סטטיסטיקה – שיעור 9

בעיית בדיקת השערות

כרגיל, חושבים על אוכלוסייה עם התפלגות על X ופרמטר θ שמאפיין את ההתפלגות ומעניין אותנו לאמוד.

בבעיית בדיקת השערות יש לנו ערך θ_0 שאנחנו "מאמינים" שהוא הנכון ל-θ – $H_0: \theta = \theta_0$
יש לנו טענה מתחרה לגבי הערך של 0 - $H_A: \theta \in R_A$. כאשר R_A קבוצת ערכים כלשהי שלא מכילה את θ_0.

הרעיון הבסיסי: נמשיך להאמין ב-H_0 אלא אם כן יש לנו "סיבה טובה" לשנות את דעתנו.
באופן מעשי: ניקח מדגם מקרי, נחשב סטטיסטי $S(X)$ ונדחה את השערת ה-0 אם הערך שנקבל לסטטיסטי הוא "לא סביר" תחת H_0.

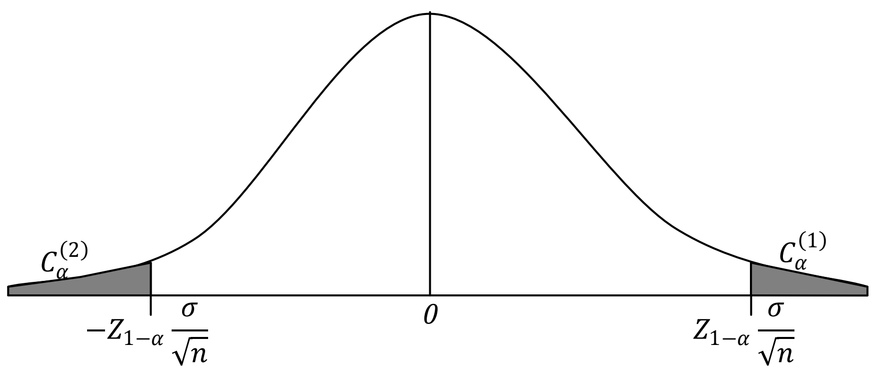
מבחן סטטיסטי

מבחן סטטיסטי ברמה α להשערה $H_0: \theta = \theta_0$ על סמך סטטיסטי $S(X)$ מוגדרת ע"י אזור דחייה $C_\alpha \subseteq Range(S)$ כך ש:

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

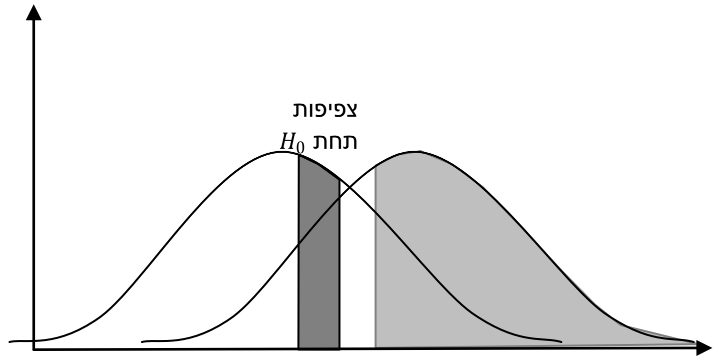
אזור הדחייה <u>אינו</u> מוגדר בצורה חד ערכית ע"י H_0, α

<u>צפיפות $S(X)$:</u>



איך נבחר בין אזורי הדחייה האפשריים:

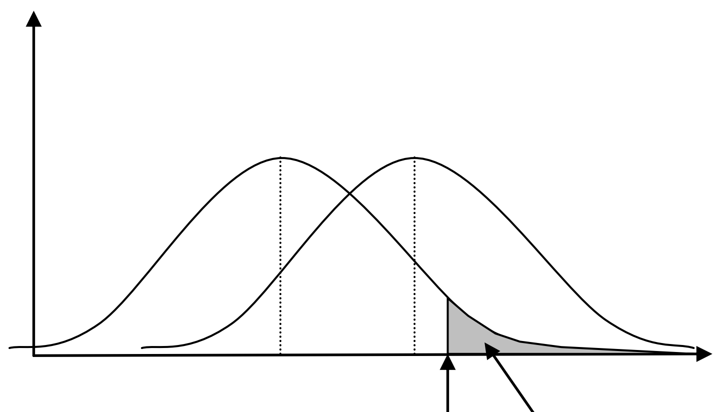
נסתכל על האלטרנטיבה שמעניינת אותנו, ונשאל את עצמנו, איזה מאזורי הדחייה "יותר סביר" תחת האלטרנטיבה.



<u>דוגמה:</u> $X\sim N(\mu,\sigma^2)$, σ^2 ידוע. $H_0: \mu=\mu_0$, משתמשים ב-$S(X)=\bar{X}$

תחת H_0: $\bar{X}\sim N\left(\mu_0,\frac{\sigma^2}{n}\right)$.

אזורי דחייה אפשריים ברמת α:



מבחנים במקרה הנורמלי:

a. מה יהיה אזור הדחייה תחת אלטרנטיבה $H_A: \mu=\mu_1\ (\mu_1>\mu_0)$?

במקרה זה $\bar{X}\sim N\left(\mu_1,\frac{\sigma^2}{n}\right)$.

$$\alpha=P_{H_0}(\bar{X}\in C_\alpha^*)=\Pr_{H_0}(\bar{X}\geq d_\alpha^*)=\mu_0+Z_{1-\alpha}\cdot\frac{\sigma}{\sqrt{n}}$$

חישוב אזור דחייה עובר אלטרנטיבה $\mu=\mu_1>\mu_0$.

1. נשיב לב שתחת האלטרנטיבה ערכים גדולים ל- $\bar{X}=S(X)$ הם יותר סבירים.
2. נבחר אזור דחייה מהסוג $C_\alpha^*=[d_\alpha^*,\infty)$
3. נמצא את d_α^*: $P_{H_0}(\bar{X}\geq d_\alpha^*)=\alpha$

$$\Pr_{H_0}\left(\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}}\geq\frac{d_\alpha^*-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right)=\alpha$$

$$\Pr_{H_0}\left(\frac{\bar{X}-\mu_0}{\frac{\sigma}{\sqrt{n}}} \le \frac{d_\alpha^*-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = 1-\alpha$$

$$\Phi\left(\frac{d_\alpha^*-\mu_0}{\frac{\sigma}{\sqrt{n}}}\right) = 1-\alpha$$

$$\frac{d_\alpha^*-\mu_0}{\frac{\sigma}{\sqrt{n}}} \le \Phi^{-1}(1-\alpha) = Z_{1-\alpha} \Longrightarrow d_\alpha^* = \mu_0 + Z_{1-\alpha}\cdot\frac{\sigma}{\sqrt{n}}$$

b. מה יהיה אזור הדחייה תחת $H_A: \mu > \mu_0$. $\Leftarrow$ תוביל לאותו אזור דחייה $C_\alpha^* = [d_\alpha^*, \infty)$.

c. באופן דומה עבור $H_A: \mu < \mu_0$ נבחר אזור דחייה מהסוג: $C_\alpha^* = \left(-\infty, \mu_0 + \underbrace{Z_{1-\alpha}}_{=-Z_{1-\alpha}} \cdot \frac{\sigma}{\sqrt{n}}\right)$

d. מה עושים אם $H_A: \mu \neq \mu_0$?

במקרה כזה, הפתרון המקובל הוא: $C_\alpha^* = \left(-\infty, \mu_0 + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}\right] \cup \left[\mu_0 + Z_{1-\frac{\alpha}{2}} \cdot \frac{\sigma}{\sqrt{n}}, \infty\right)$.

במקרה של נורמלי עם שונות לא ידועה , נעשה אומד לשונות $\hat{\sigma}^2$ ונחליף את Z – t.

במקרה הבינומי:

a. $H_0: p = p_0$, אלטרנטיבה: $H_A: p > p_0$, רוצים להשתמש ב- $S(X) = \hat{p}$.

ניח נשתמש בקרוב הנורמלי:

מהו כלל ההחלטה האם הקרוב הנורמלי טוב?

$\Leftarrow$ כלל ההחלטה צריך להתבסס על p_0: $10 \le np_0, n(1-p_0) \ge 10$.

במקרה הזה תחת H_0 אנו יודעים את שונות $\hat{p}$:

$$\hat{p} \dot{\sim} N\left(p_0, \frac{p_0(1-p_0)}{n}\right)$$

$\Leftarrow$ תחת הקרוב הנומרלי, אזור הדחייה:

$$C_\alpha^* = \left[p_0 + Z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}}, +\infty\right)$$

b. באופן דומה: $H_A: p < p_0$, נקבל אזור דחייה:

$$C_\alpha^* = \left(-\infty, p_0 + Z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}}\right]$$

c. $H_A: p \neq p_0$

$$C_\alpha^* = \left(-\infty, p_0 + Z_\alpha\sqrt{\frac{p_0(1-p_0)}{n}}\right] \cup \left[p_0 + Z_{1-\alpha}\sqrt{\frac{p_0(1-p_0)}{n}}, +\infty\right)$$

סוגי השערות

<u>השערה פשוטה</u>: השערה שכוללת ערך אחד של θ בלבד.

למשל: $H_A: \theta = \theta_A, H_0: \theta = \theta_0$.

השערה מורכבת: השערה שכוללת אוסף ערכים (1<) של θ:
למשל: $H_0: \theta > \theta_0, H_0: \theta \leq 0, H_A: \theta < 0, H_A: \theta \neq 0$.

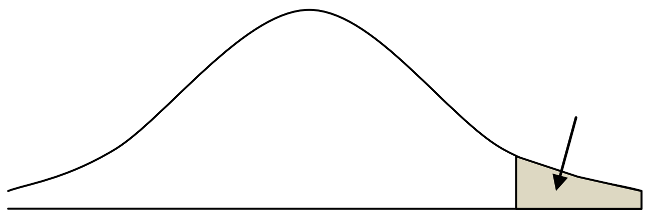
בקורס שלנו נעסוק רק בהשערות אפס פשוטות וחלק מהזמן גם באלטרנטיבות פשוטות.

אלטרנטיבה $H_A: \theta > \theta_0$ תקרא חד צדדית.
אלטרנטיבה $H_A: \theta \neq \theta_0$ תקרא דו צדדית.

ערך p או רמת מובהקות של תוצאה

נניח אנחנו רוצים לבדוק $H_0: \theta = \theta_0$ מול אלט' $H_A: \theta > \theta_0$ בעזרת סטטיסטי $S(X)$.

גישה מתבקשת: בהנתן מדגם x ותוצאה $S(\underline{X})$, נשאל מה תהיה הרמה המינימלית α^* כך ש- $S(x) \in C^*_{\alpha^*}$.



במקרה הנורמלי: שונות ידועה $H_0 = \mu = \mu_0$, $H_A: \mu > \mu_0$.
אזי רמת המובהקות של תוצאה $S(\underline{x}) = \bar{x}$

$$\alpha^*(\bar{x}) = 1 - \Phi\left(\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right)$$

נדחה את H_0 בכל רמה שגדולהאו שווה מ-$\alpha^*(\bar{x})$.
חישוב: $\bar{x} = \mu_0 + Z_{1-\alpha^*} \cdot \frac{\sigma}{\sqrt{n}}$.

$$\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}} = Z_{1-\alpha^*}$$

$$1 - \alpha^*(\bar{x}) = \Phi\left(\frac{\bar{x} - \mu_0}{\frac{\sigma}{\sqrt{n}}}\right)$$

דוגמה: גובהי סטודנטים, $n = 25, \sigma^2 = 16$. $H_0: \mu = 170, H_A: \mu > 170$.
נניח קיבלנו $\bar{x} = 175$.

$$\alpha(\bar{x}) = 1 - \Phi\left(\frac{175 - 170}{\frac{4}{5}}\right) = 1 - \Phi(6.25) = 2 \times 10^{-10}$$

מה היה קורה אם הייתי מניח $\sigma^2 = 144$?

$$\alpha(\bar{x}) = 1 - \Phi\left(\frac{175-170}{\frac{12}{5}}\right) = 1 - \Phi(2.08) = 0.019$$

במקרה הזה אדחה את השערת ה-0 כאשר $\alpha = 0.05$ ולא אדחה כאשר $\alpha = 0.01$.

השערות פשוטות: סוגי טעויות ועוצמה

$H_0: \theta = \theta_0, H_A: \theta = \theta_A$, סטטיסטי $S(X)$, רמה α, איזור דחייה C_α.

<u>הגדרה:</u> טעות מסוג I: הסיכוי לדחות את השערת האפס, אם היא בעצם נכונה.

$$\Pr_{H_0}(S(X) \in C_\alpha) \underset{\text{מהגדרת מבחן}}{=} \alpha$$

<u>הגדרה:</u> טעות מסוג II: הסיכוי <u>שלא</u> נדחה את השערת האפס, אם <u>האלטרנטיבה נכונה</u>

$$\beta \coloneqq \Pr_{H_1}(S(X) \notin C_\alpha) \stackrel{?}{=}$$

<u>הגדרה:</u> <u>עצמת המבחן</u> היא הסיכוי <u>שנדחה</u> את השערת האפס אם האלטרנטיבה נכונה.

$$\left.\begin{matrix} power \\ \pi \end{matrix}\right\} \coloneqq P_{h_1}(S(X) \in C_\alpha) = 1 - \beta$$

דוגמת הסטודנטים: $H_0: \mu = 170, H_A: \mu = 175$, רמת המבחן 0.05

$$C_\alpha = \left[170 + 1.645 \cdot \frac{4}{5}, +\infty\right) = [171.3, +\infty]$$

$$\alpha = P_{H_0}(\bar{X} \in C_\alpha) = P_{H_0}\left(\frac{\bar{X}-170}{\frac{4}{5}} \geq \frac{171.3-170}{\frac{4}{5}}\right) = 1 - \Phi\left(\frac{1.3}{\frac{4}{5}}\right) = 1 - \Phi(1.645) = 0.05$$

$$\beta = P_{H_1}(\bar{X} \notin C_\alpha) = P_{H_1}(\bar{X} \leq 171.3) = P_{H_1}\left(\frac{\bar{X}-172}{\frac{4}{5}} \leq -\frac{0.7}{\frac{4}{5}}\right) = \Phi\left(-\frac{0.7}{\frac{4}{5}}\right) = 0.19$$

$$\Longrightarrow \pi = 0.81$$

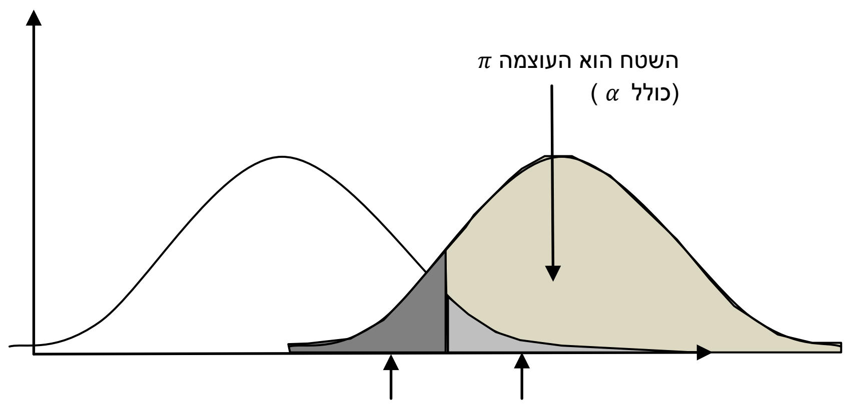
<u>שאלה:</u> מה היה קורה ל- π אם היינו לוקחים $H_A: \mu = 175$?

$$\beta = \Pr_{\mu=175}(\bar{X} < 171.3) = \Pr_{\mu=175}\left(\frac{\bar{X}-175}{\frac{4}{5}} \leq -\frac{3.7}{\frac{4}{5}}\right) = \Phi\left(-\frac{3.7}{\frac{4}{5}}\right) = 2 \times 10^{-6}$$

$$\Longrightarrow \pi = 0.99998$$

אפשר להגיד שהמקרה השני הוא בעיית בדיקת השערות <u>קלה</u>.

אם האלט' נכונה, סיכויינו לדחות גבוהים מאוד.



בחירת אזורי דחייה, הלמה של ניימן-פירסון

עד עכשיו דנו בבחירת סטטיסטי מבחן ואזורי דחייה ברמה האינטואיטיבית בלבד.
ברור שרוצים לקבל כמה שיותר עוצמה כדי שהמבחן יהיה טוב.

מכיוון שאנחנו לא יודעים מראש אילו סטטיסטיים יתנו לנו מבחנים בעוצמה גבוהה, ננסח את הדרישות שלנו ע"פ הסטטיסטי הכללי ביותר שקיים: המדגם כולו X.

<u>הגדרה</u>: מבחן ברמה α בעל עצמה מקסימלית להשערות הפשוטות: $H_0: \theta = \theta_0, H_A: \theta = \theta_A$ הוא איזור דחייה C_α^* (עבור X) כך שלכל איזור דחייה <u>אחר</u> ברמה α $\widetilde{C_\alpha}$ מתקיים: $\pi(C_\alpha^*) \geq \pi(\widetilde{C_\alpha})$.

כדי לנסח את הלמה של ניימן-פירסון שמגדירה את C_α^*, נזכר בפונקציית הנראות ונגדיר יחס נראות:

$$\Lambda(\underline{x}) = \frac{f_{\theta_0}(\underline{x})}{f_{\theta_1}(\underline{x})} \underset{\text{לפי הגדרת הנראות}}{=} \frac{L(\theta_0; \underline{x})}{L(\theta_A; \underline{x})}$$

הלמה של ניימן ופירסון

מבחן בעל עצמה מקסימלית להשערות של מעלה מוגדר על ידי:

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

כאשר d_α^* נקבע כך ש- $P_{H_0}(C_\alpha^*) = \alpha$.

<u>דוגמה</u>: המקרה נורמלי, שונות ידועה. $H_0: \mu = \mu_0, H_A: \mu = \mu_A \ (\mu_A \geq \mu_0)$

$$L(\mu_0; \underline{x}) = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{\sum (X_i - \mu_0)^2}{2\sigma^2}\right\} = \left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{\frac{\sum X_i^2}{2\sigma^2}\right\} \exp\left\{-\frac{n\mu_0\sigma^2}{\sigma^2}\right\} \exp\left\{\frac{n\bar{X}\mu_0}{\sigma^2}\right\}$$

$$L(\mu_A;\underline{x})=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{-\frac{\sum(X_i-\mu_A)^2}{2\sigma^2}\right\}=\left(\frac{1}{\sqrt{2\pi\sigma^2}}\right)^n \exp\left\{\frac{\sum X_i^2}{2\sigma^2}\right\}\exp\left\{-\frac{n\mu_A\sigma^2}{\sigma^2}\right\}\exp\left\{\frac{n\bar{X}\mu_A}{\sigma^2}\right\}$$

$$\Lambda(\underline{x})=\underbrace{\frac{\exp\left\{-\frac{n\mu_0\sigma^2}{\sigma^2}\right\}}{\exp\left\{-\frac{n\mu_A\sigma^2}{\sigma^2}\right\}}}_{\text{לא תלוי ב } \underline{x},\ C:=}\cdot\exp\left\{\frac{n\bar{X}(\mu_0-\mu_A)}{\sigma^2}\right\}$$

$$\frac{\partial\Lambda(\underline{x})}{\partial x}=\underbrace{C}_{>0}\cdot\underbrace{\frac{n(\mu_0-\mu_A)}{\sigma^2}}_{<0}\cdot\underbrace{\exp\{\dots\}}_{>0}<0$$

$\Leftarrow$ ערכים גדולים של $\bar{X}$ יתנו ערכים נמוכים של $\Lambda(\underline{x})$ $\Leftarrow$ איזור הדחייה של מבחן ניימן פירסון הוא מהסוג $C_\alpha^*=\{\underline{x}:\bar{X}\geq d_\alpha^*\}$ $\Leftarrow$ אותו אזור דחייה שהגדרנו אינטואיטיבית.