

# סיכומון סטטיסטיקה

יובל בר

בהסתברות היה נתון לנו מרחב לצד התפלגות כלשהי ומשם הסקנו מהן ההסתברויות לקבלת מדגמים. בסטטיסטיקה, ההפך הוא מה שקורה. בהינתן המדגם, נרצה לבנות את המודל של המרחב וההתפלגות.

## 1 הפרדת השערות פשוטות

מטרתנו תהיה להפריד בין שתי השערות:

• **השערת האפס**  $H_0$  - השערה שבה אנחנו מאמינים בה מלכתחילה, מייצגת את המצב הקיים או הברירת מחדל.

• **ההשערה האלטרנטיבית**  $H_1$  - החלופה ל  $H_0$ , אם נדחה את  $H_0$  נקבל את  $H_1$ .

**הגדרה 1.** השערה פשוטה היא השערה הקובעת את ההתפלגות של מ"מ. השערה מורכבת היא השערה המשייכת את ההתפלגות של מ"מ לקבוצת התפלגויות.

**דוגמה.** השערה " $X \sim \text{Geo}(p)$  לאיזשהו  $p$ " היא מורכבת והשערה " $X \sim \text{Geo}(\frac{1}{2})$ " היא פשוטה.

**הגדרה 2.** מבחן הוא הגדרה של מאורע  $\Omega \supseteq A$  כך שאם מתרחש  $\omega \in A$  אנחנו מקבלים את  $H_0$  ואחרת אנחנו מקבלים את  $H_1$ .

**דוגמה.** מטילים מטבע 100 פעמים,  $X$  מספר העצים שנקבל,  $X \sim \text{Bin}(100, p)$ , ייתכן שהמטבע הוא מטבע מזויף שנחת על עץ בסיכוי  $\frac{3}{4}$ . אנחנו מניחים ש  $p = \frac{1}{2}$ , כלומר  $H_0 : X \sim \text{Bin}(100, \frac{1}{2})$ , ייתכן שהמטבע אכן מזויף, כלומר  $H_1 : X \sim \text{Bin}(100, \frac{3}{4})$ . נבחר מבחן - אם נקבל יותר מ-75 עצים אזי נדחה את  $H_0$  ונקבל את  $H_1$ .  $A = \{X > 75\}$

בדוגמה, המספר 75 נבחר באופן שרירותי, בחירה שונה הייתה נותנת מבחן אחר, אם היינו בוחרים למשל 95 כחסם, המבחן לא היה יעיל. נרצה לפתח דרך בה נוכל לאמוד עד כמה מבחן הוא יעיל.

### הגדרה 3. טעויות:

1. **טעות מסוג ראשון** - קבלה של  $H_1$  בהינתן ש- $H_0$  נכונה. ההסתברות לטעות הזו תהיה  $P_{H_0}(H_1)$ , היא מסומנת ב- $\alpha$ . נקראת **רמת המובהקות** של המבחן.

2. **טעות מסוג שני** - קבלה של  $\mathcal{H}_0$  בהינתן ש- $\mathcal{H}_1$  נכונה. ההסתברות לטעות הזו תהיה  $P_{\mathcal{H}_1}(\mathcal{H}_0)$ , היא מסומנת ב- $\beta$ . ההסתברות המשלימה,  $1 - \beta$  נקראת **העוצמה** הסטטיסטית של המבחן.

בהינתן מבחן  $A$ , נקבל  $\alpha = P(A^c | \mathcal{H}_0)$   $\beta = P(A | \mathcal{H}_1)$ , ככלל, אנחנו רוצים לצמצם טעויות מסוג ראשון, כלומר למזער את  $\alpha$ .

$\mathcal{H}_1$ נכונה	$\mathcal{H}_0$ נכונה	
טעות סוג שני $\beta$	עוצמה $= 1 - \beta$	קיבלנו את $\mathcal{H}_0$
	טעות סוג ראשון $\alpha$ = רמת מובהקות	קיבלנו את $\mathcal{H}_1$

**דוגמה.** נוכל להעריך את גודל הטעויות מהדוגמה הקודמת (בעזרת כלי ריכוז מידה)

$$\begin{aligned} \alpha &= P(A^c | \mathcal{H}_0) = P\left(X \leq 80 \mid X \sim \text{Bin}\left(100, \frac{1}{2}\right)\right) = P(X - \mathbb{E}[X] \leq 30) \\ &\leq \frac{1}{2} \cdot \frac{\text{Var}(X)}{30^2} = 0.0139 \\ \beta &= P(A | \mathcal{H}_1) = P\left(X > 80 \mid X \sim \text{Bin}\left(100, \frac{3}{4}\right)\right) = P(X - \mathbb{E}[X] \leq 5) \\ &\leq \frac{1}{2} \cdot \frac{\text{Var}(X)}{5^2} = 0.375 \end{aligned}$$

**הגדרה 4. יחס הנראות  $\lambda$**  של תוצאה  $\omega$  (יכול להיות סדרת תוצאות) מוגדר להיות כך

$$\lambda(\omega) := \frac{P(\omega | \mathcal{H}_1)}{P(\omega | \mathcal{H}_0)} = \frac{f_{\mathcal{H}_1}(\omega)}{f_{\mathcal{H}_0}(\omega)}$$

**הגדרה 5. מבחן ניימן-פירסון** הוא מבחן מהצורה  $\lambda(\omega) \leq K$  עבור  $K$  קבוע. כלומר, משווים את יחס הנראות לקבוע  $K$ , אם הוא נמוך יותר, מקבלים את  $\mathcal{H}_0$ , אחרת מקבלים את  $\mathcal{H}_1$ . על ידי שינוי הערך  $K$  נקבל ערכי מובהקות שונים.

**משפט 6. הלמה של ניימן-פירסון** בהפרדה בין השערות פשוטות עם רמת מובהקות  $\alpha$  נתונה, מבחן ניימן-פירסון הוא המבחן בעל העוצמה המקסימלית.

איך נתכנן ניסוי המבוסס על מבחן ניימן-פירסון?

1. נבחר את רמת המובהקות  $\alpha$  בה נרצה לעבוד.
2. נחשב את הסף  $K_\alpha$  שעבורו מבחן ניימן-פירסון יהיה בעל רמת מובהקות  $\alpha$ .
3. נחשב את יחס הנראות של הדגימות הנתונות ונשווה ל- $K_\alpha$ .

## 2 אמידה

בחלק זה נניח שנתון מדגם שנוצר מתוך התפלגות ידועה, שהפרמטרים שלה לא ידועים. נרצה למצוא כלים סטטיסטיים שיאפשרו לנו לאמוד מה הם הפרמטרים האלו.

### הגדרה 7. הגדרות

1. **סטטיסטי** הוא פונקציה של המדגם שאינה תלויה בפרמטר.
  2. **אומד** הוא סטטיסטי שערכו אמור להיות קרוב לפרמטר המבוקש.
  3. **אומדן** הוא ערך של האומד עבור ערכי מדגם מסוימים.
- בהינתן דגימות  $X_1, \dots, X_n \sim F_\theta$  (כלומר מתפלגים עם פונקציית התפלגות  $F$  בעלת פרמטר  $\theta$ ) אז  $X_1$  הוא אומד ל- $\theta$ .  
גם  $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$  הוא אומד ל- $\theta$ .  
הפונקציה  $\hat{\theta} = T(X_1, \dots, X_n)$  היא אומדן ל- $\theta$  כאשר  $\hat{\theta}$  הקירוב.  
נשים לב ש- $T$  בעצמו הוא מ"מ.

**משפט 8.**  $\mathbb{E}[\bar{X}] = \mathbb{E}[X]$   
קל להראות

$$\mathbb{E}[\bar{X}] = \mathbb{E}\left[\frac{1}{n} \sum_{i=1}^n X_i\right] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X_i] = \frac{1}{n} \sum_{i=1}^n \mathbb{E}[X] = \mathbb{E}[X]$$

- הגדרה 9.** אומדן  $\hat{\theta} = T(\{X_i\}_{i=1}^n)$  לפרמטר  $\theta$  הוא **חסר הטייה** אם  $\mathbb{E}[T] = \theta$ .  
נשים לב ש- $\bar{X}$  ו- $X_1$  הם אומדנים חסרי הטייה.