סיכום מושגים ושאלות - סטטיסטיקה

כאן ירוכזו כל הדברים החשובים שיעלו במהלך השיעור, כולל שאלות ספציפיות שעלו ונקודות שכדאי לזכור.

נקודות שעלו בשיחה:

1. רמת מובהקות (α) וטעות מסוג ראשון

- **הגדרה רשמית:** רמת המובהקות (אלפא, α) היא ההסתברות לדחות את השערת האפס (H_0) כאשר היא במציאות נכונה (טעות מסוג ראשון).
- **מתמטית:** $\alpha = P(\text{Reject } H_0 \mid H_0 \text{ is true})$

2. הקשר וההבדל בין אלפא (α) לעוצמת המבחן ($1-\beta$)

- **העיקרון:** כדי לחשב את α ואת $1-\beta$, בודקים את ההסתברות של אותו אירוע (אזור הדחייה) תחת שתי הנחות שונות.
- **ההבדל בחישוב:**
  - α (טעות מסוג I): חישוב ההסתברות ליפול באזור הדחייה תחת ההנחה ש- H_0 נכונה.
  - $1-\beta$ (עוצמת המבחן): חישוב ההסתברות ליפול באזור הדחייה תחת ההנחה ש- H_1 נכונה.

3. אזורי דחייה וקבלה - האם הם משתנים או זזים?

- **העיקרון החשוב ביותר:** כלל ההחלטה (אזור הדחייה ואזור הקבלה) נקבע מראש על בסיס השערת האפס והרמה המבוקשת. הוא נותר קבוע ללא תלות בהשערה הנבחנת.

4. ניסוח השערות על סוג התפלגות

- **הצגה פורמלית:** כאשר השערות מתייחסות לסוג ההתפלגות שממנה נגזרים הנתונים:
  - $H_0 : X \sim U(0,3)$
  - $H_1 : X \sim Bin(4, 0.2)$

5. טבלת החלטות וטעויות (מטריצת הבלבולים)

	מקבלים את H0	דוחים את H0
במציאות H_0 נכונה	החלטה נכונה ($1-$)	**טעות מסוג I** (α)

במציאות H_1 **נכונה**	**טעות מסוג II** (β)	החלטה נכונה ($1-\beta$)

6. חילוץ פרמטרים ממשפט הגבול המרכזי (CLT) כשההשערה היא התפלגות מלאה

- **העיקרון:** בעת שימוש בממוצע מדגם ($n > 30$), משתמשים בקירוב נורמלי: $\bar{X} \sim N(\mu, \frac{\sigma^2}{n})$. יש לחשב את μ ו- σ^2 מתוך נוסחאות ההתפלגות הספציפיות המוגדרות בהשערות.

7. למה מחלקים ב- n כשמתקננים ממוצע (שורש שונות חלקי מדגם)?

- **העיקרון המתמטי:** השונות של ממוצע המדגם ($Var(\bar{X})$) היא $\frac{\sigma^2}{n}$. טעות התקן, שהיא סטיית התקן של הממוצע, היא $\sqrt{\frac{\sigma^2}{n}} = \frac{\sigma}{\sqrt{n}}$.

8. מתי מחלקים ב- n ומתי לא?

- **תצפית בודדת (X):** שונות: σ^2 .
- **ממוצע המדגם ($\bar{X}$):** שונות: $\frac{\sigma^2}{n}$.
- **סכום המדגם (ΣX):** שונות: $n \cdot \sigma^2$.

9. סלט של סימנים ונוסחאות: סיגמא (σ), S , ו"טעות התקן" (שגיאת תקן)

- σ היא סטיית התקן של האוכלוסייה, בעוד S הוא האומד המדגמי שלה. טעות תקן היא המונח המקצועי לסטיית התקן של הממוצע ($\frac{\sigma}{\sqrt{n}}$).

10. רווח סמך לתוחלת ואיך מוצאים את ערך ה-Z

- **הנוסחה:** $\bar{X} \pm Z_{1-\alpha/2} \cdot \text{SE}$. את ערך ה- Z מוצאים על ידי חיפוש ההסתברות המצטברת הרצויה בטבלת ההתפלגות הנורמלית הסטנדרטית.

11. אלגוריתם עבודה: השלבים לפתרון שאלת בדיקת השערות

1. **איסוף נתונים:** זיהוי $n, \bar{X}, \sigma$ או S .
2. **ניסוח השערות:** קביעת H_0 ו- H_1 (כיווני או דו-צדדי).

3. **זיהוי מטרת השאלה:** הכרעה בין מציאת כלל החלטה לבין חישוב הסתברות טעות.

12. הבלבול הנפוץ: רווח סמך מול ערך קריטי

- **רווח סמך:** מחושב סביב ממוצע המדגם ($\bar{X}$).
- **בדיקת השערות (ערך קריטי):** מחושב סביב תוחלת השערת האפס (μ_0).

13. ניסוח רשמי ואקדמי של עוצמת המבחן ($1-\beta$) וטעות מסוג שני (β)

עוצמת המבחן משקפת את יכולת המבחן לדחות השערת אפס שגויה. דיווח אקדמי יציין את ההסתברות לדחייה נכונה בהינתן ערך פרמטר חלופי מסוים.

14. כיצד נתון או מחושב S

S (או S^2) עשוי להינתן ישירות, מחושב מסכומי ריבועים בסיסיים, או מחושב מתוך מדגם גולמי קטן באמצעות הנוסחה: $S^2 = \frac{\sum(x_i - \bar{X})^2}{n-1}$.

15. הסתברות המובהקות (P-Value)

ה-P-Value מייצג את רמת המובהקות המינימלית שעבורה השערת האפס תידחה. דחייה תתבצע כאשר $\text{P-Value} \leq \alpha$.

16. אותיות גדולות לעומת קטנות

- **אות גדולה:** מייצגת את המשתנה המקרי (התפלגות תיאורטית).
- **אות קטנה:** מייצגת ערך תצפית ספציפי או מימוש קונקרטי של המשתנה.

17. יסודות תורת האומדנים

אומד הוא פונקציה של נתוני המדגם המשמשת לניחוש פרמטר לא ידוע באוכלוסייה. נהוג לסמן אומד בתוספת כובע ($\hat{\theta}$). "אומד" מתייחס לנוסחה התיאורטית, בעוד "אומדן" מתייחס לערך המספרי שהתקבל לאחר הצבת נתוני המדגם.

18. תכונות אומדים והשוואה ביניהם

כאשר משווים בין שני אומדים, משתמשים בסדר קדימויות מוגדר:

1. **חסר הטיה (Unbiasedness):** אומד מוגדר כחסר הטיה אם תוחלתו שווה לפרמטר האמיתי, כלומר $E[\hat{\theta}] = \theta$.
2. **יעילות (Efficiency):** במידה ושני אומדים חסרי הטיה, נבחר באומד בעל השונות המינימלית ($Var(\hat{\theta})$).
3. **שגיאה ריבועית ממוצעת (MSE):** במידה וקיים אומד מוטה בעל שונות נמוכה משמעותית

מאומד חסר הטיה, ניתן להעדיפו אם ה-MSE שלו נמוך יותר. ה-MSE מוגדר כ:

$$MSE(\hat{\theta}) = Var(\hat{\theta}) + [Bias(\hat{\theta})]^2$$

האומד העדיף הוא זה שממזער את ערך ה-MSE.

19. טעויות נפוצות בחישוב שונות וסטיית תקן

קיימים שני כשלים מתמטיים נפוצים בעת חישוב שונויות וביצוע סטנדרטיזציה (ציון Z):

1. **חישוב שונות של משתנה מקרי בדיד:** בחישוב שונות לפי הנוסחה $\sigma^2 = \sum(x_i - \mu)^2 \cdot P(X = x_i)$, חובה להציב בתוך הסוגריים את **ערכי המשתנה המקרי עצמו** (x_i), ולא את ההסתברויות שלהם. הצבת ערכי ההסתברות בתוך הסוגריים תוביל לתוצאה שגויה לחלוטין של השונות.
2. **תקנון ממוצע מדגם ללא שימוש בשגיאת תקן:** כאשר מחשבים הסתברות הקשורה ל**ממוצע מדגם** ($\bar{X}$), פילוג הממוצע כפוף למשפט הגבול המרכזי והשונות שלו קטנה פי n. לפיכך, חובה לחלק את סטיית התקן בשורש גודל המדגם ($\sigma_{\bar{X}} = \frac{\sigma}{\sqrt{n}}$). שימוש בסטיית התקן הרגילה (σ) במכנה של חישוב Z יתאים רק במקרה של תצפית בודדת, ויוביל להטיה חמורה בחישוב כשמדובר בממוצע מדגמי.