סטטיסטיקה למדעי המחשב - שיעור 11

שיעור חזרה - יום ד' 30/1 10-12 שרייבר 6.

הסקה על שיפוע רגרסיה

אנחנו מניחים $Y|X \sim N(a+bX, \sigma^2)$

$H_0: b=0$ (בד"כ $H_A: b\neq 0$)

אומד הריבועים הפחותים $\hat{b} = \frac{\Sigma (x_i-\bar{x})(y_i-\bar{y})}{\Sigma (x_i-\bar{x})^2}$

אמרנו בשיעור הקודם: $\hat{b} \sim N\left(b, \frac{\sigma^2}{\Sigma (x_i-\bar{x})^2}\right)$

$H_0: \hat{b} \sim N\left(0, \frac{\sigma^2}{\Sigma (x_i-\bar{x})^2}\right)$

בד"כ נניח ש- σ לא ידוע ונעריך $\hat{\sigma}^2 = \frac{\Sigma (y_i-\hat{y})^2}{n-2}$ (אומד ל- σ^2)

$\Rightarrow$ ומכאן, $\frac{\hat{b}}{\hat{\sigma}/\sqrt{\Sigma(x_i-\bar{x})^2}} \sim t_{n-2}$ (התפלגות t) (העריכו את השונות ולכן אנחנו צריכים להשתמש בהתפלגות t).

ולכן, אזור דחיה עבור $H_0: b=0$ הוא: $\left|\frac{\hat{b}}{\hat{\sigma}/\sqrt{\Sigma(x_i-\bar{x})^2}}\right| \geq t_{n-2, 1-\frac{\alpha}{2}}$.

(ערך מהטבלה)

מבחנים לקשר בין 2 התפלגויות/משתנים

המצב הכללי: $X \sim F$, ל- F יש פרמטר(ים) שמסומנים ב- θ

$Y \sim G$. ל- G יש פרמטר(ים) שמסומנים ב- ψ.

בד"כ, השערת האפס $H_0: \theta = \psi$.

דוגמאות:

$X \sim N(\mu_x, \sigma_x^2)$, $Y \sim N(\mu_y, \sigma_y^2)$.

$\theta = \mu_x$ $\psi = \mu_y$

$H_0: \mu_x = \mu_y$, $H_A: \gtrless \neq$.

נניח בה"כ $H_A: \mu_x > \mu_y$.

/* כשאנו אומרים מדגם מזווג אנו מניחים אי תלות בין ערכי X (לעצמם) ובין ערכי Y לעצמם
זה לא מספיק לנו שאין תלות בין ערכי X לערכי Y */

נתונים מדגמים מקריים:

$\underline{X} = (X_1, \ldots, X_n)$ (m ו- n יכולים להיות שונים).

$\underline{Y} = (Y_1, \ldots, Y_m)$

שאלה 1: האם $\bar{X}, \bar{Y}$ ב"ת ביניהם?

אם כן, נבצע מבחן לא מזווג לשני מדגמים.

אם לא, המקרה המעניין הוא $m=n$ ו X_i, Y_i הן תצפיות על אותו אובייקט (כלומר, X_i, Y_i תלויים אבל X_i, Y_j ב"ת עבור $j \neq i$). במקרה זה, נעשה מבחן למדגם מזווג.

1. מבחנים לא מזווגים ל-2 מדגמים

$$\begin{cases} X_1, \dots, X_n \\ Y_1, \dots, Y_m \end{cases} \text{ נ"מ ב"ת}$$

$$\bar{Y} \sim N\left(\mu_y, \frac{\sigma_y^2}{n}\right) \quad , \quad \bar{X} \sim N\left(\mu_x, \frac{\sigma_x^2}{n}\right)$$

$$\bar{X} - \bar{Y} \sim N\left(\underbrace{\mu_x - \mu_y}_{H_0 \text{ תחת } 0}, \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{n}\right)$$

אפשר לחשוב שמבחני לבדיקת השערות על $\bar{Z} = \bar{X} - \bar{Y}$: $H_0: E(\bar{Z}) = 0$

קיים UMP כמו במקרה הנורמלי הרגיל

א. אם השונויות ידועות (תחת האלטרנטיבה $H_A: \mu_x > \mu_y$)

מבחן ברמה α:

$$C_\alpha^* = \left\{ \bar{X} - \bar{Y} \geq Z_{1-\alpha} \sqrt{\frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}} \right\}$$

ב. מה קורה אם השונויות לא ידועות?

1. שוות: $\sigma_x^2 = \sigma_y^2$ לא ידוע.

2. לאו דווקא שוות: σ_x^2 לא ידוע, σ_y^2 לא ידוע.

1. שונויות שוות: $\sigma_x^2 = \sigma_y^2 = \sigma^2$

$$Var(\bar{X} - \bar{Y}) = \frac{\sigma^2}{n} + \frac{\sigma^2}{m} = \frac{(m+n)}{mn} \sigma^2$$

כאן הנחת השוויון, נעריך את σ^2 משני המדגמים יחד

$$\left(\text{אומד } \sigma^2\right) \quad \hat{\sigma}_{xy}^2 = \frac{\sum_{i=1}^{n} (X_i - \bar{X})^2 + \sum_{i=1}^{m} (Y_i - \bar{Y})^2}{n+m-2} \underset{1}{=} \frac{(n-1)\hat{\sigma}_x^2 + (m-1)\hat{\sigma}_y^2}{n+m-2}$$

$$\hat{\sigma}_x^2 = \frac{\sum (X_i - \bar{X})^2}{n-1}$$

כאן נעשה מבחן t, עם $n+m-2$ דרגות חופש:

$$C_\alpha = \left\{ \bar{X} - \bar{X} \geq t_{n+m-2, 1-\alpha} \sqrt{\frac{m+n}{mn} \hat{\sigma}_{xy}^2} \right.$$

סטטיסטיקה - שיעור 11

2. לא מניחים שונויות שוות:

$$\sigma^2_{\bar{x}-\bar{y}} := Var(\bar{x}-\bar{y}) = \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m} \quad \Rightarrow \quad \hat{\sigma}^2_{x-y} = \frac{\sigma_x^2}{n} + \frac{\sigma_y^2}{m}$$

<u>שאלה:</u> מה מספר דרגות החופש של $\hat{\sigma}^2_{\bar{x}-\bar{y}}$ (וכן של הסטטיסטי t שמחושב ממנו).

<u>תשובה:</u> הקירוב המקובל הוא תיקון Welch-Satterwhite

$$v = \frac{\left(\hat{\sigma}^2_{\bar{x}-\bar{y}}\right)^2}{\frac{\hat{\sigma}_x^4}{n^2(n-1)} + \frac{\hat{\sigma}_y^4}{m^2(m-1)}} \Rightarrow \frac{\bar{x}-\bar{y}}{\sqrt{\hat{\sigma}^2_{\bar{x}-\bar{y}}}} \sim t_v$$

אזור דחיה:

$$C_\alpha = \left\{ \bar{x}-\bar{y} \geq t_{v,1-\alpha} \sqrt{\hat{\sigma}^2_{\bar{x}-\bar{y}}} \right\}$$

<u>דוגמה לכל שלושת המקרים</u>

$H_0: \mu_{\text{מחשבים}} = \mu_{\text{מה"ר}}$
$H_A: \mu_{\text{מחשבים}} > \mu_{\text{מה"ר}}$

האם סטודנטים למחשבים יותר גבוהים מסטודנטים למדעי הרוח?

נניח $n=25$ סטודנטים מקריים ממחשבים $\bar{x}=175$ ו $m=20$ סטודנטים מקריים ממדעי הרוח $\bar{y}=172$

א. נניח $\sigma_x^2=16$ $\sigma_y^2=18$

ב. נראה $\hat{\sigma}_x^2=16$ $\hat{\sigma}_y^2=18$, ונניח שוויון שונויות.

ג. " " " , ולא נניח שוויון שונויות.

<u>פתרון</u>

א.
$$Var(\bar{x}-\bar{y}) = \frac{16}{25} + \frac{18}{20} = 1.54$$

$$std(\bar{x}-\bar{y}) = 1.24$$

$$N(0,1) \underset{H_0}{\sim} Z = \frac{\bar{x}-\bar{y}}{1.24} = \frac{3}{1.24} = 2.42$$

$$P\text{-val}: 1-\Phi(2.42) = 0.008$$

לכן, נדחה H_0 ברמה $\alpha=0.01$

ב.
$$\hat{\sigma}^2_{xy} = \frac{(n-1)\hat{\sigma}_x^2 + (m-1)\hat{\sigma}_y^2}{n+m-2} = 16.89$$

$$Var(\bar{x}+\bar{y}) = \frac{\hat{\sigma}^2_{xy}}{m} + \frac{\hat{\sigma}^2_{xy}}{n} = \frac{m+n}{mn}\hat{\sigma}^2_{xy} = \frac{45}{500} \cdot 16.89 = 1.52$$

$$t_{n+m-2} \underset{H_0}{\sim} t = \frac{3}{1.23}$$

$$P\text{-value}: P\left(t_{n+m-2} \geq \frac{3}{1.23}\right) = P\left(t_{43} \geq \frac{3}{1.23}\right) = 0.01004$$

$\Leftarrow$ לא נדחה H_0 ברמה $\alpha=0.01$,

ג. לא הנחנו שוויון שונויות: תיקון welch ל-df, $\hat{\sigma}^2_{\bar{x}-\bar{y}} = \frac{\hat{\sigma}_x}{n} + \frac{\hat{\sigma}_y}{m} = 1.54$.

$$V = \frac{(1.54)^2}{\frac{16^2}{25^2 \times 24} + \frac{18^2}{20^2 \times 18}} = \ldots = 39.72 \Rightarrow \frac{\bar{x}-\bar{y}}{\sqrt{1.54}} \overset{H_0}{\sim} t_V$$

$$\text{P-value}: 1 - P(t_{39.72} \le 2.42) = 0.00998\ldots$$

⇐ נדחה את H_0 ברמת $\alpha = 0.01$.

__הסקה במדגמים מזווגים__

נזכיר: $m=n$, X_i, Y_i אינם בת"ל זה בזה. (נניח X_i בת"ל ב-X_j $\forall i \neq j$ וכן Y_i בת"ל ב-Y_j $\forall i \neq j$)

לדוגמה: שני מדדים ניתנו לכל נבדק.

$$H_0: \mu_x = \mu_y \qquad H_A: \mu_x > \mu_y$$

מה ניתן להגיד על $Z_i = X_i - Y_i$? ⇐ מדגם מקרי מ-$N(\mu_x - \mu_y, \sigma^2_{x-y})$.

⇐ $H_0: \mu_x - \mu_y = \mu_z = 0$; $H_A: \mu_z > 0$

__שימו לב:__ $\sigma^2_{x-y} \neq \sigma^2_x + \sigma^2_y$

$$Var(x-y) = Var(x) + Var(y) - 2Cov(x,y)$$

הסקה: __א. אם__ σ^2_{x-y} ידוע אז UMP: $C^*_\alpha = \{\bar{x} - \bar{y} \ge Z_{1-\alpha}\sqrt{\frac{\sigma^2_{x-y}}{n}}\}$

__ב. אם__ σ^2_{x-y} לא ידוע, נעריך $\hat{\sigma}^2_{x-y} = \hat{\sigma}^2_z = \frac{\sum(z_i - \bar{z})^2}{n-1}$.

$$C_0^* = \{\bar{Z} = \bar{x} - \bar{y} \ge t_{n-1,1-\alpha}\sqrt{\frac{\hat{\sigma}^2_{x-y}}{n}}\}$$

__שאלה:__ מה קורה אם שוכחים לזווג מדגם מזווג?

__תשובה כללית:__ בד"כ נצפה ש- $Cov(X_i, Y_i) \ge 0$, $Var(X-Y) << Var(X) + Var(Y)$

⇐ יותר קל לדחות השערת ה-0 תחת זיווג (נכון!)

__דוגמה מספרית:__ אחים ואחיות באותה משפחה

$n = 10$ $\quad \bar{X} = 175$ $\quad \bar{Y} = 170$ $\quad \sigma_x^2 = 125$ $\quad \sigma_y^2 = 125$ $\quad Cov(x,y) = 80$.

__לא מזווג:__ נקבל סטטיסטי Z כך:

$$Var(\bar{X} - \bar{Y}) = \frac{125}{10} + \frac{125}{10} = 25$$

$$\Rightarrow Z = \frac{5}{\sqrt{25}} = 1 \Rightarrow \text{P-Value} = 1 - \Phi(1) = 0.16$$

__מזווג:__

$$Var(\bar{X} - \bar{Y}) = \frac{125}{10} + \frac{125}{10} - 2 \times \frac{80}{10} = 9$$

$$Z = \frac{5}{\sqrt{9}} = \frac{5}{3} \Rightarrow \text{P-Value} = 1 - \Phi\left(\frac{5}{3}\right) = 0.048$$

סטטיסטיקה - שיעור 11

⇐ ראינו שאם $cov>0$ ולא מזווגים אזי "מפסידים עוצמה".

<u>שאלה:</u> מה קורה אם $cov=0$ ומזווגים (נניח $n=m$)?

הערכת השונות תחת הנחת הזיווג:

$$\hat{\sigma}^2_{X-Y} = \frac{\sum (z_i-\bar z)^2}{n-1} = \frac{\sum[(x_i-y_i)-(\bar x-\bar y)]^2}{n-1} = \underbrace{\frac{\sum (x_i-\bar x)^2}{n-1}}_{\hat\sigma^2_x} + \underbrace{\frac{\sum (y_i-\bar y)^2}{n-1}}_{\hat\sigma^2_y} - 2\underbrace{\frac{[(x_i-\bar x)(y_i-\bar y)]}{n-1}}_{2\widehat{cov}(X,Y)}$$

כעת נעשה מבחן t עם $n-1$ ד"ח $\frac{\bar X-\bar Y}{\sqrt{\hat\sigma^2_{X-Y}/n}} \underset{H_0}{\sim} t_{n-1}$

אם לא היינו מזווגים:

$$\frac{\bar X-\bar Y}{\sqrt{\frac{\hat\sigma^2_x}{n}+\frac{\hat\sigma^2_y}{n}}} \underset{H_0}{\sim} t_{2n-2}$$

⇐ שני המבחנים הם תקן ברמה α, המבחן השני יותר "נכון" וינסה להיות בעל עוצמה גדולה יותר (בעל $2n-2$ ד"ח).

מה לא הסתדר של שאלה * במבחן לדוגמה.

	לא מזווגים	מזווגים
x_i, y_i ב"ת	✓	<u>נכון</u> בגלל שהרמה עדיין α מפסידים עוצמה
x_i, y_i תלוים	<u>לא נכון</u> הרמה לא α כי $cov>0$ [illegible] והשונות	✓

$$\frac{\bar X-\bar Y}{\sqrt{Var(\bar X-\bar Y)}} \sim t_{2n-2}$$

<u>הסקה על הפרש בין התפלגויות בינומיות</u>

במקרה הבינומי (לא מזווג) $X\sim Bern(p_x)$ n תצפיות, $Y\sim Bern(p_y)$ m תצפיות (לא ה- $n\neq m$ בה"כ)

$$H_0: p_x=p_y \;;\; H_A: p_x>p_y$$

$$\hat p_x-\hat p_y \sim N\left(p_x-p_y,\ \frac{p_x(1-p_x)}{n}+\frac{p_y(1-p_y)}{m}\right)$$

תחת H_0: $p_x=p_y=p$:

$$\hat p_x-\hat p_y \underset{H_0}{\sim} N\left(0,\ \frac{p(1-p)}{n}+\frac{p(1-p)}{m}\right)$$

<u>תזכורת:</u> כאשר עשינו מבחן בינומי ל- $H_0: p=p_0$. אזי השונות תחת H_0 היתה ידועה $\frac{p_0(1-p_0)}{n}$.

במקרה הנוכחי H_0 אינה <u>פשוטה</u> ולכן אנחנו צריכים לאמוד את השונות.

סטטיסטיקה - שיעור 11

1. הרכבתקני:

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

$$\hat{Var}(\hat{p}_x - \hat{p}_y) = \frac{\hat{p}(1-\hat{p})}{n} + \frac{\hat{p}(1-\hat{p})}{m}$$

2. שמרני:

$$\hat{Var}(\hat{p}_x - \hat{p}_y) = \frac{1}{4n} + \frac{1}{4m}$$

אזור הדחייה:

$$C_\alpha = \{ \hat{p}_x - \hat{p}_y \geq Z_{1-\alpha} \sqrt{\hat{Var}(\hat{p}_x - \hat{p}_y)} \}.$$

דוגמה: האם אחוז הקבלה של בנים למגמה גבוה מזה של בנות?

$n=200$ בנים, מתוכם $n_+=50$ התקבלו $(p_x=0.25)$

$m=50$ בנות, מתוכן $m_+=10$ התקבלו $(p_y=0.2)$

$$\hat{p}_x - \hat{p}_y = 0.05, \quad Var(\hat{p}_x - \hat{p}_y) = 0.00456$$

$$Z = \frac{0.05}{\sqrt{0.00456}} = 0.74$$

$$\text{P-value}: \quad 1 - \Phi(0.74) = 0.23 \Rightarrow H_0 \text{ לא נדחה}$$

מה קורה אם מס' ההצלחות (הכל) $n_+ + m_+$ ידוע מראש?

דוגמה: נקבע מראש ש-60 סטודנטים "יתקבלו": סדר השארת ה-0 הוא מדין "שפיטה" הקבלה שווה לכל X,Y (בדוגמה שווים לפי מין).

שאלה: בהינתן n,m, $n_+ + m_+$ (מס' ההצלחות). איך מתפלג n_+ תחת H_0?
אזי $n_+ \underset{H_0}{\sim} HG(n+m, n_+ + m_+, n)$

	הצלחה	כישלון	סה"כ
X	n_+	$n-n_+$	n
Y	m_+	$m-m_+$	m
סה"כ	$n_+ + m_+$	$n+m-n_+-m_+$	$m+n$

איך אפשר להסיק לגבי נכונות H_0?

1. הסקה מדויקת בעזרת התפלגות HG (Fisher Exact Test)

2. קירוב (נורמלי) $n_+ \underset{H_0}{\sim} N\left(\frac{n(n_+ + m_+)}{n+m}, \frac{n(n_+ + m_+)(n+m-n_+-m_+)m}{(n+m)^2(n+m-1)}\right)$

במקרה שלנו: $n_+ + m_+ = 60$; $n_+ \sim N(48, 7.33)$

$$\text{P-Val}: 1 - \Phi\left(\frac{50-48}{\sqrt{7.33}}\right) = 0.23$$