

## סטטיסטיקה – הרצאה 08

רווח סמך לתוחלת נורמלית עם שונות ידועה

נשתמש באומד  $\hat{\mu} = \bar{X}$ , ר"ס  $1 - \alpha$ :

$$\theta \in \left[ \bar{X} - Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}}, \bar{X} + Z_{1-\frac{\alpha}{2}} \frac{\sigma}{\sqrt{n}} \right]$$

שונות לא ידועה:

$$\hat{\sigma}^2 = \frac{1}{n-1} \sum (X_i - \bar{X})^2 \text{ אומדים שונות:}$$

$$\theta \in \left[ \bar{X} - t_{n-1, 1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}}, \bar{X} + t_{n-1, 1-\frac{\alpha}{2}} \frac{\hat{\sigma}}{\sqrt{n}} \right] \text{ ר"ס:}$$

$$\frac{N(0,1)}{\sqrt{\frac{k^2}{n-1}}} \sim t_{n-1} \text{ כאשר התפלגות } t$$

רווחי סמך לפרמטר בינומי

רוצים לאמוד הפרופורציה  $p$ , כאשר כזכור:

$X = (X_1, \dots, X_n)$ ,  $X \sim \text{Bernulli}(p)$  מדגם מקרי.

נבקש למצוא ר"ס על סמך הפרופורציה האמפירית  $\hat{p} = \frac{\sum X_i}{n}$

גישות אפשריות:

1. רווח סמך מדויק על פי ההת' הבינומית:

$$\begin{aligned} Pr(p \in [\hat{p} - \varepsilon, \hat{p} + \varepsilon]) &= Pr(\hat{p} \in [\hat{p} - \varepsilon, \hat{p} + \varepsilon]) = \sum_{k=(p-\varepsilon_2)=Pr(\hat{p}=\frac{k}{n})}^{(p+\varepsilon_1)n} Pr\left(\sum X_i = k\right) \\ &= \sum_{k=(p-\varepsilon_2)=Pr(\hat{p}=\frac{k}{n})}^{(p+\varepsilon_1)n} \binom{n}{k} p^k (1-p)^{n-k} \geq \underset{\text{דרישה מרווח סמך}}{1 - \alpha} \end{aligned}$$

זו גישה קשה ולא מבטיחה.

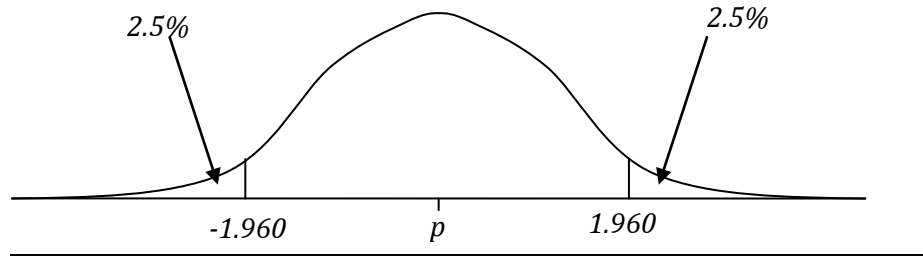
2. להשתמש בקרוב הנורמלי:

אמרנו שאם מס' ההצלחות והכשלונות שניהם מעל 10, אז ניתן להגדיר:

$$\frac{\hat{p} - p}{\sqrt{\frac{p(1-p)}{n}}} \sim N(0,1)$$

או:

$$\hat{p} \sim N\left(p, \frac{p(1-p)}{n}\right)$$



הבעיה: השונות מכילה את הפרמטר הלא ידוע  $p$ , שאותו אנחנו מנסים לאמוד.

### פתרון א:

נחליף את  $p$  ב-  $\hat{p}$  גם בשונות.

$$\text{נסמן } SE(p) = \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, SD(\hat{p}) = \sqrt{\frac{p(1-p)}{n}}$$

ונגיד

$$\frac{\hat{p} - p}{\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}} \sim N(0,1)$$

הסימן  $\sim$  מכיל 2 רמות של אי דיוק:

a. הקירוב הנורמלי לבינומי

b. שימוש באומד ל- $p$  במקום  $p$  במכנה.

פתרון לאי דיוק a:

$$Pr\left(p \in \left[\hat{p} - Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}, \hat{p} + Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\right]\right) \cong 1 - \alpha$$

### דוגמא:

אחוז הסטודנטים העובדים

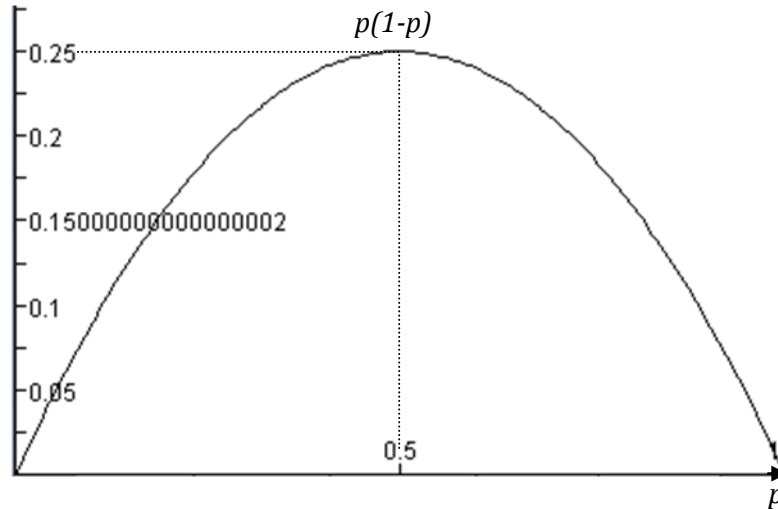
$$\alpha = 0.05, n = 25, \hat{p} = 0.6$$

ר"ס 0.95 ל- $p$ :

$$\left[0.6 - 1.96 \cdot \sqrt{\frac{0.6 \cdot 0.4}{25}}, 0.6 + 1.96 \cdot \sqrt{\frac{0.6 \cdot 0.4}{25}}\right] = [0.41, 0.79]$$

פתרון לאי דיוק b:

להחליף את  $p(1-p)$  במקסימום האפשרי שלו  $0.25 \leq p(1-p)$



$$Pr\left(p \in \left[\hat{p} - Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{1}{4n}}, \hat{p} + Z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{1}{4n}}\right]\right) \geq 1 - \alpha$$

עד כדי נכונות הקירוב הנומלי לבינומי, מובטח לנו שרמת הסמך של הרווח שלנו היא לפחות  $1 - \alpha$  זוהי גישה שמרנית לבניית רווח סמך עבור  $p$ .

בדוגמה: קודם קבלנו  $[0.41, 0.79]$ , עכשיו נקבל:

$$\left[0.6 - 1.96 \cdot \sqrt{\frac{1}{100}}, 0.6 + 1.96 \cdot \sqrt{\frac{1}{100}}\right] = [0.4, 0.8]$$

אם במקום  $\hat{p} = 0.6$  היה  $\hat{p} = 0.8$  אזי היחס בין אורך רווחי הסמך היה:  $\sqrt{\frac{0.25}{0.16}} = \frac{5}{4}$

הרעיון של גישה  $b$  השמרנית חשוב ומועיל במיוחד בסקרים פוליטיים ואחרים

- לפני עריכת הסקר ניתן לקבוע כמה נדגמים צריך לשאול כדי להגיע לרוחב "רצוי" של רווח הסמך.
- בד"כ  $p$  האמיתי קרוב ל- $1/2$ , ולכן השמרנות לא גדולה יותר מידי.

מקובל לקחת  $\pm 2 \cdot \frac{1}{2\sqrt{n}} = \pm \frac{1}{\sqrt{n}}$  ( $\approx Z_{0.975}$ ) בתור טעות הדגימה השמרנית במדגם  $\Leftarrow$  הערכה שמרנית של רוחב ר"ס 95% שמרנית.  
 $\Leftarrow$  לפני עשיית המדגם ניתן לחשב

n	טווח שגיאה
100	
400	
600	
1100	

(מקובל לקחת את השניים האחרונים)

## אינאריאנטיות ר"ס טרנ' מונוטונית

דוגמה: גובה סטודנטים, רוצים ר"ס לסטודנטים שגובהם מתחת ל-170 ס"מ.

$$\theta = Pr(X \leq 170) = \Phi\left(\frac{170 - \mu}{\sigma}\right)$$

נניח שונות  $\sigma^2 = 16$  ידועה,  $\bar{X} = 175$ ,  $n = 25$   
 אנ"מ ל- $\theta$ :

## טענה:

נתון פרמטר  $\theta$ , ר"ס על פי אומד  $S(X)$ , ברמת סמך  $1 - \alpha$ :  $I = [S(X) - \varepsilon_1, S(X) + \varepsilon_2]$   
 אם  $f(\theta)$  היא טרנספורמציה מונוטונית (עולה או יורדת) של  $\theta$  אזי  $f(I)$  הוא ר"ס ברמת סמך  $1 - \alpha$  עבור  $f(\theta)$ .

בדוגמת הסטודנטים: ר"ס עבור  $\mu$ :

$$\left[\bar{X} - 1.96 \cdot \frac{4}{5}, \bar{X} + 1.96 \cdot \frac{4}{5}\right] = [173.4, 176.6]$$

ר"ס עבור  $\Phi\left(\frac{170 - \mu}{4}\right)$  (פונ' מונוטונית יורדת של  $\mu$ ):

$$\left[\Phi\left(\frac{170 - (\bar{X} + 1.96 \cdot \frac{4}{5})}{4}\right), \Phi\left(\frac{170 - (\bar{X} - 1.96 \cdot \frac{4}{5})}{4}\right)\right] = \left[\Phi\left(-\frac{6.6}{4}\right), \Phi\left(-\frac{3.4}{4}\right)\right] = [0.049, 0.198]$$

## הוכחת הטענה:

$$Pr\left(\theta \in \underbrace{[S - \varepsilon_1, S + \varepsilon_2]}_I\right) \geq 1 - \alpha$$

אם  $f$  מונוטונית עולה אז:

$$S - \varepsilon_1 \leq \theta \leq S + \varepsilon_2 \Leftrightarrow f(S - \varepsilon_1) \leq f(\theta) \leq f(S + \varepsilon_2)$$

אם  $f$  מונוטונית יורדת אז:

$$S - \varepsilon_1 \leq \theta \leq S + \varepsilon_2 \Leftrightarrow f(S + \varepsilon_2) \leq f(\theta) \leq f(S - \varepsilon_1)$$

לכן המאורעות שקולים,  $Pr(f(\theta) \in f(I)) \geq 1 - \alpha$ .

דרך אחרת ("הדרך הרגילה") לחשב ר"ס לשיעור הסטודנטים מתחת ל-170:

- נסתכל על השיעור האמפירי במדגם
- נבנה ר"ס בינומי על סמך שיעור זה.

נניח ראינו 4 סטודנטים > 170 במדגם.  
זה אומר:

$$\hat{p} = \frac{\#\{\text{קטן מ 170}\}}{25} = 0.16$$

רווח סמך 95%:

$$\left[ \hat{p} - 1.96 \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{5}, \hat{p} + 1.96 \cdot \frac{\sqrt{\hat{p}(1-\hat{p})}}{5} \right] = [0.017, 0.303]$$

## סקרים ושקרים

### דוגמא א: הטיה במדגם

בחירות 1936 בארה"ב.  
מועמד א': רוזוולט (נשיא מכהן)  
מועמד ב': לנדון

דעת רוב הציבור: רוזוולט ינצח.

העיתון *Literary Digest* עשה סקר:

- שלח מכתבים לקוראים
- ביקש בעיתון לשלוח הצבעות
- וכו'...

⇐ מדגם בגודל 2.4 מיליון.

המסקנה:  $p < 0.5$  לנדון ינצח. (נגיד  $\hat{p} = 0.48$ )

בפועל: רוזוולט ניצח, 62% מהקולות.

הבעיה: קוראי העיתון בכלל, והעונים על השאלון בפרט, לא ייצגו את כלל האוכלוסיה.  
⇐ מדגם מאוד לא מקרי.

### דוגמא ב: הסקה לא נכונה

נובמבר 1995: רצח רבין, פרס רוה"מ  
סקר ינואר 1996: 46.5% פרס

35.8% נתניהו

17.5% אחר (לא החליט \ לא רוצה לענות \ לא יצביע 6.4%)

⇐ הוחלט להקדים את הבחירות.

ערב הבחירות: כותרת בעיתון: "פרס מוביל ב-3%: פרס 51.5%, נתניהו 48.5%, טווח השגיאה 3%"

ק: שיעור תמיכה בפרס.

מהו ההפרש בתמיכה:  $2p - 1 = p - (1 - p)$ .

$2p - 1$  טרנ' מונטונית של  $p$ , לכן ר"ס להפרש כפול מר"ס לשיעור התמיכה.  
 $\Leftarrow$  רווח סמך ל %תמיכה בפרס: [48.5%, 54.5%], ר"ס להפרש [-3%, +9%]

האם ההחלטה להקדים הייתה נכונה?

11% לא החליטו או לא מגלים.

ידוע היסטורית ש-75% מהצפים מצביעים ימין

נניח 8% נתניהו, 2.8% פרס

$\Leftarrow$  49.4% פרס

43.8% נתניהו

אם ננרמל מחדש לאחוזים מוחלטים: 53% פרס, 47% נתניהו.

$\Leftarrow$  עם טווח שגיאה של קצת יותר מ-3% לא ניתן להכריע מי ינצח.

### בדיקת \ בחינת השערות

מה למדנו מהסקה בינתיים?

1. לאמוד פרמטר

2. לבנות רווח סמך

הבעיה השלישית היא לבדוק השערות על פרמטרים.

דוגמא: סקר בחירות, אומדים אחוז התמיכה  $p$  במועמד א' מול מועמד ב'.

השאלה האמיתית שמעניינת: האם  $p > \frac{1}{2}$  בביטחון? האם  $p < \frac{1}{2}$  בביטחון? או האם לא ניתן לפסול  $p = \frac{1}{2}$ .

לשם כך נגדיר את "השערת האפס" שהיא העובדה שאנו רוצים להשתמש בנתונים כדי להפריך אותה.

אבל נפריך אותה רק אם נסיק בצורה ברורה מהנתונים שהשערת האפס אינה נכונה.

במקרה של בחירות:

$$H_0 : p = \frac{1}{2}$$

$$H_1 : p \neq \frac{1}{2} \text{ (ניתן לקבוע מי ינצח)}$$

### דוגמאות

1. אסטרונומית אומרת שהמפות מראות שבני מזל תאומים ודלי מוגנים יותר מסרטן

נניח שיעור הסרטן באוכלוסייה כולה  $p_0$  ידוע.

נסמן ב- $p$  שיעור הסרטן בקרב תאומים ודלי.

$$H_0 : p \geq p_0$$

$$H_A : p < p_0$$

2. מחקר על תרופה חדשה לאיידס.

נניח מס' תאי T באוכלוסייה בלי תרופה מתפלג נורמלי  $X \sim N(\mu_0, \sigma^2)$

רוצים לדעת האם התרופה עוזרת.

אז נניח עם התרופה  $X \sim N(\mu_1, \sigma^2)$

$$H_0 : \mu_0 = \mu_1$$

$$H_1 : \begin{cases} \mu_0 \neq \mu_1, & \text{(משפיעה בדרך כלשהי)} \\ \mu_0 > \mu_1, & \text{(עוזרת)} \end{cases}$$

### הגדרת מבחן סטטיסטי

נניח שיש לנו השערת  $H_0: \theta = \theta_0$

מבחן סטטיסטי ברמת  $\alpha$  מוגדר על ידי איזור דחיה  $C_\alpha \subset \mathbb{R}$  כך ש:

$$Pr_{H_0}(S(X) \in C_\alpha) \leq \alpha$$

המבחן: אם  $S(X) \in C_\alpha$  נדחה את  $H_0$

אם  $S(X) \notin C_\alpha$  לא נדחה את  $H_0$

קיימות הרבה הגדרות אפשריות ל- $C_\alpha$  ובד"כ נבחר ביניהן על פי האלטרנטיבה שמעניינת אותנו.

דוגמא:  $X \sim N(\mu, \sigma^2)$  ידוע.

$$H_0 : \mu = 0$$

$$H_A : \mu > 0$$

נשתמש באומד  $\bar{X}$ , מהי התפלגות תחת  $H_0$ ?

$$\bar{X} \stackrel{H_0}{\sim} N\left(0, \frac{\sigma^2}{n}\right)$$

נרצה שהשטח מתחת ל- $C_\alpha$  יהיה  $\alpha$

