סטטיסטיקה היסקית: הקשר בין גיל למשך השינה

בהמשך לתרגיל 1, נמשיך לנתח נתונים מתוך הקובץ מאתר Kaggle בקישור, כאשר המשתנים שהשתמשנו בהם הם: **X (משתנה מסביר)**- גיל האדם הנבדק (בשנים). **Y (משתנה מוסבר)**- משך השינה (בשעות).

1. א. נניח $Y \sim N(\mu, \sigma^2)$. נרצה לאמוד את הפרמטרים הלא- ידועים μ, σ^2.

כדי לאמוד אותם, נשתמש ב**שיטת המומנטים**, שעקרונה הוא שימוש בממוצע שעפ״י חוק המספרים הגדולים, מתקרב לתוחלתו כאשר המדגמים גדלים. לפי כלל אמידה זה:

נשווה מומנט תיאורטי ראשון- תוחלת, למומנט הראשון מהנתונים- ממוצע המדגם, ונקבל אומד לתוחלת: $\hat{\mu} = \overline{Y} = 7.132$

נשווה מומנט תיאורטי שני לממוצע הריבועי של המדגם, ובעזרת נוסחת העבודה, נקבל אומד לשונות:

$$\hat{\sigma}^2 = \overline{Y^2} - \hat{\mu}^2 = 0.631$$

כלומר, בהתבסס על נתוני המדגם ושימוש באומד מומנטים, נעריך כי תוחלת שעות השינה באוכלוסייה היא 7.132, ושונות שעות השינה הוא 0.631.

ב. $m = \min(X) = 27 \implies W = X - 27$. נניח כי $W \sim Gamma(\alpha, \lambda)$.

נאמוד את α, λ באמצעות **אומד מומנטים**. נשווה מומנט ראשון (תוחלת גאמה) לממוצע המדגם ומומנט שני לממוצע הריבועי, שימוש בנוסחת העבודה ופיתוח נוסף ייתן:

$$\hat{\lambda} = \frac{\overline{W}}{\overline{W^2} - (\overline{W})^2} = 0.202 \ , \ \hat{\alpha} = \frac{(\overline{W})^2}{\overline{W^2} - (\overline{W})^2} = 3.073$$

כלומר, בהתבסס על נתוני המדגם ושימוש באומד מומנטים, נעריך את התפלגות הגילאים באוכלוסייה מהם החסרנו את ערך הגיל המינימלי במדגם, כהתפלגות Gamma(3.073,0.202).

ג+ד. $\boldsymbol{Y \sim N(7.132, 0.631)}$, לפי הפרמטרים שהתקבלו מהאמידה. לפיכך, האחוזון ה- k של Y יהיה האחוזון ה- k של התפלגות נורמלית עם פרמטרים אלו.

$\boldsymbol{W \sim Gamma(3.073, 0.202)}$, לפי הפרמטרים שהתקבלו מהאמידה. לכן, האחוזון ה- k של W יהיה האחוזון ה- k של התפלגות Gamma עם פרמטרים אלו.

נחשב בעזרת R את האחוזונים הנדרשים לפי המודלים ולפי הנתונים ונציג בטבלה הבאה:

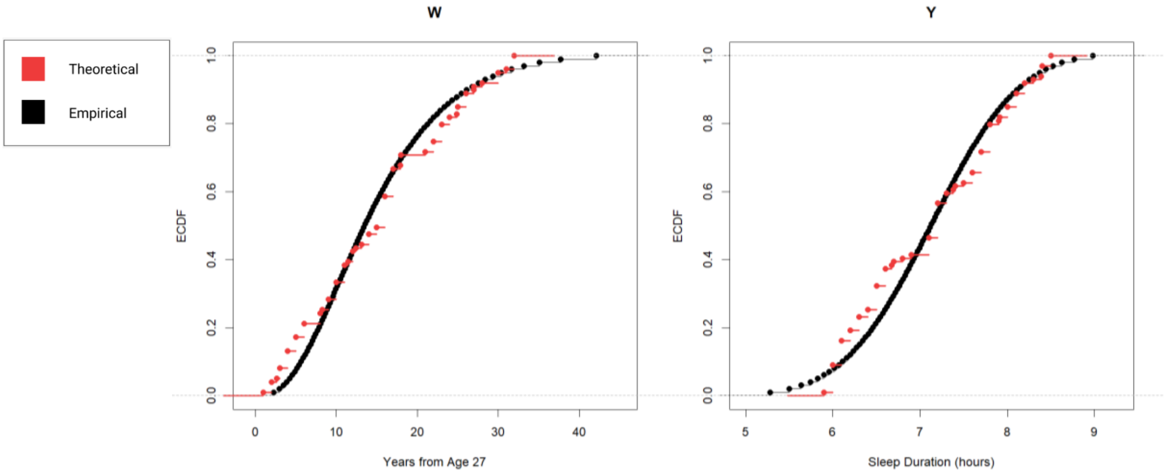
טבלת השוואה בין האחוזונים

מ"מ/ אחוזון	10		50		75		90	
	מודל	אמפירי	מודל	אמפירי	מודל	אמפירי	מודל	אמפירי
Y	6.114	6.1	7.132	7.2	7.668	7.8	8.150	8.2
W	5.668	4	13.573	16	19.806	23	26.798	27

מההשוואה ניתן לראות כי האחוזונים המתקבלים עבור שעות השינה לפי המודל הנורמלי, קרובים מאוד לאחוזונים המתקבלים מהנתונים. נתון מעניין העולה, הוא שמודל התפלגות נורמלית לשעות השינה חוזה כי מחצית מהאוכלוסייה ישנים פחות מ-7.132 שעות, ומחציתם יותר. זאת, בדומה מאוד לתוצאה מהנתונים שנותנת ערך של 7.2. הסבר אפשרי לדמיון בין האחוזונים, הוא שבכמות תצפיות גדולה, במקרה שלנו 374, התפלגות שעות השינה הממוצעות היא נורמלית בקירוב, וזאת בהתאם למשפט הגבול המרכזי. מכאן, נראה כי המודל הנורמלי הינו מודל שיכול להתאים להתפלגות שעות השינה. לצד זאת, ניתן לראות כי האחוזונים המתקבלים עבור גילאי הנבדקים (לאחר החסרת הגיל

המינימלי) לפי מודל התפלגות גאמה, לא מאוד דומים לאחוזונים האמפיריים. לכן, עפ״י אחוזונים אלו, ניתן לשער כי המודל לא מתאר בצורה מדויקת את הנתונים והתפלגות הגילאים אינה Gamma(3.073, 0.202).

ה. לאחר חישוב האחוזונים 1-99 האמפיריים והתיאורטיים, יצרנו גרפי ECDF, בעזרת R, על מנת לבחון עד כמה ההתפלגות התיאורטית מתאימה לנתונים האמפיריים. הגרפים מציגים את הפונקציה המצטברת האמפירית של הנתונים בפועל, מול הפונקציה המצטברת התיאורטית (עפ״י המודלים), עבור כל משתנה.



ניתן לראות כי המודלים התיאורטיים, גם עבור שעות השינה וגם עבור הגילאים, מתארים בצורה יחסית קרובה את הנתונים האמפיריים. אולם, אין התאמה מושלמת וישנם אחוזונים שהמודלים לא חוזים באופן מדויק, בעיקר בקצוות. למרות זאת, ההתאמה היא גבוהה, ולכן ניתן להסיק כי המודל הנורמלי מתאים להתפלגות שעות השינה ומתאר בצורה קרובה את האחוזונים המתקבלים מהמדגם. כמו כן, התפלגות גאמה מתארת את W בצורה סבירה, אך ייתכן שיהיה צורך לבדוק התאמות נוספות, במיוחד בערכים הקיצוניים יותר.

2. נחשב **רווח סמך לתוחלת של Y** ברמת סמך של 97%. נשים לב כי שונות שעות השינה באוכלוסייה אינה ידועה. כמו כן, המדגם מכיל 374 תצפיות (מדגם גדול), כך שניתן להניח כי האומד לשונות המדגם מתקרב בערכו לשונות האמיתית באוכלוסייה. לכן, נוכל להשתמש בר״ס במקרה של שונות ידועה, ע״י שימוש באומד $\hat{\sigma}^2$ לשונות שחישבנו בשאלה 1 עבור התפלגות זו. כלומר, עפ״י הנוסחה:

$$\overline{Y} \pm z_{1-\frac{\alpha}{2}} \cdot \sqrt{\frac{\hat{\sigma}^2}{n}} = 7.132 \pm 2.17 \cdot \sqrt{\frac{0.631}{374}} = [7.043, 7.221]$$

נוכל לומר כי ברמת ביטחון של 97%, כלומר רק ב- 97% מהמדגמים הפוטנציאליים, רווח הסמך שחישבנו יכיל את תוחלת שעות השינה. במדגם הספציפי, לא נוכל לדעת אם התוחלת אכן כלולה בר״ס.

נחשב כעת **רווח סמך לשונות של Y**, ברמת סמך של 92%. נוכל להשתמש באומד $\hat{\sigma}^2$ שמצאנו בשאלה 1, שכן בגלל גודל המדגם, הוא שואף לשונות האמיתית. נחשב:

$$\left[\frac{(n-1)\hat{\sigma}^2}{\chi^2_{n-1,1-\frac{\alpha}{2}}}, \cdot \frac{(n-1)\hat{\sigma}^2}{\chi^2_{n-1,\frac{\alpha}{2}}}\right] = \left[\frac{373 \cdot 0.631}{422.16}, \frac{373 \cdot 0.631}{326.59}\right] = [0.558, 0.721]$$

נקבל כי ר״ס נע בטווח הערכים 0.558-0.721. המשמעות לכך היא פיזור קטן יחסית מסביב לממוצע

בשעות השינה אצל הנבדקים. כמו כן, נוכל לומר כי רק ב92% מהמדגמים הפוטנציאליים עליהם יכולנו לחשב רווח סמך, ר"ס יכיל את השונות. נזכור כי לא נוכל להגיד בוודאות כי רווח הסמך שחישבנו אכן מכיל את השונות.

3. $med(X) = 43$. נחלק לשתי קבוצות: $B = \{(X_i, Y_i) | X_i < 43\}$,$A = \{(X_i, Y_i) | X_i \geq 43\}$.

א. שאלת המחקר שלנו היא האם גילו של אדם משפיע על שעות השינה הממוצעות שלו. נצפה לראות שבגילאים מבוגרים יותר, משך השינה יהיה גדול יותר (הגוף זקוק ליותר מנוחה). לכן, ההשערה האלטרנטיבית שלנו היא שאצל אנשים בגיל 43 ומעלה, תוחלת שעות השינה תהיה גדולה יותר מאשר אצל אנשים מתחת לגיל זה.

ב. נסמן: μ_{Y_A}- תוחלת שעות השינה בקבוצה A, μ_{Y_B}- תוחלת שעות השינה בקבוצה B.

כעת, נגדיר את ההשערות, כאשר הפרש התוחלות $\Delta = \mu_{Y_A} - \mu_{Y_B}$:

השערת האפס: $H_0:\ \Delta = 0 \iff H_0:\ \mu_{Y_A} = \mu_{Y_B}$

השערה אלטרנטיבית: $H_1:\ \Delta > 0 \iff H_1: \mu_{Y_A} > \mu_{Y_B}$

ג. $m = 188$- מס' האנשים בקבוצה A, $n = 186$- מס' האנשים בקבוצה B. נשים לב כי הקבוצות הן ב"ת והמ"מ (X_i, Y_i בתוך זוג) ב"ת, מפני שאין תלות בשעות השינה בין אדם אחד לאחר. בנוסף, השונות באוכלוסייה אינה ידועה. לא ידוע אם השונויות בשתי הקבוצות שוות, ואף סביר להניח שהן לא שוות, מפני שאנשים צעירים ואנשים מבוגרים עם סדר יום שונה- אנשים צעירים יכולים ללמוד וגם לעבוד, לעומת מבוגרים שכבר בשגרה קבועה יותר ודומה אחד לשני. לכן, נניח כי השונויות אינן שוות. כמו כן, מספר התצפיות גדול דיו בכל קבוצה. לכן, נשתמש בקירוב הנורמלי, כך שסטטיסטי המבחן יהיה:

$$T = \frac{\bar{Y}_A - \bar{Y}_B}{\sqrt{\frac{S^2_{Y_A}}{m} + \frac{S^2_{Y_B}}{n}}} \overset{H_0}{\sim} N(0,1) \implies T = 1.653$$

כאשר $\bar{Y}_A = 7.199, \bar{Y}_B = 7.064$ ממוצעי משך השינה בקבוצות A, B, והשונויות הן:

$$S^2_{Y_A} = \frac{1}{m-1}\sum_{i=1}^{m}\left(Y_{A_i} - \bar{Y}_A\right)^2 = 0.845\ , \qquad S^2_{Y_B} = \frac{1}{n-1}\sum_{i=1}^{n}\left(Y_{B_i} - \bar{Y}_B\right)^2 = 0.413$$

מכאן, נחשב באמצעות R את ערך ה- P ונקבל: $p_{val} = 1 - \Phi(1.653) = 0.05$. כלומר, ההסתברות תחת השערת האפס לקבל את ממוצע המדגם הנתון או תוצאה קיצונית יותר (בכיוון ההשערה האלטרנטיבית), היא 0.05.

ד. נתונה רמת מובהקות: $\alpha = 0.03$. לאחר ביצוע המבחן, קיבלנו $p_{val} = 0.05 > 0.03$, כלומר לא נדחה את השערת האפס. משמעות תוצאה זו היא שלא ניתן לקבוע כי תוחלת שעות השינה של אנשים בגיל 43 ומעלה, גבוהה יותר מתוחלת שעות השינה של אנשים מתחת לגיל זה. מכאן עולה כי הנתונים אינם מספיקים כדי לומר שיש קשר בין הגיל לשעות השינה.

לסיכום, בהתאם לניתוח, מתחזקת המסקנה כי אין קשר מהותי בין הגיל לשעות השינה הממוצעות. המחקר שלנו עסק במדגם של אנשים בגילאים 27-59, וככל הנראה בגילאים אלו הגיל אינו מהווה גורם מסביר באופן מספק למס' שעות השינה. לפיכך, סביר להניח כי ישנם גורמים נוספים המשפיעים על שעות השינה, וייתכן שאם היינו בוחנות טווח גילאים רחב יותר היינו מוצאות הבדל מובהק יותר.