25.1.09 (1)

סטטיסטיקה למדעי המחשב - שיעור 12

<u>שיעור חזרה (רשות)</u>
יום ד' 10-12:30 (30.1) בשעור 6.

* 2 מבחנים לדוגמא יפורסמו במודל. המבחן יהיה דומה מבחינת סגנון ורמה. אינני לעגני יותר חומר. אך יכול להיות גם חומר שלא מופיע במבחן. וההפך.

לגבי המבחן ג.ר.

מבחן 3 שעות, ללא חומר עזר, מותר מחשבון, טבלאות וכן את דף הנוסחאות. הטבלאות יימסרו עם הבחינה (χ^2, t, z). לגבי דף הנוסחאות, ייתכן ויכתב ויחולק וייתכן ויהיה אפשר להביא אותו.

<u>תוכן הבחינה</u>

- שאלות הבנה ("שכל האינטואיציה")
- חישוב + הבנה.
- ניתוח נתונים + הבנה.

סביר <u>שלא</u> יהיו הרבה חישובים מסובכים. בדר"כ התשובות לשאלות יהיו משפט - שתיים.

- הבחינה של הלמה לוותן בשנה שעברה (שאלה קודמת מהמבחן) חויינה. דגש שם היה על הסקה סטטיסטית ומשקה.
- הבחינה על נושא בינומי גם רלוונטית (את יותר "רבת").

<u>מבחנים לשוויון פרופורציות</u>

n תצפיות מ-X $X \sim Bernoli(p_x)$

m תצפיות מ-Y $Y \sim Bernoli(p_y)$

א. מבחן בינומי / נורמלי

$$\hat{p}_x - \hat{p}_y \underset{H_0}{\overset{\cdot}{\sim}} N\left(0, \frac{p(1-p)}{n} + \frac{p(1-p)}{m}\right)$$

$$\hat{p} = \frac{n\hat{p}_x + m\hat{p}_y}{n+m}$$ אומד משותף לפרופורציה

⇐ לעשות מבחן Z

ב. תנאי (או נניח ידוע) גם מספר ההצלחות <u>הכולל</u>.

n מ-X, m מ-Y, N_+ סה"כ הצלחות

$$H_0: \underset{(n\hat{p}_x)}{n_+} \sim HG(n+m, N_+, n) \approx N\left(\frac{nN_+}{n+m}, \frac{n\cdot N_+(n+m-N_+)(n+m-n)}{(n+m)^2(n+m-1)}\right)$$ ⇐ נשתמש בקירוב נורמלי (HG.

דוגמה: קבלה למכמ"ח

$n=200$ בנים, $m=50$ בנות, $N_+=60$ מקומות לקבלה.

נניח שבשנה מסוימת התקבלו 50 בנים (25%), ו-10 בנות (20%).

שאלה: האם סיכויי הקבלה של בנות נמוכים יותר?

$H_0: P_X = P_Y$ (במילים: סיכויי הקבלה שווים).

מבחנים

א. קירוב נורמלי לבינומי (לא משתמשים ב-N_+). $Z=0.74$, P-val: 0.23.

ב. קירוב נורמלי ל-HG: P-val: 0.23.

התפלגות של HG בתור טבלה 2×2

	✓	×	סה"כ
בן	50	150	200
בת	10	40	50
סה"כ	60	190	

⇐ מספר הבנים שהתקבלו (n_+) הוא שמתפלג HG לפי שחישבנו קודם.

לפני שפותחים את הנתונים, ורואים רק את הסה"כ, כלומר, התפלגויות שוליות אמפיריות.

* ב-HG הסיכוי של כל תא בטבלה הוא מכפלת ההתפלגויות השוליות, ננסה לבדוק האם הערכים באמת נראים מתכונה זו *

⇐ דרך אחרת להסתכל על זה:

מ"מ U עם ערכים: $U = \begin{cases} 0 & \text{נקבה} \\ 1 & \text{זכר} \end{cases}$

מ"מ Z עם ערכים: $Z = \begin{cases} 0 & \text{לא} \\ 1 & \text{כן} \end{cases}$

השערת ה-0 שלנו היא ש-U ו-Z בלתי תלויים.

* ניתן להמשיך ולבדוק בדרך כזו גם עם טבלאות גדולות יותר. אך מכיוון שאין הסבר בין 2 המשתנים, לא כדאי להמשיך ולחקור זאת.

התפלגות χ^2

הגדרה: אם $X_1,\dots,X_k$ מ"מ ב"ת בעלי התפלגות נורמלית סטנדרטית, כלו' $X_i\sim N(0,1)$ אזי:

$$\sum_{i=1}^{k} X_i^2 \sim \chi^2_k$$

← מספר דרגות החופש

שאלה: מה אם $X\sim\chi^2_k$ ו- $Y\sim\chi^2_l$ והם ב"ת. מה ניתן לומר על הת' $X+Y$?

תשובה: $X+Y\sim\chi^2_{k+l}$ (כי X ו-Y הם סכומים של מ"מ X_i ב"ת).

נניח שיש לנו מדגם בגודל N, עם 2 ערכים של משתנים דיסקרטיים.

U: עם k ערכים אפשריים. $\{u_1,\dots,u_k\}$

Z: עם L ערכים אפשריים. $\{z_1,\dots,z_L\}$

25.1.09 3

נתון מדגם לפי u,z כך: H_0. (בדוגמה הקודמת: Z - מין, u - קבלה ללימודים)

נסדר את הנתונים בטבלה בגודל k×l.

n_{kl} - מספר האובייקטים שערך U שלהם u_k וערך Z הוא z_l.

	$Z=z_1$	$Z=z_2$	......	$Z=z_l$	סה"כ
$U=u_1$	n_{11}	n_{12}	----	n_{1l}	n_1
$U=u_2$	n_{21}	-	----	n_{2l}	⋮
⋮	⋮	⋮	⋮	⋮	⋮
$U=u_k$	n_{k1}	-	----	n_{kl}	n_k
סה"כ	m_1	m_2	----	m_l	

נחשב את הערכים השוליים $n_1, n_2, \ldots, n_k$ ו- $m_1, m_2, \ldots, m_k$

נבנה טבלה שניה, טבלת ה- expected:

$$e_{ij} = \frac{n_i \cdot m_j}{N} = \frac{n_i}{N} \cdot \frac{m_j}{N} \cdot N$$

דוגמא: (המשך ממקודם 2x2)

observed

	V	X
בנים	50	150
בנות	10	40

expected

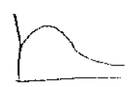
	V	X	
בנים	$\frac{200\times 60}{250} = 48$	152	200
בנות	12	38	50
	60	190	250

שימו לב: אם כל התאים ב- expected מכילים לפחות 5 תצפיות אזי

$$\sum_{\substack{i \text{ עובר על} \\ \text{כל תאי הטבלה}}} \frac{(o_i - e_i)^2}{e_i} \underset{\substack{H_0 \\ \text{(בקירוב)}}}{\sim} \chi^2_{(k-1)\cdot(l-1)}$$

*/ הסיבה שמורידים בדרגות החופש 1 מ- k ו- l היא כיוון שברגע שיש k-1 ו- l-1 נתונים, השאר ניתן להשלים ע"י חיסורים והסתכלות בשוליים */

*/ כשמחשבים את הערך p-val, וקיבלנו ערך שונה מ- 0.23 כיוון שערך זה מושג עם אלטרנטיבה חד צדדית ובמקרה זה אין משמעות לאלטרנטיבה חד צדדית. היא תמיד כוללת. כלומר אנו צריכים לקבל ערך 0.46 (פי 2) */



התפלגות χ^2 עם דרגה 1

$$\underset{\text{הסטטיסטי}}{\chi^2} = \frac{(50-48)^2}{48} + \frac{(150-152)^2}{152} + \frac{(10-12)^2}{12} + \frac{(40-38)^2}{38} = 0.55 \underset{H_0}{\sim} \chi^2_1$$

המעלות (דרגות החופש)

$$P\text{-val} = P(\chi^2_1 \geq 0.55) = \underline{\underline{0.46}}$$

סטטיסטיקה למדעי המחשב - שיעור 12

25.1.09 (9)

רגרסיה מול חי-בריבוע

נתונים זוגות $(x_n,y_n) \dots (x_1,y_1)$ (מספרים). אני רוצה לדעת אם יש קשר בין x,y. הדבר להשתמש ברגרסיה $y \sim \hat{a}+\hat{b}x$ ולבחון $H_0: b=0$ לפי $\hat{b}$.

גישה אחרת: נעשה דיסקרטיזציה ל-x,y. נניח: $x \in \{$נמוך, בינוני, גבוה$\}$ $y \in \{$נמוך, בינוני, גבוה$\}$

ונצייר טבלה:

	y=א	y=ב	y=ג	
x=א	$n_{אא}$			
x=ב				
x=ג			$n_{גג}$	
				n

ונעשה מבחן χ^2 לאי תלות.

דוגמה מספרית יותר גדולה

בודקים 3 טיפולים על 50 חולים כל אחד, ובודקים את רמת השיפור.

observed:

	תרופה א	תרופה ב	תרופה ג	
שיפור רב	5	13	20	38
שיפור כלשהו	32	28	27	87
ללא שינוי	13	9	3	25
	50	50	50	

H_0': אי תלות בין סוג התרופה לרמת השיפור (כלומר כל התרופות טובות באותה המידה).

$$\chi^2=\frac{(5-12\frac{2}{3})^2}{12\frac{2}{3}}+\frac{(13-12\frac{2}{3})^2}{12\frac{2}{3}}+\dots$$

נניח שיצא $\chi^2=16.83$ (לא בדיוק נכון למספרים הללו)

expected:

	א	ב	ג
רב	$12\frac{2}{3}$	$12\frac{2}{3}$	$12\frac{2}{3}$
כלשהו	29	29	29
ללא	$8\frac{1}{3}$	$8\frac{1}{3}$	$8\frac{1}{3}$

אזור הדחייה:

$$C_\alpha=\{\chi^2\geq\chi^2_{4,0.05}\}=\{\chi^2\geq 9.49\}$$

חישוב P-value:

$$P(\chi^2_4\geq 16.83)=0.002$$

מבחן χ^2 של פירסון לטיב התאמה

נניח שיש לנו מ"מ דיסקרטי X עם ערכים $x_1,\dots,x_k$, ונניח שרוצים לבדוק $H_0: X\sim F(=multi(p_1,\dots p_k))$ כאשר F היא התפלגות מולטינומית נתונה על $x_1,\dots,x_k$.

נניח כעת אני עושה מדגם מקרי בגודל n ורואים n_1 פעמים x_1, ... n_k פעמים x_k.
$n_1+\dots+n_k=n$

ניתן לחשוב על זה כעל טבלה עם שורה אחת

	x_1	x_2	...	x_k	סה"כ
	n_1	n_2	...	n_k	n
expected	np_1	np_2	...	np_k	n

25.1.09 (5)

באנלוגיה לטבלה הדו-מימדית, אם $np_1, \dots, np_k \ge 5$ כולם אזי $\sum_{i=1}^{n}\frac{(n_i-np_i)^2}{np_i} \underset{H_0}{\overset{\cdot}{\sim}} \chi^2_{k-1}$

שלושה נושאים

א. השוואות מרובות - נניח שיש לנו אוסף גדול של השערות אותן שנרצה לבדוק בו זמנית.

נ.ב. לא יהיה במבחן למרות שזה נושא חשוב אך ניתן לקרוא לבד עליו לא נחשב

דוגמה חשובה במיוחד: בחיפוש אחר גורמים גנטיים למחלות, משווים חולים לבריאים מבחינת רמת ההתבטאות של 5000 גנים.

לכל גן יש השערת 0: H_0: לא משפיע ⇔ רמת ההתבטאות בחולים ובבריאים דומה.

למשל: $H_0: \mu_{ח}^{(i)} = \mu_{ב}^{(i)}$

אם נבדוק כל H_0 ברמה $\alpha=0.05$ ונניח שאין אף גן שקשור למחלה, צפויין $0.05\times 5000 = 250$ השערות "דחו בממוצע".

לצורך בעיית הגילוי השגוי False Discover. כדי להתמודד, עושים תיקוני השוואות מרובות ל-α. הנאיבי ביותר: תיקון בונפרוני $\alpha = \frac{\alpha}{\text{מספר הבדיקות}}$.

במקרה שלנו: אם נבדוק את כל ההשערות ברמה $\frac{0.05}{5000}$.

ב. מבחנים סטטיסטיים לא-פרמטריים - עד עכשיו הנחנו (בד"כ) $X\sim F_\theta$, $Y\sim G_\psi$ ורצינו $H_0: \theta=\psi$ (לדוגמה, הנחת נורמליות המובילה למבחן t).

מה קורה אם נרצה לבדוק שוויון התפלגויות בלי הנחות פרמטריות? כלומר $X\sim F$, $Y\sim G$. $H_0: F=G$ (יכול להיות נורמלי, פואסוני, מעריכי התפלגות כלשהי)

הרעיון הכללי: נמצא סטטיסטי שאנו יודעים את התפלגותו תחת H_0 לכל F.

ג. גישה בייסיאנית - כאשר אנו אומדים או בודקים השערות על פרמטר θ, הנחנו כל הזמן ש-θ הוא מספר ולא משתנה מקרי.

הגישה הבייסיאנית:

א. הפרמטרים הם גם משתנים מקריים.

ב. יש התפלגות של הפרמטר לפני שראינו הנתונים, הנקראת התפלגות מקדימה (a-priori), ומבטאת את ה"אמונה" שלנו לגבי θ לפני שראינו הנתונים. נסמנה כ-$P(\theta)$.

ג. בהינתן הנתונים, אנחנו "מעדכנים" התפלגות θ ומקבלים התפלגות א-פוסטריורית.
כלומר: נתון $P(\theta)$, ומשפחה פרמטרית $F(X|\theta)$ עם צפיפות $f(X|\theta)$.

$$Pr(\theta|X) \underset{\text{הגדרה}}{=} \frac{P(x,\theta)}{P(x)} \underset{\text{בייס}}{=} \frac{f(x|\theta)P(\theta)}{P(x)} \propto f(x|\theta)P(\theta).$$

יתרונות: א. אם יש לנו ידע טוב על ערכי θ אזי $P(\theta)$ יהיה טוב ונרוויח מגישה זו.

ב. ניתן לפרש הסקה הסתברותית על θ ישירות כי הוא מ"מ.

חסרונות: מה אם לא יודעים כלום על θ מראש?

25.1.09 6

פ. קשה מאוד מבחינה חישובית לעיתים.

<u>דוגמא:</u> נניח $X|\mu \sim N(\mu, \sigma_x^2)$, σ_x^2 לא ידוע.

עושים מדגם בגודל n, ומעריכים את μ.

<u>שכיחותי:</u> $\hat{\mu}_{MLE} = \bar{X}$, רווח סמך: $[\bar{x} \pm Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}]$

<u>בייסיאני:</u> נניח $\mu \sim N(0, \sigma_\mu^2)$, σ_μ^2 גם ידועה.

(פרשנות: אני מצפה למצוא את μ קרוב ל-0)

$$P(\mu|X) \propto \underbrace{P(\mu)}_{\text{נורמלי }(0,\sigma_\mu^2)} \underbrace{f(X|\mu)}_{\text{נורמלי }(\mu,\frac{\sigma_x^2}{n})}$$

עושים חישובים ומקבלים (במקרה $n=1$)

$$\mu|X \sim N\left(\frac{\sigma_\mu^2}{\sigma_x^2+\sigma_\mu^2}X, \frac{\sigma_x^2 \cdot \sigma_\mu^2}{\sigma_x^2+\sigma_\mu^2}\right)$$

לכל n: $\mu|X \xrightarrow[\sigma_\mu^2 \to \infty]{D} N(X, \frac{\sigma_x^2}{n})$