סטטיסטיקה / תרגיל 2#

אריאל סטולרמן
קבוצה 03

(1)

	סטודנטים	סגל	סה״כ
רכב אמריקאי	107	91	198
רכב אירופאי	212	120	332
רכב יפני	130	170	300
סה״כ	449	381	830

(i) ההתפלגות השולית של רכבים לפי מדינת ייצור היא:

- רכב אמריקאי: 23.85%
- רכב אירופאי: 40%
- רכב יפני: 36.15%

ההתפלגות השולית של הרכבים בהתניה על תפקיד הבעלים:

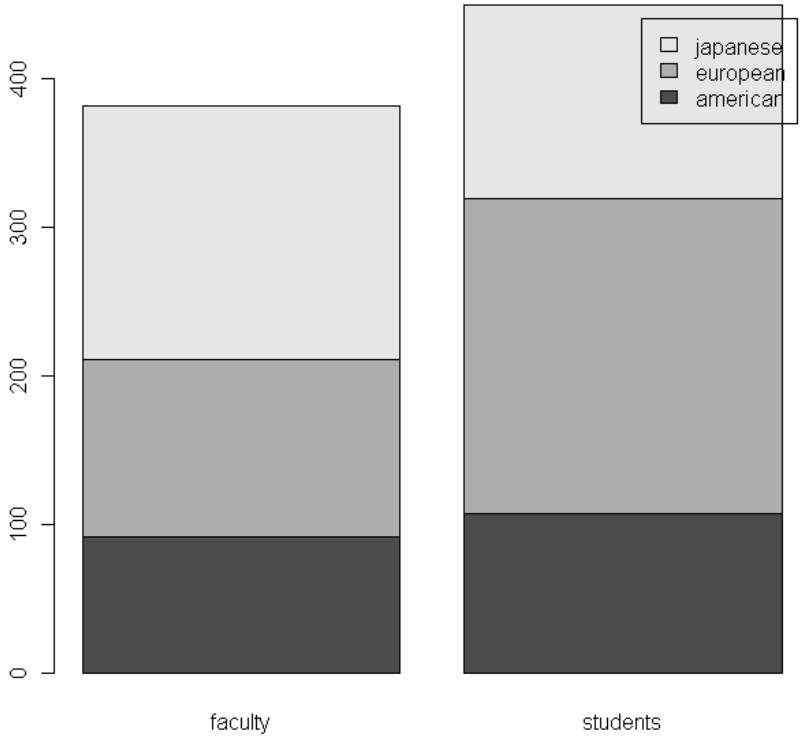
```
> owners=c(rep('students', 449), rep('faculty', 381))
>  country_man=c(rep('american',  107),  rep('european',  212),  rep('japanese',  130),
rep('american', 91), rep('european', 120), rep('japanese', 170))
> vehicles_table=table(country_man, owners)
> print(vehicles_table)
           owners
country_man faculty students
   american      91      107
   european     120      212
   japanese     170      130
> prop.table(vehicles_table, 2)
           owners
country_man   faculty  students
   american 0.2388451 0.2383073
   european 0.3149606 0.4721604
   japanese 0.4461942 0.2895323
```

מכאן מקבלים כי:

- חלוקת ארץ ייצור הרכבים עבור בעלים סטודנטים: 23.83% אמריקאי, 47.22% אירופאי, 28.95% יפני.
- חלוקת ארץ ייצור הרכבים עבור בעלים מהסגל: 23.88% אמריקאי, 31.5% אירופאי, 44.62% יפני.

(ii) להלן תרשים להצגת הנתונים:

```
> barplot(vehicles_table, legend.text=T)
```



(2)

(i) – דוגמא אפשרית לפרדוקס סימפסון - יתכן שעונה שעברה שיחקו רק שחקנים טובים, והעונה משחקים רק השחקנים הכי פחות טובים, שלמרות שהשתפרו, עדיין לא הגיעו למה שהיו פעם השחקנים הטובים – כלומר ההתניה בהסתכלות על כלל הקבוצה לא היתה נכונה, ולא הסתייחסה לאחוזי ההשתתפות של כל שחקן.

(ii) – פשוט בלתי אפשרי.

(iii) – אפשרי אך לא בגלל הפרדוקס – אין קשר בין נתוני ההרשמה ובין נתוני הקבלה – כלל אין כאן פרדוקס.

(3)

(i) להלן 3 דוגמאות לגדלים הנמדדים בסולם לוגריתמי:

- סולם ריכטר למדידת עוצמתן של רעידות אדמה (עוצמתה של רעידת האדמה החזקה ביותר שנמדדה (9.5 בסולם ריכטר) גדולה פי 1,600,000,000 מזו של הרעידה החלשה ביותר שניתן למדוד).
- דציבל למדידת עוצמת הצליל שהאוזן האנושית קולטת.
- מדד pH למדידת חומציות של תמיסה.

(ii) גובהו של אדם בגובה 175 ס״מ לפי סולם לוגריתמי בבסיס 2 הוא 7.45 ס״מ

(iii) אם גובהו של אדם אחר פי שניים משלי בסולם ללינארי, את גובהי כ-x נקבל כי ההפרש בסולם לוגרתימי הוא:

$$\log_2(2x) - \log_2(x) = \log_2(2x/x) = \log_2(2) = 1$$

כיוון שהתשובה אינה תלויה ב-x, אזי לדעת את גובהי האמיתי אינו פרט שצריך לדעת לצורך החישוב.

(iv) בבסיס 10:

$\log_{10}10(2x) - \log_{10}(x) = \log_{10}(2x/x) = \log_{10}(2) = 0.301$

(v) נסמן את הגובה שלי ב-x, ולפי הנתונים מתקיים:

$\log_2(x+20) - \log_2(x) = \log_2([x+20]/x) = \log_2(1 + 20/x)$

קיבלנו תוצאה תלויה ב-x ולפיכך לדעת את גובהי האמיתי הוא כן פרט שחשוב לדעת לצורך החישוב.

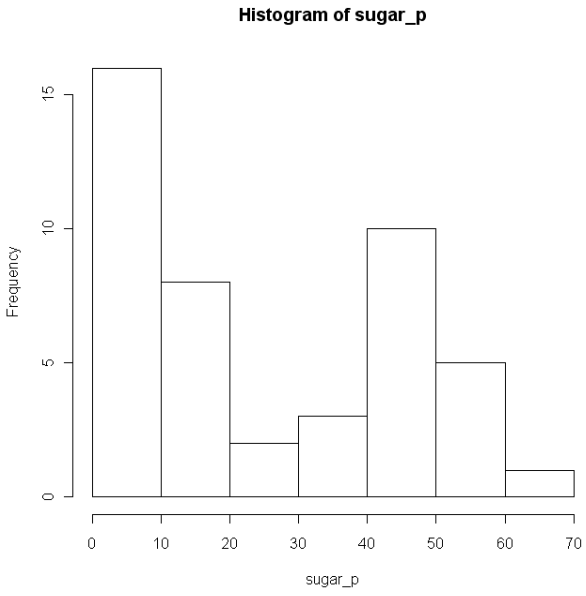
(2 השניה)

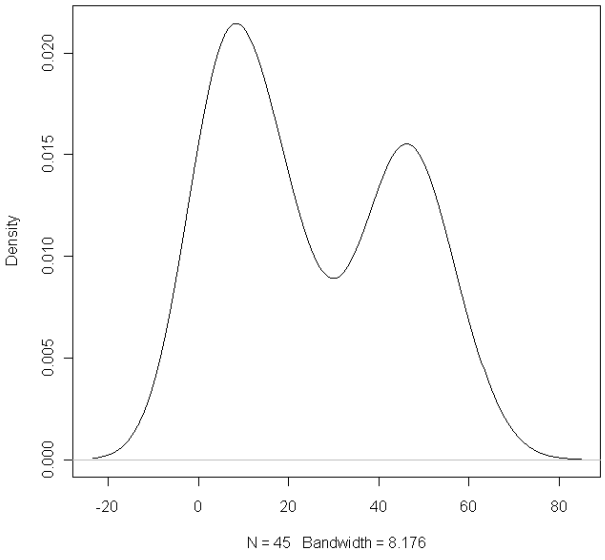
(i)

(a)

```
>
sugar_p=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)
> hist(suger_p)
> plot(density(sugar_p), main='')
```

היסטוגרמה ותרשים צפיפות לנתונים:





(b) נראה כי ישנן 2 קבוצות עיקריות.

(c) סיבה אפשרית לחלוקה יכולה להיות שחלק מדגני הבוקר הינם דגנים מסוכרים (בעיקר לילדים), וחלק הם דגני בריאות בהם אחוז סוכר נמוך מהראשונים.

(ii)

(a) מתיקות ממוצעת:

```
> mean(sugar_p)
[1] 24.96
```

(b) מתיקות חציונית:

```
> median(sugar_p)
[1] 18.4
```

(c) ממוצע קצוץ $\alpha=25$:

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

(d) כמות הסוכר ש-10% מהדגנים מתחתיה והיתר מעליה:

```
> quantile(sugar_p, 0.1)
 10%
3.38
```

(e) כמות הסוכר ש- 90% מהדגנים מתחתיה והיתר מעליה:

```
> quantile(sugar_p, 0.9)
  90%
50.36
```

(f) ההפרש בין הדגנים הכי מתוקים להכי תפלים:

```
range(sugar_p)[2]-range(sugar_p)
 [1] 59.3
```

(g) טווח בין רבעוני:

```
> IQR(sugar_p)
[1] 36.5
```

(h) MAD:

```
> mad(sugar_p)
[1] 22.38726
```

חישוב זה שונה מאשר חישוב ע"פ הנוסחה הנלמדה בכיתה:

```
> median(abs(sugar_p-median(sugar_p)))
[1] 15.1
```

(i) סטיית התקן של כמות הסוכר - בשאלה מצויינת הפקודה var, אך זו פקודה לחישוב השונות, בעוד הפקודה לחישוב סטיית התקן היא sd (והיא שווה גם לשורש השונות):

```
> var(sugar_p)
[1] 378.3206
> sd(sugar_p)
[1] 19.45047
```

(iii) חלון ההחלקה בו משתמש R בפקודה `density` כברירת מחדל הוא `"gaussian"`.

(3 השניה)

סוג נתונים	תרשים	סוג נתונים	תרשים
גבהים	היסטוגרמה, ענף וגזע	סוג אוטו	תרשים עוגה
מגדר	תרשים עמודות	ספירת תקלות במכשיר	היסטוגרמת חלון נע

הערה: זו חלוקה מינימלית אפשרית, יתכנו שימושים עבור חלק מהתרשימים ליותר מסוג נתונים אחד.

(4)

(i) הציון הממוצע:

$$(70 + 60 + 50 + 30 + 60 + 40)/6 = 51.66$$

(ii) הציון החציוני:

30 40 50 ↓ 60 60 70

החציון יהיה במקרה זה 55 (ממוצע פרופורציוני של 50 ו-60, שני הציונים הקרובים אליו ביותר). זה קורה כאשר מספר הנתונים זוגי, ופתרון לכך יכול להיות כמו לעיל – לקיחת המקסימום של החצי התחתון של הנתונים ומינימום של החצי העליון של הנתונים ולמצע באופן פרופורציוני (בהתייחסות למרחק שלהם מהיכן שאמור להיות ה"אמצע האמיתי"). להלן פירוט החישוב:

$q=0.5,\ n=6,\ x=(30,40,50,60,60,70)$:

$i = \lfloor qn + 0.5 \rfloor = 3$ ➔ $x_{0.5} = (3.5 - 3)*50 + (4-3.5)*60 = 55$

(iii)

נשתמש בנוסחה המתוארת לעיל בכדי למצוא את האחוזון ה-25 וה-75:

$q=0.25,\ n=6,\ x=(30,40,50,60,60,70)$

$i = qn + 0.5 = 2$ ➔ $x_{0.25} = \mathbf{40}$ - זהו האחוזון ה-25

$q=0.75...$

$i = qn + 0.5 = 5$ ➔ $x_{0.75} = \mathbf{60}$ - זהו האחוזון ה-75

(iv) כבירית מחדל התוכנה בוחרת ב-type 7 שהוא $(k-1)/(n-1)$. השיטה בתוכנה המתאימה לגישה שהוצגה בהרצאה היא כנראה type 5, לפחות על הנתונים בשאלה זו. Type 5 היא היחידה שהניבה תוצאות זהות לנוסחה עבור שברונים שונים (ולא רק "חלקים" כמו 0.5, 0.25...). על 0.5 למשל, עוד types הניבו תוצאות זהות לנוסחה, ועל נתונים אחרים, type 5 הניבה תוצאות שונות מהנוסחה, כך שקשה להתחייב על כך ש-type 5 היא המתאימה לנוסחה מהכיתה.

(v) הממוצע במקרה שנחליף את 70 ב-100 יהיה 56.66, כלומר הוא יגדל. לעומתו, החציון לא ישתנה וישאר 55.

(vi) כיוון שנקודת השבירה של ממוצע היא $1/n$, שינוי תוצאה אחת ל-∞ תשנה את הממוצע גם ל-∞. החציון, שוב, לא ישתנה במקרה זה, ולפיכך **החציון** מייצג במקרה זה טוב יותר את ביצועי בכיתה.

(vii) המרצה יכול לתת סה"כ 2 "אינסופים" במקרה זה שהם 33.3% מהציונים (וגם זה תחת הנחה שמשנה בנוסף ל-70 אך ורק את אחד ה-60) , וכל שינוי נוסף כבר יפגע בחציון. תכונה זו של המדד נקראת נקודת שבירה.

(viii)

סטיית התקן:

$$\sqrt{[(1/(n-1))*\Sigma(x_i-x)^2]} = \sqrt{[(1/5)*(1083.33)} = \mathbf{14.7196}$$

MAD:

$$\text{median}\{|x_i - \text{median}(x)|\} = \text{median}\{25, 15, 5, 5, 5, 15\} =$$
$$= \text{median}\{5, 5, 5, 15, 15 ,25\} = 0.5*5 + 0.5*15 = \mathbf{10}$$

IQR: טווח בין רבעונים:

$$x_{0.75} - x_{0.25} = 60 - 40 = \mathbf{20}$$

(ix) עבור ציון נמוך ביותר 0, נקבל את התוצאות הבאות (הממוצע השתנה ל-46.66):

- סטיית תקן: 25.0333
- MAD: 10 = median{55,15,5,5,5,15} – **ללא שינוי**
- IQR: 20 – **ללא שינוי**

לדעתי MAD ו-IQR מייצגות את פיזור הציונים באופן הכי טוב, כיוון שלא הגיבו לשינוי, ולכן מייצגים את כלל הכיתה באופן טוב יותר.

(5) אתייחס כאן למספרים טבעיים כרציפים:

משתנה	סוג משתנה	אופן הצגה
הכנסה פר אדם	רציף	ערך קורדינטת x
תוחלת חיים	רציף	ערך קורדינטת y
גודל האוכלוסיה במדינה	רציף	קוטר, כמו כן הסקאלה משתנה בהתאם לגודל האוכלוסיה (דיוק יחידה, מליון, מיליארד...)
אזור גיאוגרפי/יבשת	קטגוריאלי	צבע – צהוב לאמריקה, כחול לאפריקה...

(6)

(i) ממוצע קצוץ α או חציון

(ii) סטיית התקן האמפירית

(iii) מדד לנטיה כלשהו (לבדוק נטיה ימינה) כגון Yule

(iv) MAD

(v) מקסימום על כל ממוצעי הציונים של התלמידים בכל כיתה – הממוצע הגבוה ביותר מבין שתי הכיתות הוא של התלמיד המוכשר ביותר (כאן נתייחס לממוצע ציונים אישי כמדד ליכולותיו של תלמיד)

(vi) חישוב $x_{0.2}$ על כל ציוני התלמיד– אם מתקבל ציון שלילי אז הסיכוי גדול או שווה ל-20%, אחרת הוא קטן מ-20% (הסקה מציוניו הנוכחיים של התלמיד על סיכוייו להצלחה בעתיד).

(7)

(vii)

(a) log(MAD(x))-log(median(x)) היא הצעה טובה:

= log[MAD(x)/median(x)]

- המונה והמכנה מיוצגים באותן יחידות בדיוק וכל שינוי יחידות ישנה את המונה והמכנה באותו יחס כך שהפרופורציה תישמר.
- הגדלת כל התצפיות בקבוע אינה משנה את המונה כיוון ש-MAD הוא חציון של הפרשי התצפיות מהחציון, והאחרונים לא ישתנו עם הגדלת כל התצפיות בקבוע. לעומת זאת, המכנה כן ישתנה כיוון ש-median ישתנה – הוא יגדל, ולכן כל המנה תקטן. כיוון ש-log פונקציה מונוטונית, אזי המדד יקטן.
- הקטנת פיזור ההכנסה תקרב את כל התצפיות לחציון, ולכן ההפרשים מהחציון יקטנו, ולפיכך ה-MAD יקטן. לעומת זאת, החציון לא ישתנה (או ישתנה מעט) כאשר נקרב את התצפיות אליו (שזה למעשה צמצום הפיזור). לכן המדד יקטן.

(b) var(x)/mean(x) הינה הצעה לא טובה כיוון שלא מקיים את הדרישה הראשונה. כל הגדלה פי קבוע (נניח הכפלה ב-4 למעבר מדולרים לשקלים) תגדיל את המונה פי אותו קבוע בריבוע, ואת המכנה רק פי הקבוע. לכן המדד ישתנה.

(c) Q3/Q1 הינה הצעה טובה:

- המונה והמכנה מיוצגים באותן יחידות בדיוק (כי שניהם תצפיות), וכל שינוי יחידות ישנה את המונה והמכנה באותו יחס כך שהפרופורציה תישמר.
- תוספת שווה לכולם תגרור תוספת שווה למונה ומכנה יחד. כיוון ש-Q3 גדול (נניח ממש) מ-Q1 אזי חלוקה זו גדולה מ-1, וכל תוספת זהה למונה ולמכנה תשאיף את הביטוי ל-1 מלמעלה, משמע המדד יקטן כרצוי.
- הקטנת פיזור ההכנסה גם כן תקרב את Q3 מטה ואת Q1 מעלה, ולכן תשאיף את המדד ל-1 מלמעלה, משמע המדד יקטן כרצוי.

(8)

הפונקציה מקבלת שני וקטורים x,y ומחשבת את **הקו החסין** לקורדינטות המיוצגות ע״י שני הוקטורים האלה. הפונקציה מחזירה:

- b0: חיתוך הקו החסין עם ציר ה-y.
- b1: שיפוע הקו החסין.
- pred: קורדינטות ה-y של כל איברי x על הקו החסין לפי סדר עולה של x.
- resid: פיזור – הפרשי כל אברי y מהקו החסין לפי סדר עולה של x.
- x: הצגת וקטור x (לפי סדר עולה)
- xb: קורדינטת ה-x של חציון השליש הראשון של הנתונים.
- yb: קורדינטת ה-y של חציון השליש הראשון של הנתונים.
- xt: קורדינטת ה-x של חציון השליש האחרון של הנתונים.
- yt: קורדינטת ה-y של חציון השליש האחרון של הנתונים.