מטלת פרויקט 2

מטרת מטלה 2 היא ליישם את מבחני ההשערות שלמדתם וכן לבדוק את ההשפעה של גודל המדגם על אורך רווחי הסמך ותוצאות המבחנים.

במשימה הקודמת, התבקשתם לנסח את שאלת המבחן – האם הערך של משתנה X שונה בין קטגוריות שונות של משתנה בינארי Y. אם יש לכם יותר משתי קטגוריות, הגבילו את עצמכם לשתי קטגוריות לטובת המשימה הבאה.

<u>תזכורת</u>: ניתן להחליף את שאלת המבחן שבחרתם במטלת פרויקט 1.

ודאו כי ההסברים שלכם מלווים בתוצאות המוצגות בקובץ ההגשה, ניתן להציג בשנית תוצאות ממשימה קודמת.

שימו לב כי עבור הרצת הקוד לצורך בדיקה, נשתמש גם בקוד שהגשתם במטלת פרויקט 1. כלומר, הקוד במטלה צריך לרוץ בהצלחה על מאגר הנתונים שהגשתם במטלה הקודמת לאחר הטרנספורמציות שביצעתם בו.

1. <u>אמידה:</u>

א. חשבו את הממוצע של כל קטגוריה.

ב. חשבו רווח סמך מקורב לתוחלת המשתנה X בכל קטגוריה (תוך שימוש באנ"מ לתוחלת). ניתן להניח לצורך החישוב כי $\bar{X}_n \approx \mathcal{N}\left(\mu_x, \frac{\hat{\sigma}^2}{n}\right)$ כאשר $\hat{\sigma}^2$ היא שונות הדגימה בכל קטגוריה. הסבירו מדוע ניתן להניח זאת.

ג. האם רווחי הסמך של שתי הקטגוריות חופפים? הסבירו את המשמעות של תוצאה זאת.

2. <u>מבחני השערות:</u>

נרצה לבצע מבחן השערות כדי לבחון אם הממוצעים שונים בין הקטגוריות שבחרתם.

א. נסחו את השערת האפס ואת האלטרנטיבה.

ב. האם ההנחות של מבחן t מתקיימות? ניתן לבצע בדיקת נורמליות באופן איכותי בעזרת היסטוגרמה. כמו כן, ניתן לבדוק שוויון שונויות באמצעות מבחן F (ראו תזכורת בסוף התרגיל).

ג. בדקו את ההשערה בעזרת מבחן t (ללא תלות בתוצאות של סעיף ב'). השתמשו בסטטיסטי המבחן עבור הפרש תוחלות כאשר השונויות שוות אך לא ידועות.

ד. בדקו את ההשערה בעזרת מבחן וולד.

<u>הדרכה</u>: השתמשו במבחן שהוצג בהרצאה 2.17, שקף 89 (דוגמה 10.8 בספר).

ה. השוו בין התוצאות ובין ערכי הp-value.

ו. כעת הניחו שהנתונים מגיעים מהתפלגות נורמלית. גם כאן נניח כי השונויות בשתי הקטגוריות זהות אך אינן ידועות. בצעו מבחן יחס נראות כדי לבדוק את ההשערות הבאות:

$$H_0: \mu_1 - \mu_2 = 0$$
$$H_1: \mu_1 - \mu_2 \neq 0$$

<u>הדרכה:</u>

- סמנו ב-$\{X_{1i}\}_{i=1}^{n_1}$ את הנתונים בקטגוריה הראשונה, וב- $\{X_{2j}\}_{j=1}^{n_2}$ את הנתונים בקטגוריה השנייה. כמו כן, $n = n_1 + n_2$

- חשבו את הנראות בכל מרחב הפרמטרים לפי
$$\mathcal{L}(\hat{\mu}_1,\hat{\mu}_2)=\frac{1}{\left(\sqrt{2\pi S_p^2}\right)^n}\prod_{i=1}^{n_1}e^{-(X_{1i}-\hat{\mu}_1)^2/2S_p^2}\prod_{j=1}^{n_2}e^{-(X_{2i}-\hat{\mu}_2)^2/2S_p^2}.$$
כאשר S_p^2 היא ה-pooled variance שחישבתם בסעיף ג.
- חשבו את הנראות תחת H_0 לפי:
$$\mathcal{L}(\hat{\mu})=\frac{1}{\left(\sqrt{2\pi\hat{\sigma}^2}\right)^n}\prod_{i=1}^{n}e^{-(X_i-\hat{\mu})^2/2\hat{\sigma}^2}$$
כאשר כאן $\hat{\sigma}^2$ היא השונות הדגימה של כלל הנתונים (כלומר של שתי הקטגוריות יחד).

3. <u>גדלי מדגם שונים:</u>

א. בחרו באופן אקראי מדגמים בגדלים $30, 50, 100, 500$ מקובץ הנתונים המקורי (שימו לב שמספר התצפיות מכל קטגוריה יכול להיות שונה).

ב. חשבו רווח סמך לתוחלת של המשתנה X בכל קטגוריה עבור כל גודל מדגם. השוו לתוצאה בשאלה 1.

ג. בצעו את מבחן וולד עבור כל גודל מדגם. השוו לתוצאה בשאלה 2.

ד. חזרו על סעיפים 3א-3ג 100 פעמים והיעזרו בגרפים ו/או טבלאות לתאר את ההתפלגות של אורך רווחי הסמך ושל ה p-values כפונקציה של גודל המדגם. מה אחוז הפעמים שהממוצעים שמצאתם בשאלה אחת נמצאים ברווחי הסמך?

הפעילו שיקול דעת ובחרו בעצמכם כיצד להציג את התוצאות של הסעיפים השונים.

<u>תזכורת: מבחן F לשוויון שונויות</u>

ניתן להשתמש במבחן F כדי לבדוק שוויון שונויות בין שני מדגמים של **תצפיות המתפלגות נורמלית** באופן הבא:

$$H_0:\sigma_X^2=\sigma_Y^2$$
$$H_1:\sigma_X^2\neq\sigma_Y^2$$
$$T.S.:F=\frac{S_X^2}{S_Y^2}=\frac{\frac{1}{n-1}\sum_{i=1}^{n}(X_i-\bar{X})^2}{\frac{1}{m-1}\sum_{i=1}^{n}(Y_i-\bar{Y})^2}\sim F_{(n-1,m-1)}$$
$$R.R.:F>F_{(n-1,m-1),1-\alpha}$$

כדי להשתמש במבחן בתצורה זו יש להציב במונה את השונות המדגמית הגדולה מבין שתי הדגימות (כלומר צריך להתקיים $S_X^2>S_Y^2$).