

# סטטיסטיקה – שיעור 10

## בדיקת השערות

### הלמה של ניימן פירסון

בוחנים:  $H_0: \theta = \theta_0, H_A: \theta = \theta_A$  על פי מדגם בגודל  $n$ .  
המבחן בעל עוצמה מקסימלית ברמה  $\alpha$  הוא:

$$\Lambda(\underline{x}) \leq d_\alpha^* \Leftrightarrow H_0 \text{ את דחה}$$

$$P_{H_0}(\Lambda(X) \leq d_\alpha^*) = \alpha \text{ כאשר } d_\alpha^* \text{ נקבע כך:}$$

### הוכחה:

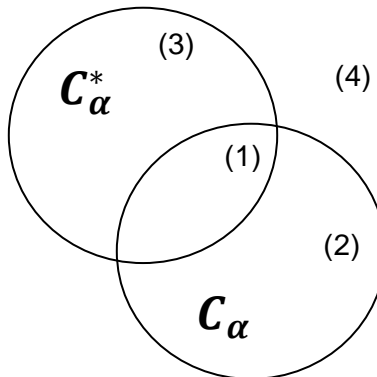
$$P_{H_0}(X \in C_\alpha^*) = \alpha \text{ ונניח } C_\alpha^* = \{\underline{x}: \Lambda(\underline{x}) \leq d_\alpha^*\}$$

יהי  $C_\alpha$  איזור דחייה אחר ברמה  $\alpha$ , ז"א:

$$P_{H_0}(X \in C_\alpha) = \alpha$$

נסתכל על שתי הצגות גרפיות של היחס בין  $C_\alpha^*, C_\alpha$ :

	$\underline{x} \in C_\alpha^*$	$\underline{x} \notin C_\alpha^*$
$\underline{x} \in C_\alpha$	(1) שני המבחנים דחו את $H_0$	(2) מבחן NP לא יידחה את $H_0$ , המבחן $C_\alpha$ ידחה את $H_0$
$\underline{x} \notin C_\alpha$	(3) מבחן NP יידחה את $H_0$ , המבחן $C_\alpha$ לא ידחה את $H_0$	(4) שני המבחנים לא דחו את $H_0$



לפי ההגדרות:

$$\alpha = P_{H_0}(X \in C_\alpha^*) = \underbrace{P_{H_0}(X \in C_\alpha^* \cap X \in C_\alpha)}_1 + \underbrace{P_{H_0}(X \in C_\alpha^* \cap X \notin C_\alpha)}_3$$

$$\alpha = P_{H_0}(X \in C_\alpha) = \underbrace{P_{H_0}(X \in C_\alpha^* \cap X \in C_\alpha)}_1 + \underbrace{P_{H_0}(X \notin C_\alpha^* \cap X \in C_\alpha)}_2$$

$$\Rightarrow P_{H_0}((2)) = P_{H_0}((3))$$

$$C_\alpha^* \text{ עוצמת} = P_{H_A}(X \in C_\alpha^*) = \underbrace{P_{H_A}(X \in C_\alpha^* \cap X \in C_\alpha)}_1 + \underbrace{P_{H_A}(X \in C_\alpha^* \cap X \notin C_\alpha)}_3$$

$$C_\alpha \text{ עוצמת} = P_{H_A}((1)) + P_{H_A}((2))$$

$$\begin{aligned} C_\alpha^* \text{ עוצמת} - C_\alpha \text{ עוצמת} &= P_{H_A}((1)) + P_{H_A}((3)) - P_{H_A}((1)) - P_{H_A}((2)) \\ &= \underbrace{P_{H_A}(X \in C_\alpha^* \cap X \notin C_\alpha)}_3 - \underbrace{P_{H_A}(X \notin C_\alpha^* \cap X \in C_\alpha)}_2 \stackrel{::}{\geq} \end{aligned}$$

הגדרת  $\{x: \Lambda(x) \leq d_\alpha^*\} : C_\alpha^*$

$$\Rightarrow P_{H_A}((3)) = \int_{\underline{x} \in C_\alpha^*, \underline{x} \notin C_\alpha} f_{\theta_A}(\underline{x}) d\underline{x} \stackrel{f_{\theta_A}(\underline{x}) \geq \frac{f_{\theta_0}(\underline{x})}{d}}{\geq} \frac{1}{d_\alpha^*} \int_{\underline{x} \in C_\alpha^*, \underline{x} \notin C_\alpha} f_{\theta_0}(\underline{x}) d\underline{x}$$

$$\Rightarrow P_{H_A}((2)) = \int_{\underline{x} \notin C_\alpha^*, \underline{x} \in C_\alpha} f_{\theta_A}(\underline{x}) d\underline{x} < \frac{1}{d_\alpha^*} \int_{\underline{x} \notin C_\alpha^*, \underline{x} \in C_\alpha} f_{\theta_0}(\underline{x}) d\underline{x}$$

$$\stackrel{::}{\geq} \frac{1}{d_\alpha^*} \left( \frac{P_{H_0}((3))}{\alpha - P_{H_0}((1))} - \frac{P_{H_0}((2))}{\alpha - P_{H_0}((1))} \right) = 0$$

$\Rightarrow$  ל.ש.מ

### מבחנים בעלי עצמה מקסימלית לפי NP

a. נורמלי עם שונות ידועה:  $H_0: \mu = \mu_0, H_A: \mu = \mu_A (\mu_A > \mu_0)$

בשיעור שעבר ראינו כי  $C_\alpha^* = \{\bar{X}: \bar{X} \geq \mu_0 + Z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}\}$  לפי NP.

אם  $H_0: \mu = \mu_0, H_A: \mu = \mu_A (\mu_A > \mu_0)$  אז  $C_\alpha^* = \{\bar{X}: \bar{X} \leq \mu_0 - Z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}\}$

b. נורמלי עם שונות  $\sigma^2$  לא ידועה:  $H_0: \mu = \mu_0, H_A: \mu = \mu_A (\mu_A > \mu_0)$

$\Leftarrow$  יחס הנראות עדיין מונוטוני.

$C_\alpha^* = \{\bar{X}: \bar{X} \geq \mu_0 + t_{n-1, 1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}\} \Leftarrow \hat{\sigma}^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2$  כאשר

אם  $H_0: \mu = \mu_0, H_A: \mu = \mu_A (\mu_A < \mu_0)$  אז  $C_\alpha^* = \{\bar{X}: \bar{X} \leq \mu_0 - t_{n-1, 1-\alpha} \cdot \frac{\sigma}{\sqrt{n}}\}$

c. ברנולי:  $H_0: p = p_0, H_A: p = p_A (p_A > p_0)$

$$f_{p_0}(\underline{x}) = p_0^{\sum X_i} (1 - p_0)^{n - \sum X_i}$$

$$f_{p_A}(\underline{x}) = p_A^{\sum X_i} (1 - p_A)^{n - \sum X_i}$$

$$\Lambda(\underline{x}) = \left(\frac{p_0}{p_A}\right)^{\sum X_i} \left(\frac{1 - p_0}{1 - p_A}\right)^{n - \sum X_i} = \left(\frac{p_0(1 - p_A)}{p_A(1 - p_0)}\right)^{n\hat{p}} \cdot \left(\frac{1 - p_0}{1 - p_A}\right)^n$$

$$\log \Lambda(\underline{x}) = n\hat{p} \log \left(\frac{p_0(1 - p_A)}{p_A(1 - p_0)}\right) + n \cdot \log \left(\frac{1 - p_0}{1 - p_A}\right)$$

$$\frac{\partial}{\partial \hat{p}} \log \Lambda(\underline{x}) = n \cdot \underbrace{\log \left(\frac{p_0(1 - p_A)}{p_A(1 - p_0)}\right)}_{< 0} < 0$$

$\Leftarrow$  מבחן NP ל-  $H_0: p = p_0, H_A: p = p_A (p_A > p_0)$  הוא מהצורה  $C_\alpha^* = \{\underline{x}: \hat{p} \geq d_\alpha^*\}$

איך קובעים את אזור הדחייה? (מוציאים את  $d_\alpha^*$ )

i. נקבע  $d$  כלשהו ואז נאמר:

המבחן שדוחה את  $\hat{p} \geq d$  הוא מבחן בעל עוצמה מקסימלית ברמה:

$$\begin{aligned} P_{H_0}(\hat{p} \geq d) &= P_{H_0}\left(\sum_{i=1}^n X_i \geq dn\right) = \sum_{z=\lceil nd \rceil}^n P_{H_0}\left(\sum_{i=1}^n X_i = z\right) \\ &= \sum_{z=\lceil nd \rceil}^n \binom{n}{z} p_0^z (1-p_0)^{n-z} \end{aligned}$$

ii. נרצה לקבוע את  $\alpha$  ולהשתמש בקירוב נורמלי,  $\hat{p}_{H_0} \sim N\left(p_0, \frac{p_0(1-p_0)}{n}\right)$ , ולכן ניקח:

$$C_\alpha^* = \left\{ \underline{x}: \hat{p} \geq p_0 + Z_{1-\alpha} \sqrt{\frac{p_0(1-p_0)}{n}} \right\}$$

$\Leftarrow$  מבחן בעל עוצמה מקסימלית ברמה "בקירוב"  $\alpha$ .

d. פואסוני.  $X \sim Pois(\lambda)$ ,  $H_0: \lambda = \lambda_0$ ,  $H_A: \lambda = \lambda_A$  ( $\lambda_A > \lambda_0$ )

הערה: סכום פואסונים מתפלג פואסוני.

לכן: שקול להגיד  $X_i \sim Pois(\delta)$ , ואנו רוצים לבדוק  $H_0: \delta = \delta_0$ ,  $H_A: \delta = \delta_A$  לפי מדגם בגודל  $n$

או להגיד: ניקח  $X = \sum_{i=1}^n X_i$ , ונניח  $X \sim Pois(\lambda)$ , אז  $H_0: \lambda = n\delta_0$ ,  $H_A: \lambda = n\delta_A$

$$\Lambda(x) = e^{-(\lambda_A - \lambda_0)} \left(\frac{\lambda_0}{\lambda_A}\right)^x$$

$$\frac{\partial}{\partial x} \log(\Lambda(x)) = \log\left(\frac{\lambda_0}{\lambda_A}\right) < 0$$

$\Leftarrow$  איזור הדחייה למבחן עצמה מקסימלית:  $C_\alpha^* := \{x: x \geq d_\alpha^*\}$

## הכללת NP

מבחנים בעלי עוצמה מקסימלית במידה שווה (UMP)

נניח אנחנו רוצים לבחון:  $H_0: \theta = \theta_0$ ,  $H_A: \theta \in H_A$

אזור דחייה  $C_\alpha^*$  מגדיר מבחן ברמה  $\alpha$  בעל עוצמה מקסימלית במידה שווה עם:

$$P_{H_0}(X \in C_\alpha^*) = \alpha \quad .a$$

$$H_A: \theta \in \theta_A, \forall \theta_A \in H_A \quad .b$$

## דוגמאות

a. נורמלי, שונות ידועה  $H_0: \mu = \mu_0$ ,  $H_A: \mu > \mu_0$

במקרה הפשוט:  $H_A: \mu = \mu_A$  קיבלנו לפי NP איזור דחייה שאינו תלוי ב- $\mu_A$ :

$$C_\alpha^* = \left\{ \underline{x}: \bar{X} \geq \mu_0 + Z_{1-\alpha} \cdot \frac{\sigma}{\sqrt{n}} \right\}$$

מכיוון שאינו תלוי ב- $\mu_A$  הוא בעל עוצמה מקסימלית מול כל:

$$\mu_A \in H_A = \{\mu: \mu > \mu_0\}$$

$\Leftarrow C_\alpha^*$  ניתן UMP.

b-d: באותו באופן, המבחנים שנתנו בשיעור הקודם הם UMP עבור השארות מורכבות חד-צדדיות למקום:

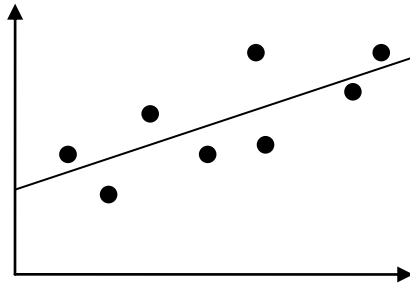
- נורמלי,  $\sigma^2$  לא ידוע
- בנומי
- פואסוני

במקרה הדו צדדי: מבחן MP תלוי בכיוון האלט' הפשוטה. לכן אם האלטרנטיבה המורכבת כוללת את שני הכיוונים אין UMP.

### הסקה סטטיסטית בבעיות רגרסיה

נזכר בבעיית הרגרסיה:

- a. נתונות תצפיות  $(x_1, y_1), \dots, (x_n, y_n)$
- b. מעריכים מודל:  $\hat{y} = a + bx$



c. הדרך המקובלת: ריבועים פחותים.

נגדיר:  $f(a, b) = \sum_{i=1}^n (y_i - a - bx_i)^2$  כערכים שמביאים את  $f(a, b)$  למינימום. ניתן לחשוב על הבעיה כבעיית אמידה של  $a, b$

- $a, b$  הפ פרמטרים שאומדים
- מהו התאור של אוכלוסיה, התפלגות, פרמטרים שמתאים כדי לתאר את מציאת  $\hat{a}_{LS}, \hat{b}_{LS}$  כבעיית אמידה פורמלית?

תשובה: אנחנו מנסים לעשות הסקה סטטיסטית לגבי ההתפלגות המותנה  $P(Y|X)$ .

נניח  $(Y|X = x) = N(a + bx, \sigma^2)$  לא ידוע במילים אחרות:  $Y|X = a + bx + \varepsilon$  כאשר  $\varepsilon \sim N(0, \sigma^2)$ .

ואנחנו רוצים לעשות הסקה סטטיסטית לגבי  $a, b$  על פי מדגם מקרי בגודל  $n$ :

$$(x_1, y_1), \dots, (x_n, y_n)$$

איך אפשר לאמוד את  $a, b$ ? נראות מקסימלית.

$$L(a, b; (x_1, y_1), \dots, (x_n, y_n), \sigma^2) = \prod_{i=1}^n \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i - a - bx_i)^2}{2\sigma^2}}$$

$$= (2\pi\sigma^2)^{-\frac{n}{2}} \exp\left\{-\frac{\sum_{i=1}^n (y_i - a - bx_i)^2}{2\sigma^2}\right\}$$

$$l(a, b; \dots) = \underbrace{-\frac{n}{2} \log(2\pi\sigma^2)}_{\text{לא מכיל } a, b} - \frac{1}{2\sigma^2} \underbrace{f(a, b)}_{\text{סכום ריבועי השאריות}}$$

$l$  לא מכיל את  $f(a, b)$  בסימן מינוס.

$\Leftarrow$  להביא למקסימום את  $l$  שקול למינימום את  $f(a, b)$ .

$\Leftarrow$  א.נ.מ:

$$(\hat{a}, \hat{b}) = \underset{a, b}{\operatorname{argmin}} f(a, b)$$

נזכר בפתרון שנתנו בכיתה:

$$\hat{b} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$\hat{a} = \bar{Y} - \hat{b}\bar{X}$$

מודל:  $y_i = a + bx_i + \varepsilon_i$  כאשר  $\varepsilon_i \sim N(0, \sigma^2)$ .

מאחר שאנחנו במסגרת של הסקה סטטיסטית, מה עוד אנחנו יכולים לעשות?

- לעשות רווחי סמך ל- $a, b$  על סמך  $\hat{a}, \hat{b}$ .
- לבדוק השערות על  $a, b$ .

ההשערה הקלאסית במקרה זה:  $H_0: b = 0, H_A: b \neq 0$ .

בכדי לבצע הסקה זו, נשאל את עצמנו לגבי התפלגות  $\hat{b}$ .

$$\hat{b} = \frac{\sum_{i=1}^n (X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})Y_i - \bar{Y} \sum_{i=1}^n (X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2} = \frac{\sum_{i=1}^n (X_i - \bar{X})(a + bx_i + \varepsilon_i)}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$= \frac{a \sum_{i=1}^n (X_i - \bar{X}) + b \sum_{i=1}^n X_i(X_i - \bar{X}) + \sum_{i=1}^n \varepsilon_i(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}$$

$$E(\hat{b}) = b \cdot \underbrace{\frac{\sum_{i=1}^n X_i(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}}_{=1} + E\left(\frac{\sum_{i=1}^n \varepsilon_i(X_i - \bar{X})}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$

מהמודל שלנו  $E(\varepsilon_i) = 0 \forall i$ .

$$\sum X_i(X_i - \bar{X}) = \sum X_i^2 - \left(\sum X_i\right)\bar{X} = \sum X_i^2 - n\bar{X}^2 = \sum (X_i - \bar{X})^2 \Rightarrow E(\hat{b}) = b$$

$\hat{b}$  אומד ח"ה ל-b.

לאחר עוד חישובים שלא נראה כאן, נקבל:

$$\hat{b} \sim N\left(b, \frac{\sigma^2}{\sum_{i=1}^n (X_i - \bar{X})^2}\right)$$

איך נבחן עכשיו:

$H_0: b = 0, H_A: b \neq 0$

בעזרת  $\hat{b} \stackrel{b=b_0}{\sim} N\left(\hat{b}, \frac{\sigma^2}{\sum (X_i - \bar{X})^2}\right), \hat{b} \stackrel{H_0}{\sim} N\left(0, \frac{\sigma^2}{\sum (X_i - \bar{X})^2}\right)$ :

אומד חסר הטייה ל- $\sigma^2$ :  $\hat{\sigma}^2 = \frac{1}{n-2} \sum (y_i - \bar{y}_i)^2$

← איזור דחייה לבדיקת השערות דו כיווניות:

$$C_\alpha^* = \left[ -t_{n-2, 1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{\sum (X_i - \bar{X})^2}}, t_{n-2, 1-\frac{\alpha}{2}} \cdot \frac{\hat{\sigma}}{\sqrt{\sum (X_i - \bar{X})^2}} \right]$$