<u>מטלה 1: סטטיסטיקה להנדסת תוכנה</u>

שאלה 1

שאלה זאת עוסקת בנתונים מטאורולוגים שנמצאים בבסיס נתונים שנקרא Ozone.
כדי לקבל את הנתונים עליכם להתקין את החבילה mlbench, לייבא אותה בקוד, ולהעלות את קובץ הנתונים (data("Ozone")).
אוסף הנתונים מכיל את שלושה עשר 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

מתוכם, התרגיל מתייחס לנתוני הלחות (V7) ולחודשי השנה (V1). עיבוד מקדים:

- העתיקו את העמודות הללו למבנה נתונים חדש (Ozone[,c("V1","V7")]).
- נקו את כל השורות בהן מופיעים חוסרים (NA) ע"י הפונקציה na.omit().
- וודאו שעמודות אלו הן נומריות ולא קטגוריות (y = as.numeric(y)).

א) תארו את נתוני הלחות בהיסטוגרמה עם 20 יחידות מדידה. שימו לב לעמודה הגבוהה שמופיעה בערכי הלחות הנמוכים. מה לדעתכם גורם אפשרי לעמודה יוצאת דופן זאת?

ב) צרו היסטוגרמה חדשה ללא נתוני הלחות הנמוכים (לצורך כך צרו משתנה חדש ללא נתונים אלו). האם ההתפלגות נראית נורמלית כעת?
האם ישנם גורמים בהתפלגות שנראים שונה מהתפלגות נורמלית?

ג) בחרו רק את חודשי הקיץ (יולי-ספטמבר כולל).
(i) הדגימו התפלגות נורמלית של הנתונים בהיסטוגרמה.
(ii) הניחו שסטיית התקן באוכלוסיה (σ) היא 7%.
קיימת השערה שהלחות הממוצעת בחודשים אלו (μ) היא 74%.
חשבו את סטיית התקן במדגם ואת הממוצע במדגם.
חשבו את הסיכוי שהנתונים תומכים בהשערה הקיימת, לעומת הסברה שהמדגם נובע ממוצע לחות שונה מ-74% (P_v) ברמת מובהקות $\alpha=0.05$.
הסבירו את משמעות הסיכוי שקיבלתם להשערה שהנתונים תומכים בהשערת האפס.

(iii) חשבו את רווח הסמך סביב ממוצע המדגם בחודשי הקיץ ברמת בטחון של 95%. האם התוחלת (μ) נמצאת בטווח זה?

(iv) מה הקשר בין התשובה לשאלה (ii) לתשובה לשאלה (iii)?

שאלה 2

עבור מיני יונקים מסויימים, הנקבות ממליטות שגר (מספר גורים) בכל המלטה. גודל השגר הנפוץ הוא ארבעה גורים. ידוע, שיחס הנקבות לזכרים ביונק מסויים הוא 1:1. לא ידוע האם מין הצאצאים נקבע בצורה בלתי תלויה, או אם יש נטייה למין זהה בהמלטות מרובות גורים, כך שבלידה נתונה יהיו יותר נקבות או לחלופין יותר זכרים.

אם מין העובר נקבע בצורה בלתי תלויה במין שאר העוברים בשגר, אזי מספר הנקבות (או הזכרים) פר שגר יתפלג בצורה בינומית. זוהי השערת האפס.

נאספו נתונים מ-90 המלטות של ארבעה גורים בכל המלטה.

מתוכן, 8 המלטות היו של זכרים בלבד. ב-27 המלטות היתה נקבה אחת מתוך ארבעת הגורים, ב-20 המלטות היו שתי נקבות, ב-25 המלטות היו שלוש נקבות וב-10 היו ארבע נקבות בשגר.

א) הציגו את טבלת ההסתברויות לכל אחד מהמקרים בקבוצה הצפויה ובקבוצה הנצפית. באותה הטבלה הציגו עמודה למספר הנקבות הנצפה בהמלטה (x_i), ועמודה למספר הנקבות הנצפה בהמלטה (m_i) בהנתן n=90.

ב) הציגו את התפלגות התדירויות הנצפות בעקומת עמודות (probability distribution).

ג) הציגו את ערך ה-x^2 המחושב עבור כל קבוצה בטבלה שב-א', וחשבו את x^2 הכללי.

ד) בעזרת הנתונים הללו, הראו האם ניתן לדחות את השערת האפס ברמת מובהקות של 5%.

ה) לפי המבחן, האם הנתונים תומכים בהשערה כי יש נטייה למין אחיד בהמלטות מרובות גורים?

תזכורת: ההסתברויות להתפלגות בינומית יכולות להיות מחושבות לפי הנוסחא P(x). אם x הוא מספר ההצלחות (מספר הנקבות מ-0 עד 4) ב-n נסיונות בלתי תלויים (n מייצג את מספר הגורים בהמלטה), וקיים סיכוי ρ להצלחה וסיכוי q לכשלון בכל נסיון, אזי:

$$\boldsymbol{P(X=x) = \binom{n}{x} \rho^x q^{n-x}}$$