<u>תרגיל כיתה שבוע 4</u>

התקינו את ספריית mlbench.

יבאו אותה ב-editor.

תוכלו ללמוד על תוכן הספריה ע"י העלאת התיעוד שלה באמצעות הפקודה: library(help = "mlbench").

אנו נעבוד על קובץ הנתונים Ozone. יבאו אותו על ידי הפקודה data(Ozone).

הוא מכיל את המשתנים הבאים:

1	Month: 1 = January, ..., 12 = December
2	Day of month
3	Day of week: 1 = Monday, ..., 7 = Sunday
4	Daily maximum one-hour-average ozone reading
5	500 millibar pressure height (m) measured at Vandenberg AFB
6	Wind speed (mph) at Los Angeles International Airport (LAX)
7	Humidity (%) at LAX
8	Temperature (degrees F) measured at Sandburg, CA
9	Temperature (degrees F) measured at El Monte, CA
10	Inversion base height (feet) at LAX
11	Pressure gradient (mm Hg) from LAX to Daggett, CA
12	Inversion base temperature (degrees F) at LAX
13	Visibility (miles) measured at LAX

אנו נעבוד רק על משתנה האוזון, "V4" כתלות בחודש, "V1".

צרו dataframe חדשה שמכילה רק את העמודות הללו. חיתוך dataframe נעשה בצורה הבאה:

DataName[c(selected rows), c(selected columns)].

עמודה "V1" היא עמודת פקטור. יש להפוך אותה לנומרית כדי שנוכל לטפל בה בהמשך. הפכו אותה לנומרית על ידי הפקודה as.numeric() והשמה מחדש למשתנה.

הסירו את השורות שבהן מופיעים חוסרים על ידי הפקודה na.omit() והשמה מחודשת אל בסיס הנתונים.

ציירו את עקומת ההתפלגות של משתנה האוזון. האם המשתנה מתפלג נורמלית? למה כן/לא?

בחרו רק את נתוני האוזון מחודש יוני בלבד. (בחירת שורות עם תנאי מתוך בסיס נתונים ובחירת עמודה). הציבו את הנתונים שבחרתם במשתנה וקטורי חדש בשם y_summer.

הציגו את התפלגות המשתנה y_summer בדיאגרמת שכיחויות עם 8 bins (שמונה נקודות דגימה בציר x).

האם ההתפלגות של משתנה האוזון בחודש יוני נראית לכם נורמלית? למה כן/לא?

<u>בדיקת השערות עבור התוחלת:</u>

1. בדקו את ההשערה, שרמת האוזון הממוצעת (תוחלת האוזון) היא 15 ברמת מובהקות α=.05.

H_0: μ=15

H_1: μ≠15

סטיית התקן הכללית (סטיית התקן באוכלוסיה) אינה ידועה.

לצורך כך, הגדירו את גודל המדגם במשתנה, n.

דווחו את ממוצע המדגם, sample_mean

דווחו את סטיית התקן במדגם, sample_sd.

הדפיסו את המשתנים הללו אל המסך (השתמשו בפונקציה (print(paste(,,)).

לצורך בדיקת ההשערות יש לבצע מבחן t דו צדדי.

ניתן לבצע מבחן זה באמצעות מבחן t מובנה ב-R (t.test). בצעו את המבחן ודווחו את האומדן הסטיסטי ואת ערך ה-P_v המתקבל ממנו. שימו לב, שהפלט של המבחן הינו אובייקט, המכיל מספר attributes. בצעו הדפסה למסך של המשתנים הדרושים.

ניתן גם לחשב אנליטית את האומדן הסטיסטי על ידי המשוואה: $t = |\mu - \bar{x}| \cdot \frac{\sqrt{n}}{s}$

את השטח התואם תחת הגרף, ניתן לחשב באמצעות הפונקציה (pt(, ללא הזנב השמאלי (lower.tail = F), עם מספר דרגות חופש מתאים (n-1), ו-t שחישבנו (בערך חיובי) קודם. את השטח הזה יש להכפיל ב-2, כדי לחשב את הסיכוי לשגיאה משני הקצוות.

האם P_v שקיבלנו מהחישוב זהה לזה שחושב על ידי הפונקציה t.test?

מה המסקנה שניתן להסיק ממבחן זה?

2. בנו רווח סמך סביב ממוצע המדגם ברמת מובהקות α=.05.

לצורך כך, מצאו את t ברמת מובהקות מתאימה על ידי הפונקציה (qt(.

חשבו את גבולות רווח הסמך על ידי הצבה במשוואה:

$$C.I = \bar{x} \pm t_{1-\alpha/2} \cdot \frac{s}{\sqrt{n}}$$

בדקו האם התוחלת (הממוצע באוכלוסיה) נופלת בגבולות רווח הסמך?

מה ניתן להסיק מכך?

האם תוצאה זאת מתיישבת עם בדיקת ההשערות בשאלה 1?