תרגיל בית 1

שאלה 1:

בקובץ ״Prices.csv״ מרוכזים מחירי 200 בתים (במיליוני $) בני שישה חדרים אשר נמכרו בארה״ב, במחוז קינג (King County) - מחוז במדינת וושינגטון שבירתו היא סיאטל, בין מאי 2014 עד מאי 2015. הנתונים נלקחו מהאתר:

https://www.kaggle.com/harlfoxem/housesalesprediction

את הסעיפים שלהלן יש לפתור בשפת R – יש לצרף את הפלטים ובנוסף להציג את הקוד.

א. בנו את גרף ההיסטוגרמה כאשר בציר האופקי ערכים מ-0.1 עד 7.1 וגודל האינטרוול הינו 0.1. מה ניתן ללמוד על הנתונים על סמך ההיסטוגרמה? על-סמך ההיסטוגרמה, האם סביר כי הנתונים באים מהתפלגות נורמלית? הסבירו.

ב. מצאו את המדדים הבאים:

- ממוצע
- חציון
- רבעון ראשון (תחתון) ורבעון שלישי (עליון)
- שונות וסטיית תקן
- טווח הנתונים ותחום בין-רבעוני

ג. שרטטו דיאגרמת QQ-Plot. האם סביר כי הנתונים באים מהתפלגות נורמלית?

ד. חשבו LF ו-UF, LW ו-UW. ציירו דיאגרמת Box-Plot. מה ניתן לומר על התפלגות הנתונים על סמך דיאגרמה זו? האם קיימות תצפיות חריגות? אם כן, כמה?

שאלה 2:

הגרילו ב-R שלושה מדגמים בגדלים שונים (למשל, 5, 15 ו-150) מהתפלגות אקספוננציאלית עם פרמטר λ בעזרת פונקציית rexp. בחרו את גדלי המדגמים וכן את הפרמטר λ כרצונכם כך ש- $0 < \lambda < 0.5$. הגדירו set.seed(a) כך ש-a היינו 4 ספרות אחרונות של ת.ז. המגיש.

עבור הסעיפים שאתם פותרים בשפת R – יש להציג את הקוד ואת הפלטים.

א) עבור כל אחד מהמדגמים חשבו ב-R את המדדים הסטטיסטיים הבאים: ממוצע, שונות, חציון, השברון ה- p עבור $0 < p < 0.5$ לבחירתכם.

ב) עבור המדגם הקטן ביותר חשבו את פונקציית ההתפלגות המצטברת האמפירית באופן ידני. פרטו חישובים.

ג) נגדיר X – משתנה מקרי מהתפלגות אקספוננציאלית עם פרמטר λ כפי שבחרתם. מהי פונקציית ההתפלגות המצטברת של X? חשבו תוחלת, שונות, חציון והשברון ה- p של X (עבור אותו ערך של p שנבחר בסעיף א').

ד) שרטטו ב-R על פני גרף אחד את פונקציית ההתפלגות המצטברת התאורטית שציינתם בסעיף הקודם בעזרת פונקציית pexp ואת שלוש הפונקציות האמפיריות עבור המדגמים שהגרלתם בעזרת פונקציית ecdf.

רמז: היעזרו בקוד שהוצג בהרצאה.

ה) דונו בהבדלים בין המדדים המדגמיים לבין המדדים התאורטיים המתאימים, תוך התייחסות לתוצאות הסעיפים הקודמים. מה אתם מצפים שיקרה להבדלים אלה אם תקחו מדגמים יותר גדולים? תנו הצדקה תאורטית עבור הממוצע המדגמי ופונקציית ההתפלגות האמפירית.

שאלה 3:

בשאלה זו לגבי כל טענה יש לציין האם היא נכונה, לנמק ולהוכיח את הטענות הנכונות.

יהי X משתנה מקרי מהתפלגות בעלת פונקציית התפלגות מצטברת $F_X(x;\theta)$ המקיימת: $F_X(\theta+d)=1-F_X(\theta-d)$ לכל d. נגדיר $Y=X-\theta$.

יהי ξ_p שברון p של ההתפלגות של X. להלן שלוש טענות:

I. $2\theta-\xi_p$ הינו שברון $1-p$ של ההתפלגות של X.

II. $\xi_p-\theta$ הינו שברון p של ההתפלגות של Y.

III. 0 הינו חציון של ההתפלגות של Y.

שאלה 4 (לא להגשה):

א) בסעיף זה יש להוכיח את טענה 2 המופיעה במצגת 1 בשקף 26 עבור מקרה מסוים (ההוכחה במקרים אחרים דומה). להלן המשימה:

יהיו $x_1,\ldots,x_n$ מספרים ממשיים. נניח כי n הינו זוגי וכי $x_{\left(\frac{n}{2}\right)}<x_{\left(\frac{n}{2}+1\right)}$.

יהי m מספר ממשי כלשהו המקיים $x_{\left(\frac{n}{2}\right)}<m<x_{\left(\frac{n}{2}+1\right)}$. הוכח כי עבור כל a ממשי,

$$\frac{1}{n}\sum_{i=1}^{n}|x_i-m|\le\frac{1}{n}\sum_{i=1}^{n}|x_i-a|$$

הדרכה: צריך להוכיח כי

$$\frac{1}{n}\sum_{i=1}^{n}\left(|x_i-a|-|x_i-m|\right)\geq 0$$

נניח, ללא הגבלת הכלליות, כי $a<m$ (המקרה $a>m$ הוא סימטרי). נגדיר שלוש קבוצות של אינדקסים:

$$A=\{i:\ x_i<a\},\quad B=\{i:\ a<x_i\leq m\},\quad C=\{i:\ x_i>m\}$$

i. הוכח כי לכל $i\in A$, $|x_i-a|-|x_i-m|=a-m$.

ii. הוכח כי לכל $i\in B$, $|x_i-a|-|x_i-m|=2x_i-a-m\geq a-m$.

iii. הוכח כי לכל $i\in C$, $|x_i-a|-|x_i-m|=m-a$.

הסק כי

$$\frac{1}{n}\sum_{i=1}^{n}\left(|x_i-a|-|x_i-m|\right)\geq\frac{m-a}{n}\left[|C|-\left(|A|+|B|\right)\right]=\frac{m-a}{n}\left(\frac{n}{2}-\frac{n}{2}\right)=0$$

כאשר $|A|,|B|,|C|$ מציינים את מספר האיברים בקבוצות A, B ו-C בהתאמה.

ב) הוכח את טענה 3 המופיעה במצגת 1 בשקף 26, כלומר הוכח כי אם C^* הינו שכיח, אזי לכל a ממשי,

$$\frac{1}{n}\sum_{i=1}^{n}I\left(x_i\neq C^*\right)\leq\frac{1}{n}\sum_{i=1}^{n}I\left(x_i\neq a\right)$$

<u>רמז</u>: התייחס להשוואה בין $\sum_{i=1}^{n}I\left(x_i=C^*\right)$ לבין $\sum_{i=1}^{n}I\left(x_i=a\right)$.