

## סטטיסטיקה למדעי המחשב – פתרון תרגיל 2

1. הגדרת טבלת הנתונים ב-R:

```
manStatus = c(rep('Students',449),rep('Staff',381))  
carType = c(rep('American',107),rep('European',212),rep('Japanese',130),rep('American',91),rep('European',120),rep('Japanese',170))  
table1 = table(manStatus,carType)
```

i.

a. ההתפלגות השולית של רכבים לפי מדינת ייצור:

```
carType.n=apply(table1,2,sum)  
carType.n/sum(carType.n)
```

American	European	Japanese
0.2385542	0.4000000	0.3614458

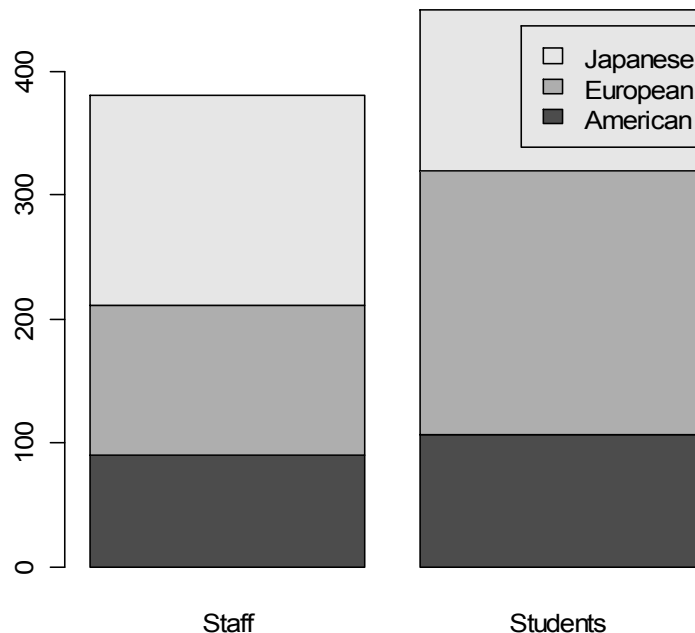
b. ההתפלגות השולית של רכבים לפי מדינת ייצור בהתניה על תפקיד הבעלים:

```
prop.table(table1,1)
```

manStatus	carType	American	European	Japanese
Staff	American	0.2388451	0.3149606	0.4461942
Students	American	0.2383073	0.4721604	0.2895323

ii. הצגה גראפית של הנתונים:

```
barplot(t(table1),legend.text=T)
```



.2

- A .i
- B .ii
- C .iii

.3

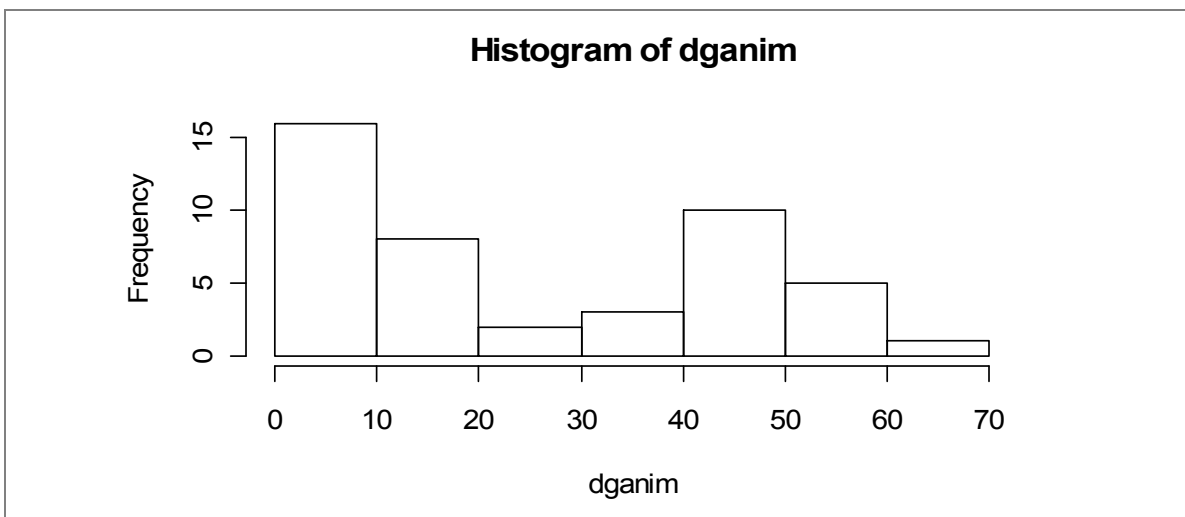
- i. 3 דוגמאות לגדלים אשר נמדדים בסקאלה לוגריתמית:
- a. עוצמה של רעידות אדמה (סולם ריכטר).
  - b. מדידת עוצמות צליל שהאוזן האנושית קולטת (דציבל).
  - c. מדידת חומציות של תמיסה (Ph).
- ii. הגובה של אדם שגובהו 175 ס"מ בסולם לוגריתמי (בסיס 2) הוא **7.45**.
- iii. הפרש הגבהים הוא 1. לא צריך לדעת את הגובה שלי כדי לענות על השאלה כיוון ש  $\log_2(x \cdot 2) = \log_2 x + \log_2 2 = (\log_2 x) + 1$
- iv. הפרש זה בסולם לוגריתמי בבסיס 10 יהיה  $\log_{10}(x \cdot 2) = \log_{10} x + \log_{10} 2 = (\log_2 x) + 0.301$
- v. הפרש גבהים בסולם לוגריתמים בין הגובה שלי לגובה של אדם הגבוה ממני ב10 ס"מ הוא  $\log_2(187) - \log_2(177) = \log_{10} x + \log_{10} 2 = 7.54 - 7.46 = 0.08$ . צריך לדעת את הגובה שלי כדי לחשב זאת כי סולם לוגריתמי מבטא שינויים אקספוננציאליים. לכן, ההפרש בין 2 ל12 הוא 2.58 לעומת זאת, כאמור, ההפרש בגבהים הוא רק 0.08.

.4 הגדרת הנתונים ב-R:

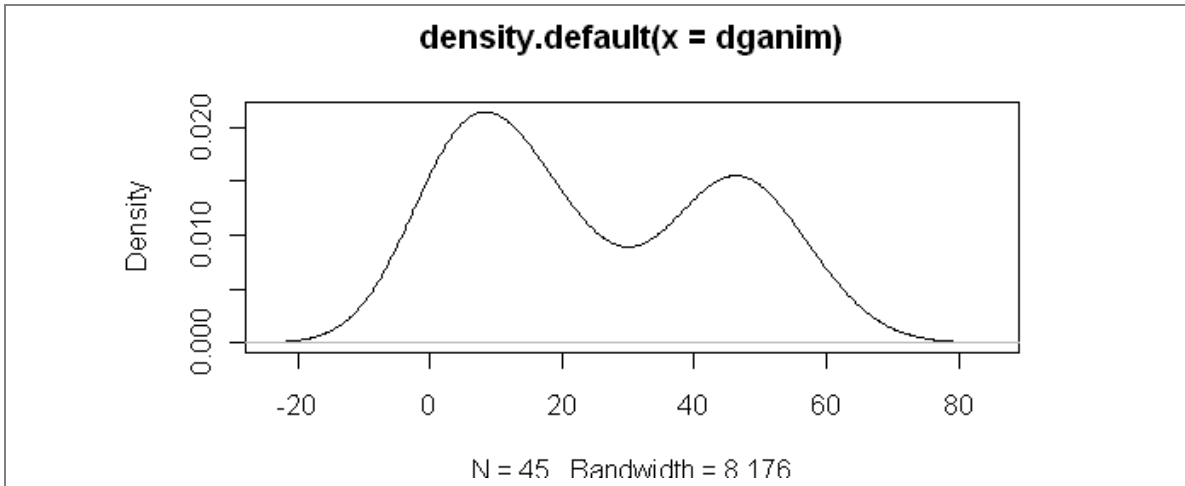
```
dganim=c(40.3, 55, 45.7, 43.3, 50.3, 53.5, 43, 44.2, 44, 47.4, 44, 33.6, 55.1, 48.8, 50.4, 37.8, 60.3, 46.6, 20, 30.2, 2.2, 7.5, 22.2, 16.6, 14.5, 21.4, 3.3, 6.6, 7.8, 10.6, 16.2, 14.5, 4.1, 15.8, 4.1, 2.4, 3.5, 8.5, 10, 1, 4.4, 1.3, 8.1, 4.7, 18.4)
```

i. שרטוט ההיסטוגרמה:

```
hist(dganim)
```



```
plot(density(dganim))
```



a. כן. ניתן לראות זאת בברור בהיסטוגרמה ובתרשים הצפיפות.  
b. ע"פ התרשימים הדגנים מתחלקים ל-2 קבוצות.

c. ייתכן וההבדלים נובעים כתוצאה מהגדרות רגולציה של כמות הסוכר המותרת בחלוקה לדגנים רגילים ודגנים דיאטטיים.

.ii

a. ממוצע

```
mean(dganim)  
[1] 24.96
```

b. חציון

```
median(dganim)  
[1] 18.4
```

c. ממוצע קצוץ

```
mean(dganim, trim=0.25)  
[1] 22.96957
```

d. כמות הסוכר ש-10% מהדגנים מתחתיה – 3.38% (או 3.3% ע"פ השיטה שנלמדה בשיעור).

```
quantile(dganim, probs = c(0.1,0.9))  
10%    90%  
3.38   50.36
```

או בחישוב התואם את שיטת החישוב שנלמד בכיתה

```
quantile(dganim, probs = c(0.1,0.9), type=5)  
10%    90%  
3.3     50.4
```

e. כמות הסוכר ש-90% מהדגנים מתחתיה – ע"פ הסעיף הקודם: 50.36% (או 50.4% ע"פ השיטה שנלמדה בשיעור).

f. הפרש בין הדגנים הכי מתוקים והכי תפלים

```
diff(range(dganim))  
[1] 59.3
```

g. טווח בין רבעוני

```
IQR(dganim)  
[1] 36.5
```

MAD .h

```
mad(dganim)  
[1] 22.38726
```

i. סטיית התקן של כמות הסוכר

```
sqrt(var(dganim))  
[1] 19.45047
```

iii. חלון ההחלקה בו משתמש R בפקודה density הוא חלון גאוסיאני (Gaussian).  
(\*) הפרמטר של החלון הוא bandwidth והוא נשלט ע"י המתג bw.

.5

- i. גבהים – היסטוגרמת חלון נע.
- ii. מגדר (מין) – תרשים עוגה.
- iii. סוג אוטו – תרשים עמודות או תרשים עוגה.
- iv. ספירת תקלות במכשיר (בהנחה שהכוונה לאורך זמן) – היסטוגרמת חלון נע.

.6 . וקטור הציונים – 40 60 30 50 60 70

i. הציון הממוצע בקורס הוא  $51.66 = \frac{70 + 60 \cdot 2 + 50 + 30 + 40}{6}$ .

ii. האינדקס של החציון הוא  $3.5 = \frac{1}{2} \cdot 6 + \frac{1}{2} = q \cdot n + \frac{1}{2}$  ולכן החציון הוא הסיבה לכך שהחציון אינו אחת מהתצפיות הוא כיוון שמספר התצפיות הוא זוגי. לכן, החציון "נפל" בין 2 תצפיות. הפתרון של הבעיה, כאמור, הוא לבצע ממוצע של שני האיברים האמצעיים.

iii. האינדקס של אחוזן ה-25 הוא  $2 = 0.25 \cdot 6 + \frac{1}{2} = q \cdot n + \frac{1}{2}$  ושל האחוזן ה-75 הוא

$$5 = 0.75 \cdot 6 + \frac{1}{2} = q \cdot n + \frac{1}{2}. \text{ לכן האחוזן ה-25 הוא 40 והאחוזן ה-75 הוא 60.}$$

iv. כברירת מחדש התוכנה משתמש באופציה 7 ועושה ממוצע משוקלל ע"פ האינדקס הדרוש, בין שתי התצפיות השכנות. סוג החישוב התואם את מה שלמדנו בכיתה הוא סוג 5.

v. הממוצע יקפוץ ל-56.66.  $\frac{100 + 60 \cdot 2 + 50 + 30 + 40}{6}$ . החציון לא ישתנה.

vi. במקרה כזה הממוצע גם ישאף לאינסוף. החציון לעומת זאת לא ישתנה. מנתונים אלו ניתן להסיק שהחציון יבטא טוב יותר את ביצועי הכיתה כיוון שרוב הכיתה קיבלה ציונים הנמוכים בהרבה מאינסוף.

vii. המרצה יכול לתת אינסוף כציון לעד מחצית מתלמידי הכיתה מבלי לפגוע בחציון (ליתר דיוק עד ל  $\frac{n-1}{2}$  מהתלמידים). תכונה זו נקראת נקודת השבירה.

a. סטיית התקן

$$\begin{aligned} & \left\{ \frac{1}{6-1} \cdot [(70-51.66)^2 + (60-51.66)^2 + (50-51.66)^2 + (30-51.66)^2 + (60-51.66)^2 + \right. \\ & \left. (40-51.66)^2] \right\}^{0.5} = \left\{ \frac{1}{5} \cdot [18.33^2 + 8.33^2 + (-1.66)^2 + (-21.66)^2 + 8.33^2 + \right. \\ & \left. (-11.66)^2] \right\}^{0.5} = \left\{ \frac{1}{5} \cdot [336.11 + 69.44 + 2.77 + 469.44 + 69.44 + 136.11] \right\}^{0.5} = \\ & \sqrt{\frac{1}{5} \cdot 1083.31} = \sqrt{216.662} = 14.71 \end{aligned}$$

b. IQR

$$q_{0.75} - q_{0.25} = 60 - 40 = 20$$

c. MAD

$$\begin{aligned} \text{median}\{ |70-55| + |60-55| + |50-55| + |30-55| + |60-55| + |40-55| \} &= \text{median}\{15, 5, 5, 25, 5, 15\} \\ &= [3 + 1 - 3.5]k_3 + [3.5 - 3]k_4 = [0.5] \cdot 5 + [0.5] \cdot 15 = 10 \end{aligned}$$

a. סטיית התקן

$$\begin{aligned} & \left\{ \frac{1}{6-1} \cdot [(70-51.66)^2 + (60-51.66)^2 + (50-51.66)^2 + (0-51.66)^2 + (60-51.66)^2 + \right. \\ & \left. (40-51.66)^2] \right\}^{0.5} = \left\{ \frac{1}{5} \cdot [18.33^2 + 8.33^2 + (-1.66)^2 + (-51.66)^2 + 8.33^2 + \right. \\ & \left. (-11.66)^2] \right\}^{0.5} = \left\{ \frac{1}{5} \cdot [336.11 + 69.44 + 2.77 + 2669.44 + 69.44 + 136.11] \right\}^{0.5} = \\ & \sqrt{\frac{1}{5} \cdot 3283.31} = \sqrt{656.662} = 25.62 \end{aligned}$$

b. IQR

$$q_{0.75} - q_{0.25} = 60 - 40 = 20$$

c. MAD

$$\begin{aligned} \text{median}\{ |70-55| + |60-55| + |50-55| + |0-55| + |60-55| + |40-55| \} &= \text{median}\{15, 5, 5, 55, 5, 15\} \\ &= [3 + 1 - 3.5]k_3 + [3.5 - 3]k_4 = [0.5] \cdot 5 + [0.5] \cdot 15 = 10 \end{aligned}$$

מבין שלושת המדדים המדד שלדעתי מייצג הכי טוב את פיזור הציונים בכיתה הוא IQR כי הוא אינו רגיש לחריגים כמו סטיית התקן, אך בו בזמן יותר מושפע מהשינויים מאשר MAD (נקודת השבירה שלו יותר נמוכה מאשר נקודת השבירה של MAD).

7. המשתנים המוצגים בתרשים הם:

- i. אזור גיאוגרפי - מוצג ע"י צבע העיגול. זהו משתנה קטגוריאלי (המכיל את האזורים השונים בעולם כקטגוריות).
- ii. הכנסה לנפש - מיוצג ע"י ציר X. זהו משתנה רציף.
- iii. תוחלת חיים צפויה בלידה - מיוצג ע"י ציר Y. זהו משתנה רציף.
- iv. גודל האוכלוסייה - מיוצג ע"י גודל העיגול. זהו משתנה רציף.
- v. זמן - מיוצג ע"י אנימציה בגרף (תנועה של הנתונים). זהו משתנה רציף.



c. מדד זה אינו מתאים. נראה זאת ע"י דוגמה נגדית:

```
> x = c(10, 10, 40, 50, 50, 10)
> y = c(15, 10, 40, 50, 50, 10)
> quantile(x, probs = c(0.75))/quantile(x, probs = c(0.25), type=5)
4.75
> quantile(y, probs = c(0.75))/quantile(x, probs = c(0.25), type=5)
4.75
```

הקטנו את הפיזור של הנתונים אך המדד לא השתנה ולכן הוא לא מתאים.

ii. (\*) יתרון של המדד הוא בכך שהוא פחות פגיע לחריגים (היות ומדובר בחציון ולא ממוצע). עם זאת, זהו גם חיסרון מסוים היות וייתכן שישנו פער גדול מאוד בין הרבעון העליון לעומת 3 הרבעונים הנותרים, ובמקרה כזה, לפער הנ"ל לא יהיה ביטוי במדד.

10. הפונקציה מחשבת נתונים המייצגים את הגרף החסין עבור 2 וקטורים של נתונים: x ו-y.

.Arl - b0

.Brl - b1

pred - ערכי הקו החסין בהתאם לנקודות x.

resid - הפרשי הערכים בין ערכי y לקו החסין.

x - וקטור x ממזין.

xb - חציון שליש תחתון של וקטור x.

yb - חציון שליש תחתון של וקטור y.

xt - חציון שליש עליון של וקטור x.

yt - חציון שליש עליון של וקטור y.