תרגיל בית 2

מעבר לעזרה שבתוכנה עצמה, מומלץ להיעזר באתר: http://wiki.r-project.org/rwiki/doku.php

ניתן ורצוי להתשמש בפורום הקורס להתייעצות.

יש לצרף את הקוד ששימש לפתרון.

סעיפים בדרגת קושי גבוהה יותר סומנו בכוכבית(*). כדאי להשתדל לענות עליהם אבל **אין הכרח**.

הנתונים בתרגיל בדויים.

---

1. סקר של מדור בטחון של האוניברסיטה בדק את סוגי הרכב שבמגרשי החנייה בקמפוס. להלן התוצאות:

	סטודנטים	סגל
רכב אמריקאי	107	91
רכב אירופאי	212	120
רכב יפני	130	170

   i. מהי ההתפלגות השולית[1] של רכבים לפי מדינת ייצור (אמריקאי/אירופאי/יפני)? ובהתנייה על תפקיד הבעלים (סטודנט/סגל)?
   אפשר לפתור ידנית אך מומלץ ב R על ידי prop.table.

   ii. הציגו את הנתונים בדרך הגרפית הנראית לכם הכי אינפורמטיבית (ייתכנו מספר דרכים מתאימות אבל אחת תספיק לצורך התרגיל).

2. סווג את שלושת הדוגמאות הבאות לפי:
   A- דוגמה אפשרית לפרדוקס סימפסון.
   B- פשוט בלתי אפשרי.
   C- אפשרי, אבל לא בגלל הפרדוקס.

   i. בקבוצת הכדורסל של מכבי תל אביב, בית שנת 2001 ל 2002, לא התחלפו שחקנים ואחוז הקליעה של **כל** השחקנים מן השדה השתפר. מאידך, אחוז הקליעה הכלל של הקבוצה ירד.

   ii. יוסי קיבל ציון גבוה יותר מרחל בשיעורי הבית, הבחינה והפרוייקט בקורס. מאידך, ציונה הסופי של רחל גבוה יותר (הציון הסופי נקבע כמובן אך ורק לפי שלושת ציונים אלו).

   iii. במחקר נוסף באוניברסיטת ברקלי, התגלה שמבין הפונים אחוז הנרשמים ללימודי משפטים גבוה יותר מאשר אחוז הנרשמים ללימודי מתימטיקה. מאידך, אחוז המתקבלים במתימטיקה גבוה יותר מאשר במשפטים.

---

1 כלומר איזה אחוז מהרכבים מגיע מכל מדינה?

3. טרנספורמציית לוג היא לא רק פרוצדורה טכנית להתבונן בקבצי נתונים אלא גם בעלת הצדקה נורמטיבית במקרים רבים.

i. תנו דוגמאות לשלושה גדלים אשר נמדדים בדרך כלל בסולם לוגריתמי.
אפשר להתחיל את החיפוש כאן- http://en.wikipedia.org/wiki/Logarithmic_scale

ii. מהו הגובה של אדם שגובהו 175 ס"מ בסולם לוגריתמי (בבסיס 2)?

iii. מה **הפרש** הגבהים בינכם ובין אדם שגובהו **פי 2** מכם בסולם לוגריתמי (בסיס 2)? האם בכלל צריך לדעת את הגובה שלכם בשביל לענות?

iv. מה יהיה הפרש זה בסולם לוגריתמי בבסיס 10?

v. מה יהיה הפרש הגבהים (בסולם לוגריתמי) בינכם ובין אדם שגבוה מכם בעשרים סנטימטרים? האם בכלל צריך לדעת את הגובה שלכם בשביל לענות?

2. להלן כמות הסוכר (באחוזים מהמשקל) בסוגים שונים של דגני בוקר
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. שרטטו היסטוגרמה ותרשים צפיפות לנתונים ב- R . העזרו בפקודות hist, density, plot

(a) האם תסכימו עם האמירה שיש מספר קבוצות דגני בוקר?

(b) כמה קבוצות יש?

(c) הציעו סיבה אפשרית לחלוקה זו.

ii. חשבו את

(a) המתיקות הממוצעת בדגני בוקר- mean

(b) המתיקות החציונית- median

(c) ממוצע קצוץ $(\alpha=0.25)$ - mean

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

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

(f) ההפרש בין הדגנים הכי מתוקים והכי תפלים- range

(g) טווח בין רבעוני- iqr

(h) MAD - mad [2]

(i) סטיית תקן של כמות הסוכר- var

iii. מהו חלון ההחלקה בו משתמש R בפקודה density כברירת מחדל?
(*)מהו הפרמטר של חלון זה וכיצד שולטים עליו?

---

2 הפקודה mad ב R מחזירה תוצאה שונה מהצפוי. בשביל להבין את מקור ההבדל, מוטב לחכות לדיון על "עקיבות" של אומדים.

3. התאימו בין סוג נתונים לתרשים שיתאים להצגתם:
   i. סוג נתונים: גבהים, מגדר (מין), סוג אוטו, ספירת תקלות במכשיר.
   ii. תרשימים אפשריים: היסטוגרמה(histogram), תרשים עמודות (barplot), תרשים עוגה (pie), היסטוגרמת חלון נע (density plot), תרשים ענף וגזע (stem and leaf plot).

4. תרגיל זה נועד לתת תחושה של מדדים/תמציות נתונים דרך הידיים. מאוד **לא** מומלץ לפתור אותו במחשב.
להלן וקטור הציונים בקורס היוקרתי "תהליכים סטוכסטיים מרחביים" (אשר נלמד כזכור רק בטכניון): 40 60 30 50 60 70
   i. מהו הציון הממוצע?
   ii. מהו הציון החציוני? בוודאי תשימו לב שהחציון אינו אחת מהתצפיות. מתי זה קורה? הציעו פתרון לבעייה.
   iii. מהו האחוזון ה 25 וה 75 של הציונים? שימו לב לבעייה דומה לסעיף קודם. הציעו פתרון.
   iv. תוכנת R מחשבת אחוזונים בדרכים רבות. הציצו בעזרה של הפקודה quantile... מה עושה התוכנה כברירת מחדל? ומה השיטה בתוכנה, המתאימה לגישה שהוצגה בהרצאה?
   v. מה יהיה הממוצע אם הציון הגבוה הוא למעשה 100 ולא 70? מה ייקרה לחציון?
   vi. מה ייקרה לממוצע אם המבחן הכי טוב (עם הציון הכי גבוה) היה כל כך טוב שהמרצה נתן לו ציון אינסופי (לדוגמה אם התלמיד מצא מודל שמסביר את השפעת האם הפולנייה בזוגיות). מה יהיה החציון?איזה משני מדדים אלו מייצג טוב יותר את ביצועי הכיתה?
   vii. איזה אחוז של "אינסופים" יכול המרצה לתת לפני שיהיה שינוי בציון החציוני של הכיתה? כיצד נקראת תכונה זו של המדד?
   viii. מהי סטיית התקן של הציונים? מהו ה IQR? מהו ה MAD?
   ix. חשבו את הסעיף הקודם אם הציון הכי נמוך הוא אפס? איזה מהם מייצג טוב יותר את הפיזור בכיתה (לדעתכם)?

5. ויזואליזציה של הרבה משתנים בו זמנית:
היכנסו לאתר http://www.gapminder.org/world
לאחר שהשתעשתם בתרשימים, החלפת המשתנים בצירים, מעבר מסקאלה לוגריתמית ללינארית, והתקדמו בזמן...
איזה משתנים מוצגים בתרשים ומה הסוג של כל אחד מהם? רציף? קטיגוריאלי? כיצד מוצג כל אחד מהם? (רמז- ישנם ארבעה לא כולל משתנה הזמן).

6. איזה מדד/תמצית תעדיפו בשביל לענות כמותית על השאלות הבאות?
ייתכן יותר ממדד אחד (תשובות יצירתיות יתקבלו בברכה):

i. הביצועים של כיתה A טובים משל כיתה B.
ii. תלמידי כיתה A דומים יותר ביכולותיהם (בתוך הכיתה) מאשר תלמידי כיתה B.
iii. בכיתה A יש כמה גאונים ואין טמבלים כבדים.
iv. במבחן הבית האחרון כל כיתה A עבדה יחד וכולם עשו את אותן הטעויות.
v. התלמיד הכי מוכשר בשכבה נמצא בכיתה B.
vi. הסיכוי שתלמיד יכשל במבחן גדול מ 20%.

7. חוקרי כלכלה ומדעי החברה דורשים שמדד לפער החברתי על התפלגות ההכנסות באוכלוסייה יקיים לפחות את שלוש הדרישות הבאות:
- שינוי ביחידת המדידה של ההכנסות (מעבר משקלים לדולרים) לא ישנה את ערכו של המדד.
- תוספת הכנסה שווה לכל האוכלוסייה תוריד את ערכו של המדד.
- הקטנת פיזור ההכנסה תוריד את ערכו של המדד.

vii. אלו מההצעות הבאות למדד מקיים דרישות אלה? נמקו:

(a) $\log(MAD(x)) - \log(median(x))$

(b) $variance(x)/mean(x)$

(c) Q_3/Q_1

viii.(*) מדד העוני במדינת ישראל הוא מחצית ההכנסה החציונית במדינה. מהן היתרונות והחסרונות של המדד. נמקו.

8. מה עושה הפונקציה הבאה?

```
myf  = function(x, y) {
    x ← x[!is.na(y)]
    y ← y[!is.na(y)]
    o ← order(x)
    x ← sort(x)
    y ← y[o]
    n ← length(x)
    n3 ← round(n/3)
    xb ← median(x[c(1:n3)])
    yb ← median(y[c(1:n3)])
    xt ← median(x[c((n - n3 + 1):n)])
    yt ← median(y[c((n - n3 + 1):n)])
    b1 ← (yt-yb)/(xt-xb)
    b0 ← mean(yt,yb)-b1*mean(xt,xb)
    pred ←b0 + b1 * x
    mr ← median(y - pred)
    b0 ← b0 + mr
    pred ← b0 + b1 * x
    resid ← y - pred
    return(list(b0=b0, b1=b1, pred=pred, resid=resid, x=x, xb=xb, yb=yb, xt=xt, yt=yt))
}
```