה

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

$\chi^2 = 42.7 \quad df = 6 \quad \text{p-value} < 0.000$

ברור שיש קשר, אך מהי עוצמתו?

© Yossi Levy 2021

www.sci-princess.info

מקדמי קשר מבוססי חי בריבוע

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

Pearson's Contingency coefficient $P = \sqrt{\frac{\chi^2}{\chi^2 + n}} = \sqrt{\frac{42.7}{42.7 + 275}} = 0.367$

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

מקדם קשר מבוסס ניבוי - λ

האם אזור המגורים מנבא את רמת העבריינות?

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

ניבוי ללא ידיעת איזור המגורים: 101

ניבוי בהינתן שכונת מצוקה: 75

ניבוי בהינתן שכונת יוקרה: 25

ניבוי בהינתן עירוני: 25

ניבוי בהינתן עירוני: 4

www.sci-princess.info

© Yossi Levy 2021

הקשר בין איזור המגורים ורמת הפשיעה

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

ניבויים מוטעים ללא ידיעת איזור המגורים	$275 - 101 = 174$
ניבויים מוטעים עם ידיעת מקום המגורים	$275 - (75 + 25 + 25 + 4) = 146$
סך הכל שיפור בניבוי	$174 - 146 = 28$
אחוז שיפור בניבוי	$\lambda = 28/174 = 0.161$

גודל האפקט

Table 2.3. Cross Classification of Aspirin Use and Myocardial Infarction

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

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

האם יש קשר בין נטילה קבועה של אספירין וסיכון להתקף לב?

www.sci-princess.info

© Yossi Levy 2021

הקשר בין נטילת אספירין וסיכון להתקף לב

Table 2.3. Cross Classification of Aspirin Use and Myocardial Infarction

	Myocardial Infarction		
Group	Yes	No	Total
Placebo	189	10,845	11,034
Aspirin	104	10,933	11,037

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

$$p_{PL} = P(\text{MI} \mid \text{Placebo}) = \frac{189}{11034} = 0.0171$$

$$p_{AS} = P(\text{MI}|\text{Aspirin}) = \frac{104}{11037} = 0.0094$$

רווח סמך להפרש הפרופורציות:

$$p_{PL} - p_{PS} = 0.0077 \pm 0.0003$$

יחס הסיכונים (עם רווח סמך):

$$\text{Relative risk} = \frac{0.0094}{0.0171} = 0.550 \pm 0.132$$

www.sci-princess.info

© Yossi Levy 2021