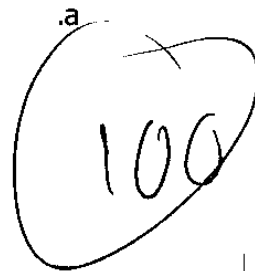


# סטטיסטיקה למדעי המחשב – פתרון תרגיל 5

1.

- i. אוכלוסייה
- ii. מדגם מקרי
- iii. מדגם
- iv. מדגם מקרי
- v. מדגם
- vi. אוכלוסייה (בהנחה שהכוונה היא שאלו עוגות הכבש היחידות שאי פעם ייצרו במאפייה, אחרת, מדובר במדגם מקרי).

a. 

אוכלוסייה (בהנחה שהכוונה היא שאלו עוגות הכבש היחידות שאי פעם ייצרו במאפייה, אחרת, מדובר במדגם מקרי).

b. X - מספר הצימוקים בעוגה.  
נחשב תוחלת ושונות:

$$E[X] = \frac{54 + 60 + 44 + 56}{4} = \frac{214}{4} = 53.5$$

$$V(X) = \frac{(54 - 53.5)^2 + (60 - 53.5)^2 + (44 - 53.5)^2 + (56 - 53.5)^2}{4} =$$

$$\frac{0.25 + 42.25 + 90.25 + 6.25}{4} = \frac{139}{4} = 34.75$$

c.

i. כדי לבחור 2 עוגות שונות, בפעם הראשונה נוכל לבחור 4 עוגות, לאחר מכן מתוך 3. סה"כ 12 אפשרויות. כיוון שסדר הבחירה אינו משנה, נחלק במספר הסידורים האפשריים ונקבל שקיימות **6 אפשרויות**. דגימה שכזו אינה מקיימת את הגדרות הדגימה המקרית כיוון שעל פי ההגדרה עלינו לבצע בחירה עם החזרה כדי לשמור על עיקרון אי-התלות.

ii. אם ניתן לבחור עוגה פעמיים, הרי שבכל בחירה קיימות לנו 4 אפשרויות. לכן, סה"כ קיימות **16 אפשרויות**. בחירה שכזו מקיימת את הגדרות הדגימה המקרית שכן דרך זו שומרת על אי תלות בין התצפיות.

iii.

A.

ממוצע גיאומטרי	ממוצע חשבוני	סיכוי	עוגה ב'	עוגה א'
54	54	$\frac{1}{16}$	1	1
56.92	57		2	1
48.74	49		3	1
54.99	55		4	1
56.92	57		1	2
60	60		2	2
51.38	52		3	2
57.96	58		4	2
48.74	49		1	3
51.38	52		2	3
44	44		3	3
49.64	50		4	3
54.99	55		1	4
57.96	58		2	4
49.64	50		3	4
56	56		4	4

.B

$$E(T) = \frac{54 + 2 \cdot 57 + 2 \cdot 49 + 2 \cdot 55 + 60 + 2 \cdot 52 + 2 \cdot 58 + 44 + 2 \cdot 50 + 56}{16}$$

$$= \frac{856}{16} = 53.5$$

תוחלת האומד שווה לפרמטר אותו אנו רוצים לאמוד (תוחלת מספר הצימוקים בעוגה) ולכן אומד זה הוא **חסר הטיה**.

.C. נחשב את שונות האומד:

$$V(T) = \frac{0.25 + 2 \cdot 24.5 + 2 \cdot 40.5 + 2 \cdot 4.5 + 42.25 + 90.25 + 5}{16}$$

$$\frac{276.75}{16} = 17.296875$$

.D. נחשב MSE של האומד:

$$MSE(T) = bias^2(T) + Var(T) = 0 + 17.296875 = 17.296875$$

.E. (\*) נגדיר מדד אלטרנטיבי  $M + (T) = |bias(T)| + \sqrt{Var(T)}$   
נחשב את ערך המדד החדש:

$$M + (T) = |bias(T)| + \sqrt{Var(T)} = 0 + \sqrt{17.296875} = 4.15895$$

.iv. (\*) נחשב את MSE של הממוצע הגיאומטרי:

$$E(Tg) = \frac{54 + 2 \cdot 56.92 + 2 \cdot 48.74 + 2 \cdot 54.99 + 60 + 2 \cdot 51.38 + 2 \cdot 57.96 + 44 + 2 \cdot 49.64 + 56}{16}$$

$$= \frac{853.26}{16} = 53.32875$$

$$V(Tg) = \frac{290.25}{16} = 18.140625$$

$$MSE(Tg) = bias^2(Tg) + Var(Tg) = 0.029 + 18.140625 = 18.169625$$

קיבלנו מדד MSE גבוהה מהמדד של הממוצע החשבוני ולכן ממוצע גיאומטרי הוא אומד פחות טוב ממוצע חשבוני. כלומר, נעדיף את הממוצע החשבוני על פני ממוצע גיאומטרי.

2. נפתר בתרגול.

3.

a.

i.

$$E(Tiv) = E\left(\frac{\sum X_i}{n}\right) = \frac{1}{n} \sum E(X_i) = \frac{n}{n} E(X) = E(X) = p$$

$$bias(Tiv) = E(Ti) - p = p - p = 0$$

האומד חסר הטיה עבור  $p=0.5$  וכן עבור  $p$  כללי.

$$V(Ti) = V\left(\frac{\sum X_i}{n}\right) = \frac{1}{n^2} \sum V(X_i) = \frac{n}{n^2} V(X) = \frac{V(X)}{n} = \frac{p(1-p)}{n}$$

היות והאומד הוא חסר הטיה, MSE תלוי בשונות בלבד. כיוון שהשונות תלויה ב-n (כלומר גודל המדגם) כך שככל שנגדיל את n, ערך MSE ישאר ל-0, נקבע כי האומד עקיב.

ii. במקרה של התפלגות בינומית אין הבדל בין פרופורציית ההצלחות לממוצע.

iii.

$$E(T_{iii}) = E\left(\frac{\sum_{i=1}^n X_{i+1}}{n+2}\right) = \frac{1}{(n+2)} [E(\sum_{i=1}^n X_i) + 1] = \frac{1}{(n+2)} (\sum_{i=1}^n E(X_i) + 1) = \frac{np+1}{(n+2)}$$

$$bias(T_{iii}) = \frac{np+1}{(n+2)} - p = \frac{np+1-np-2p}{(n+2)} = \frac{1-2p}{n+2}$$

עבור  $p=0.5$  האומד אינו מוטה. לעומת זאת, עבור  $p$  כללי, האומד מוטה.

נמצא את שונות האומד:

$$V(T_{iii}) = V\left(\frac{\sum_{i=1}^n X_{i+1}}{n+2}\right) = \frac{1}{(n+2)^2} V\left(\sum_{i=1}^n X_{i+1}\right) = \frac{1}{(n+2)^2} V\left(\sum_{i=1}^n X_i\right) = \frac{p(1-p)}{(n+2)^2}$$

נשים לב ש-MSE מורכב מה-bias והשונות. שני ערכים אלו הולכים וקטנים ככל שגודל המדגם גדל (כיוון שקיים n במכנה). לכן גם MSE ישאר ל-0 כאשר המדגם גדל ומכאן שאומד זה הוא אומד עקיב.

iv.

$$E(T_{iv}) = E(X_1) = p$$

$$bias(T_{iv}) = E(T_{iv}) - p = p - p = 0$$

האומד חסר הטיה עבור  $p=0.5$  וכן עבור  $p$  כללי.

נמצא את שונות האומד

$$V(T_{iv}) = V(X_1) = p(1-p)$$

היות והאומד הוא חסר הטיה, MSE תלוי בשונות בלבד. כיוון שהשונות אינה תלויה ב-n (כלומר גודל המדגם) נקבע כי האומד אינו עקיב.

b. 1,0,0,1,1,0,1,0,0,0

i. מתוך 10 איברים במדגם, 4 הסתיימו בהצלחה. לכן, פרופורציית ההצלחות היא **0.4**.

ii. כאמור, במקרה זה הממוצע זהה לפרופורציית ההצלחות, ואכן הממוצע הוא **0.4**.

.iii

$$\frac{\sum_{i=1}^n X_i + 1}{10+2} = \frac{1+0+0+1+1+0+1+0+0+0+1}{12} = \frac{5}{12} = 0.4166$$

iv. תוצאת הניסוי הראשון היא 1 ולכן זהו גם ערך האומד.

אף אחד מהאומדים לא הצליח להגיע קרוב ממש לערך האמיתי של  $\theta$  (בהנחה שההפרש בין 0.4 ל-0.5 הוא גדול מאוד כשאר מסתכלים על התחום שבין 0 ל-1 בלבד). מבין כל האומדים, העומד השלישי, ניבא את התוצאה הקרובה ביותר ל- $\theta$  והעומד הרביעי ניבא את התוצאה הגרועה ביותר.

.4

$$v(x) = \frac{\theta^2}{12}, E(x) = \frac{\theta}{2} \quad .a$$

.i

$$E(T_1) = E(2X_1) = 2E(X_1) = 2E(x) = 2 \cdot \frac{\theta}{2} = \theta$$

$$V(T_1) = V(2X_1) = 4V(X_1) = 4 \cdot \frac{\theta^2}{12} = \frac{\theta^2}{3}$$

- תוחלת האומד  $T_1$  שווה לפרמטר ולכן זהו אומד **חסר הטיה**.

- היות וזהו אומד חסר הטיה, MSE שלו תלוי רק בשונות. כיוון שהשונות אינה תלויה בגודל המדגם, הרי שהאומד **אינו עקיב**.

.ii

$$E(T_2) = E\left(\frac{2}{n} \sum_{i=1}^n X_i\right) = \frac{2}{n} \sum_{i=1}^n E(X_i) = \frac{2}{n} \cdot n \cdot \frac{\theta}{2} = \theta$$

$$V(T_2) = V\left(\frac{2}{n} \sum_{i=1}^n X_i\right) = \frac{4}{n^2} \sum_{i=1}^n E(X_i) = \frac{4}{n^2} \cdot n \cdot \frac{\theta}{12} = \frac{\theta^2}{3n}$$

- תוחלת האומד  $T_2$  שווה לפרמטר ולכן זהו אומד **חסר הטיה**.

- היות וזהו אומד חסר הטיה, MSE שלו תלוי רק בשונות. כיוון שהשונות תלויה בגודל המדגם, וכן ככל שנגדיל את גודל המדגם, ערך השונות ירד ל-0 (משמע גם ערך MSE ירד) אזי זהו אומד **עקיב**.

.iii

$$E(T_3) = E(X_n^2) = V(X_n) + E^2(X_n) = \frac{\theta^2}{3}$$

$$V(T_3) = \frac{4}{45} \theta^4 \quad (\text{חושב בתרגול})$$

- תוחלת האומד  $T_3$  אינה שווה לפרמטר ולכן זהו **אינו אומד חסר הטיה**.

- MSE של אומד מורכב מריבוע הטיה והשונות. כיוון ששני נתונים אלו אינם תלויים ב-n, אומד זה **אינו עקיב**.

b. האומד השני הוא העדיף מבין השלושה. הסיבה לכך היא כיוון שהוא היחיד מבינם שהוא גם חסר הטיה וגם עקיב.

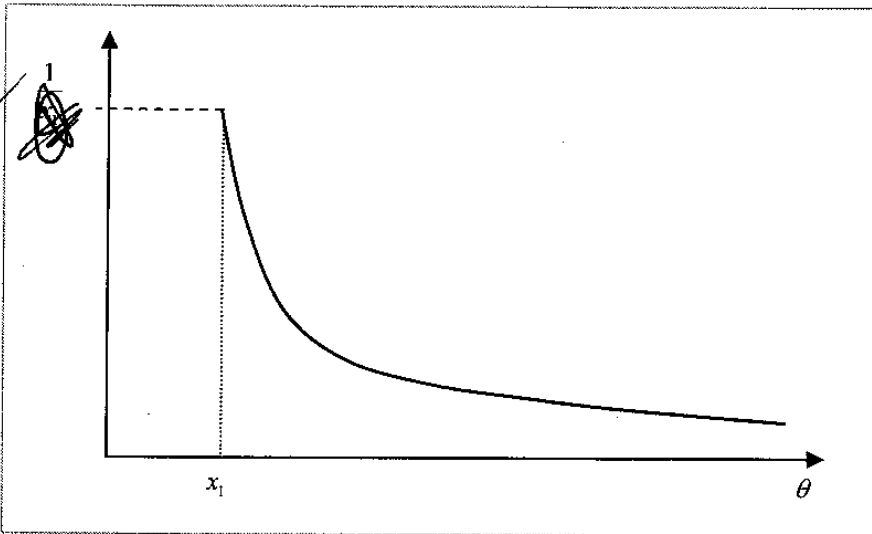
c.

i. מצב זה כמובן שאינו אפשרי כיוון שלא ייתכן שהסוללה תפעל יותר מאורך החיים המקסימאלי שלה.

ii. בצורה דומה, (בהנחה ואורך החיים של סוללה מתייחס למקסימום) גם במצב כזה לא ייתכן שערך הפרמטר יהיה קטן מהערך של אחת משתיים.

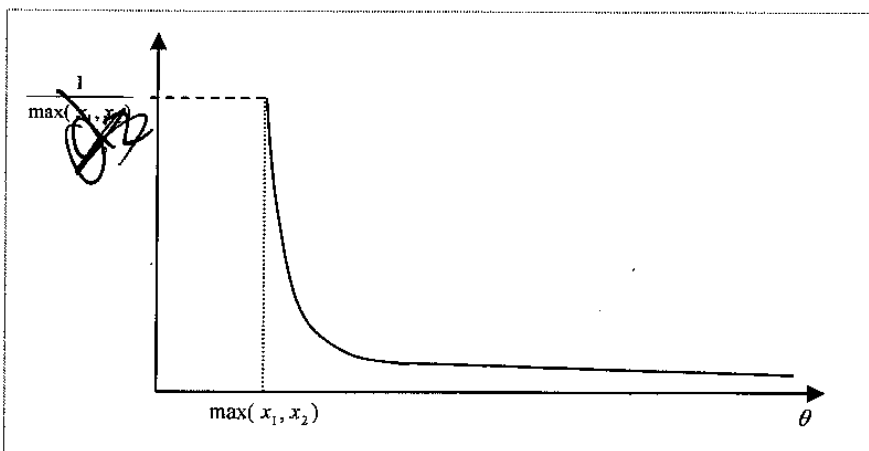
iii. נסתכל על פונקציית הצפיפות וממנה נשליך על פונקציית הנראות:

$$f_{\theta}(x) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \phi \\ 0 & \text{o.w.} \end{cases} \Rightarrow L(\theta) = \begin{cases} \frac{1}{\theta} & \theta \geq x_1 \\ 0 & \theta < x_1 \end{cases} \Rightarrow$$



iv. נסתכל על פונקציית הצפיפות וממנה נשליך על פונקציית הנראות:

$$f_{\theta}(x) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & \text{o.w.} \end{cases} \Rightarrow L(\theta) = \begin{cases} \frac{1}{\theta^2} & \theta \geq \max(x_1, x_2) \\ 0 & \theta < \max(x_1, x_2) \end{cases} \Rightarrow$$



d. אורך החיים המקסימאלי הוא **אומד עקיב**. הסיבה לכך היא כיוון שככל שנגדיל את גודל המדגם, יש סיכויי יותר גדול שנתקל בסוללה שאורך החיים שלה הוא  $\theta$ . כיוון שמבין כל התצפיות במדגם, האומד בוחר את הערך המקסימאלי, הרי שאם גודל המדגם ישאף לאינסוף הסיכוי שאחת הסוללות (לפחות) תהיה בעלת אורך חיים  $\theta$  הוא קרוב מאוד ל-1.

(\*) נחשב את תוחלת האומד כדי לבדוק האם הוא חסר הטיה:

$$F_{T_4}(x) = \begin{cases} \frac{x^n}{\theta^n} & 0 \leq x \leq \theta \\ 0 & o.w. \end{cases} \Rightarrow f_{T_4}(x) = \begin{cases} \frac{n \cdot x^{n-1}}{\theta^n} & 0 \leq x \leq \theta \\ 0 & o.w. \end{cases}$$

$$E(T_4) = \int_0^{\theta} \frac{n \cdot x^{n-1}}{\theta^n} \cdot x dx = \int_0^{\theta} \frac{n \cdot x^n}{1} dx = n \int_0^{\theta} x^n dx = n \left[ \frac{x^{n+1}}{n+1} \right]_0^{\theta} = \frac{n\theta^{n+1}}{n+2}$$

מהתוצאה שקיבלנו ניתן לראות בבירור כי האומד מוטה.

e. ראינו בסעיף c שככל שגודל המדגם גדל, המהירות בה פונקצית הנראות קרבה ל-0 כתלות ב- $\theta$  נעשה מהיר יותר. הסיבה לכך היא מאופן הגדרת פונקצית הנראות עצמה במקרה של אומד זה:

$$f_{\theta}(x) = \begin{cases} \frac{1}{\theta} & 0 \leq x \leq \theta \\ 0 & o.w. \end{cases} \Rightarrow L(\theta) = \begin{cases} 1 & \theta \geq \max(x_1, \dots, x_n) \\ \theta^n & \theta < \max(x_1, \dots, x_n) \end{cases}$$

כלומר, עבור ערכים הקטנים ממקסימום המדגם, הסיכוי הוא 0 ועבור ערכים הגדולים מ- $\theta$  הסיכוי קטן ממש ל-0 ככל ש- $\theta$  (או גודל המדגם) גדלים. כלומר, ערך האומד הוא המקסימום של הפונקציה ולכן זהו הערך הסביר ביותר.

5.

a.

$$E\left(\frac{\sum X_i}{n}\right) = \frac{1}{n} \sum E(X_i) = \frac{n}{n} E(X) = \lambda$$

נשים לב שהפרמטר  $\lambda$  מוגדר להיות כמות הפניות בדקה. לכן, הפרמטר לכמות הפניות בשנייה הוא  $\frac{\lambda}{60}$ . מכאן שבאופן ברור האומד מוטה. כיוון שהתוחלת והשונות בהתפלגות פואסונית שווים שניהם לפרמטר, ממוצע המדגם מוטה גם לשונות.

b. כן. סכום של מ"מ בלתי תלויים המתפלגים פואסונית, מתפלג פואסונית.

c. לא. כיוון שסכום התצפיות מתפלג פואסונית, הרי שממוצע המדגם יתפלג פואסונית כפול קבוע וזו אינה משפחה מוכרת.

d.

$$V\left(\frac{\sum X_i}{n}\right) = \frac{1}{n^2} \sum V(X_i) = \frac{n}{n^2} V(X) = \frac{\lambda}{n}$$

e. אוקיי ☺

.f

$$\lambda \approx 60 \cdot \frac{\sum_{i=1}^{11} X_i}{11} = 60 \cdot \frac{1+1+4+10+2+3+7+3+5+6+6}{11} = 60 \cdot 4.36 = 261.81$$

.i

.ii

$$E(X_i) \approx \frac{\lambda}{60} = \frac{261.81}{60} = 4.36$$

.iii כיוון שבהתפלגות פואסונית, התוחלת והשונות שווים, גם שונות מספר השאלות בשנייה תהיה **4.36**.

.iv

$$V\left(\frac{\sum X_i}{n}\right) = \frac{1}{n^2} \sum V(X_i) = \frac{n}{n^2} V(X) = \frac{261.81}{11} = 23.8009$$

.g חישבנו בסעיף הקודם כי  $\lambda \approx 261.81$ . נסמן  $X \sim \text{Pois}(261.81)$  ונחשב  $P(X > 200)$ .

$$\begin{aligned} P(X > 200) &= P\left(\frac{X - \lambda}{\sqrt{\lambda}} > \frac{200 - \lambda}{\sqrt{\lambda}}\right) = P\left(\frac{X - 261.81}{\sqrt{261.81}} > \frac{200 - 261.81}{\sqrt{261.81}}\right) = \\ &= 1 - P\left(\frac{X - 261.81}{\sqrt{261.81}} \leq -3.82\right) = P\left(\frac{X - 261.81}{\sqrt{261.81}} \leq 3.82\right) = \Phi(3.82) = 0.9999333 \end{aligned}$$

נחשב במדויק באמצעות R:

```
> ppois(200,261.81,FALSE)
[1] 0.9999597
```

.h (\*)

$$\begin{aligned} P(X \leq x) = 0.9 &= P\left(\frac{X - \lambda}{\sqrt{\lambda}} \leq \frac{x - \lambda}{\sqrt{\lambda}}\right) \Rightarrow \frac{x - \lambda}{\sqrt{\lambda}} = X_{0.9} \Rightarrow x = \sqrt{\lambda} X_{0.9} + \lambda \\ \Rightarrow x &= 1.282\sqrt{\lambda} + \lambda \end{aligned}$$

.6 (\*)

.a תחילה נחשב MSE ל- $T_2$ 

$$E(T_2) = E\left(\frac{2}{n} \sum_{i=1}^n T_i\right) = \frac{2}{n} E\left(\sum_{i=1}^n T_i\right) = \frac{2}{n} \sum_{i=1}^n E(T_i) = \frac{2}{n} \cdot n \cdot \frac{1}{2} = 1$$

$$\text{bias}(T_2) = E(T_2) - \theta = 1 - 1 = 0$$

$$V(T_2) = V\left(\frac{2}{n} \sum_{i=1}^n T_i\right) = \frac{4}{n^2} V\left(\sum_{i=1}^n T_i\right) = \frac{4}{n^2} \sum_{i=1}^n V(T_i) = \frac{4}{n^2} \cdot n \cdot \frac{\theta^2}{12} = \frac{\theta^2}{3n} = \frac{1}{3n}$$

$$MSE(T_2) = \text{Bias}^2(T_2) + \text{Var}(T_2) = \frac{1}{3n}$$

עתה, נחשב MSR ל- $T_4$ :

$$E(T_4) = \frac{n\theta}{n+1} = \frac{n}{n+1} \quad (\text{ראינו בכיתה})$$

$$\text{bias}(T_4) = E(T_4) - \theta = \frac{n}{n+1} - 1 = \frac{n - n - 1}{n+1} = -\frac{1}{n+1}$$

נחשב את השונות:

$$F_{T_4}(x) = \begin{cases} \frac{x^n}{\theta^n} & 0 \leq x \leq \theta \\ 0 & o.w. \end{cases} \Rightarrow f_{T_4}(x) = \begin{cases} \frac{n \cdot x^{n-1}}{\theta^n} & 0 \leq x \leq \theta \\ 0 & o.w. \end{cases}$$

$$E(T_4^2) = \int_0^1 \frac{n \cdot x^{n-1}}{\theta^n} \cdot x^2 dx = \int_0^1 \frac{n \cdot x^{n+1}}{1} dx = n \int_0^1 x^{n+1} dx$$

$$= n \int_0^1 x^{n+1} dx = n \left[ \frac{x^{n+2}}{n+2} \right]_0^1 = \frac{n}{n+2}$$

$$V(T_4) = \frac{n}{n+2} - \left( \frac{n}{n+1} \right)^2 = \frac{n(n+1)^2 - n^2(n+2)}{(n+2)(n+1)^2} = \frac{n}{(n+2)(n+1)^2}$$

מכאן, שעריך MSE של האומד הוא:

$$MSE(T_4) = \frac{1}{(n+1)^2} + \frac{n}{(n+2)(n+1)^2} = \frac{2}{(n+2)(n+1)}$$

נשווה בין ערכי ה-MSE:

$$\frac{MSE(T_2)}{MSE(T_4)} = \frac{(n+2)(n+1)}{2} \cdot \frac{1}{3n} = \frac{n^2 + 3n + 2}{6n} = \frac{n + 3 + \frac{2}{n}}{6} \xrightarrow{n \rightarrow \infty} \infty$$

לכן  $T_4$  שואף ל-0 מהר יותר ומכאן שהוא האומד העדיף.

b.

i-viii

```
midgAv=vecotr(length=10)
midgMAX=vecotr(length=10)
for (i in 1:10)
{
    midg=runif(10,0,1)
    midgAv[i]=2*mean(midg)
    midgMAX[i]=max(midg)
}
SEav=(midgAv-1)^2
SEmax=(midgMAX-1)^2

MSEav=mean(SEav)
MSEmax=mean(SEmax)
```

viii. הסטייה הריבועית הממוצעת של  $T_2$  היא:

```
> MSEav
[1] 0.03700171
```

הסטייה הריבועית הממוצעת של  $T_4$  היא:

```
> MSEmax
[1] 0.0199235
```



ix. אם נציב  $n=10$  בערכי ה-MSE שמצאנו בסעיף a, נראה כי האומדים אכן דומים. האומד העדיף הוא  $T_4$  כיוון שערכי MSE שלו קטנים מערכי MSE של  $T_2$ . כיוון שערכי ה-MSE קרובים בסימולציה ובפיתוח האנליטי, אכן הבחירה הייתה זהה.

x.

```
midgAv=vecotr(length=100)
midgMAX=vecotr(length=100)
for (i in 1:100)
{
  midg=runif(10,0,1)
  midgAv[i]=2*mean(midg)
  midgMAX[i]=max(midg)
}
SEav=(midgAv-1)^2
SEmax=(midgMAX-1)^2

MSEav=mean(SEav)
MSEmax=mean(SEmax)

> MSEav
[1] 0.03254116
> MSEmax
[1] 0.01334547
```

קיבלנו שהאומדים קרובים יותר לערך התיאורטי וזאת מתוקף תכונת העקיבות.

.7

a. ביחס למדינת הוותיקן, נתונים אלו הם **אוכלוסייה** כיוון שמתארים את הגבהים של כל חברי הקבוצה.

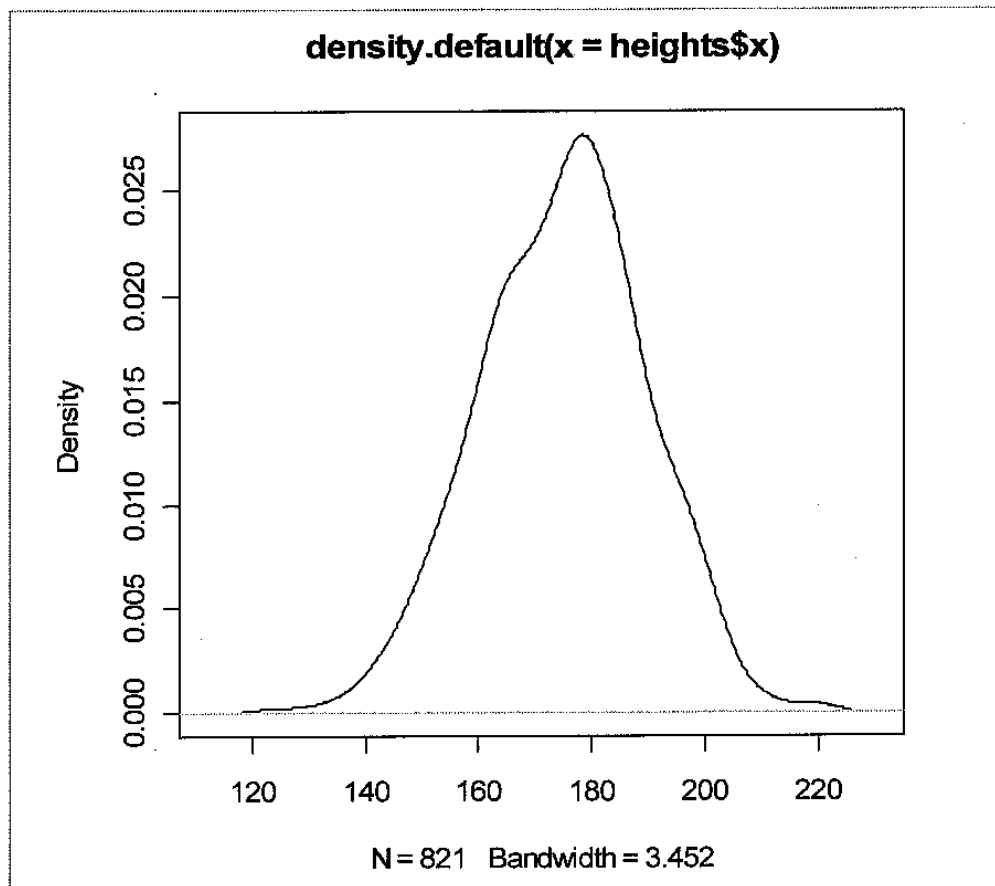
b. ביחס לכלל העולם, נתונים אלו הם **מדגם** כיוון שבשנת 2007 מספר בני אדם שחיו בעולם היה הגדול בהרבה מ-821. נשים לב, שכיוון שמספר בני האדם הוא כל כך גדול ביחס למדגם שנלקח, שלעובדה שבמדגם זה אין בחירה עם החזרה אין משמעות ריאלית.

c. נחשב תוחלת ושונות באמצעות R:

```
> heights=read.csv('d:/stat/heights.txt',sep=" ", header=T)
> mean(heights$x)
[1] 174.9610
> mH=mean(heights$x)
> sum((heights$x-mH)^2)/length(heights$x)
[1] 215.2271
```

d. נשרטט תרשים צפיפות:

```
plot(density(heights$x))
```



מתרשים הצפיפות ניתן להעריך כי הנתונים מתפלגים נורמלית.

e.

```
> choose(821,50)
[1] 3.738084e+80
```

f. המדגמים מהסעיף הקודם **אינם** מקיימים את ההנחות של מדגם מקרי כיוון שזו בחירה ללא החזרה (ולכן יש תלות בין התצפיות).

g. נשתמש בנוסחה לחישוב מספר הצירופים האפשריים עם החזרה ונקבל שניתן לבחור  $\binom{n+k-1}{k} = \binom{870}{50} = 7.403584e+81$  מדגמים שונים של אנשים שאינם בהכרח שונים במדינת הוותיקן.

h.

$$V(X) = V\left(\frac{\sum X_i}{50}\right) = \frac{1}{2500} \cdot V(\sum X_i) = \frac{1}{2500} \cdot \sum V(X_i) = \frac{215.4896}{50} = 4.309792$$

i.

```
> samH=sample(heights$x,replace=T,size=50)
> mean(samH)
[1] 174.8871
```

ממוצע זה הוא **אומד** לתוחלת הגבהים של אוכלוסיית הוותיקן. האומד שקיבלנו קרוב לתוחלת שקיבלנו בסעיף c כיוון שממוצע הוא אומד **חסר הטיה** אך לא מדויק.

```

> hgtData = vector(length = 1000)
> for (i in 1:1000)
+ {
+ hgtData[i]=mean(sample(heights$x,replace=T,size=50))
+ }
> var(hgtData)
[1] 4.12715

```

k. כפי שניתן לראות, השונות שקיבלנו קרובה מאוד לשונות מסעיף (h).

.8

a.  $X \sim B(1, p)$

סכום התצפיות על פני איברים המתפלגים בינומית, הוא מספר ההצלחות שקיבלנו במדגם. כלומר, כל איבר במדגם הוא כמו ניסוי ברנולי. לכן, סכום ההצלחות מתפלג  $B(n, p)$ . כמו כן, ע"פ משפט הגבול המרכזי, סכום התצפיות עבור  $n$  גדולים יתפלג נורמאלי בקירוב.

מכאן שהממוצע מתפלג בינומית כפול קבוע וזו אינה משפחה מוכרת. נשים לב שגם כאן, ע"פ משפט הגבול המרכזי, ניתן לקרב באמצעות התפלגות

$$\bar{x}_n \sim N\left(p, \frac{p(1-p)}{n}\right)$$

b.  $X \sim P(\lambda)$

סכום של משתנים מקריים ב"ת המתפלגים פואסונית הוא מ"מ המתפלג פואסונית ולכן גם הסכום יתפלג פואסונית עם הפרמטר  $\lambda n$ .

מכאן שהממוצע יתפלג פואסונית כפול קבוע וזו אינה משפחה מוכרת. נשים לב שגם כאן, ע"פ משפט הגבול המרכזי, ניתן לקרב את הממוצע באמצעות

$$\bar{x}_n \sim N\left(\lambda, \frac{\lambda}{n}\right)$$

c.  $X \sim G(p)$

סכום של משתנים מקריים ב"ת המתפלגים גיאומטרית הוא מ"מ המתפלג בינומית שלילית עם פרמטרים  $n$  ו- $p$ .

מכאן שהממוצע יתפלג בינומית שלילית כפול קבוע וזו אינה משפחה מוכרת. נשים לב שגם כאן, ע"פ משפט הגבול המרכזי, ניתן לקרב את הממוצע

$$\bar{x}_n \sim N\left(\frac{1}{p}, \frac{(1-p)}{np^2}\right)$$

d.  $X \sim N(u, q^2)$

סכום של משתנים מקריים ב"ת המתפלגים נורמאליה הוא מ"מ המתפלג נורמאליה. לכן, את הסכום יתפלג נורמאליה  $Y \sim N(nu, nq^2)$ .

כיוון שהתפלגות נורמאליה סגורה על טרנספורמציות ליניאריות, גם הממוצע

$$\bar{x}_n \sim N\left(u, \frac{q^2}{n}\right)$$



## תרגיל 5- דגימה ואמידה

מעבר לעזרה שבתוכנה עצמה, מומלץ להיעזר באתר: <http://wiki.r-project.org/rwiki/doku.php>  
 הדרכה בעברית ניתן למצוא באתר הקורס.  
 ניתן ורצוי להתשמש בפורום הקורס להתייעצות.  
 יש לצרף את הקוד ששימש לפתרון אך אין הוא תחליף לתשובה סופית.  
 סעיפים בדרגת קושי גבוהה יותר סומנו בכוכבית (\*). כדאי להשתדל לענות עליהם אבל אין הכרח.

1. הניחו שאלו הם מספר הצימוקים בארבע עוגות במאפיית "פחמימות ובניו":

עוגה	1	2	3	4
צימוקים	54	60	44	56

(a) האם ארבעת עוגות אלו הן **אוכלוסייה** או **מדגם** או **מדגם מקרי** בתרחישים הבאים:

i. אלו העוגות היחידות שיצרו אי פעם במאפייה.

ii. אלו עוגות שבחרתם באקראי בביקור האחרון שלכם במאפייה כמייצגות את מספר הצימוקים בעוגות אותו היום.

iii. אלו עוגות שבחרתם באקראי בביקור האחרון שלכם במאפייה כמייצגות את מספר הצימוקים בכלל העוגות במאפייה.

iv. אלו עוגות אקראיות שנאספו במועדים אקראיים מאז שהמאפייה נפתחה כמייצגות את מספר הצימוקים בכלל העוגות במאפייה.

v. אלו העוגות היחידות בטעם כבש שנבחרו לייצג את מספר הצימוקים בכלל העוגות במאפייה.

vi. אלו העוגות היחידות בטעם כבש שנבחרו לייצג את מספר הצימוקים בכלל העוגות בטעם כבש במאפייה.

בהמשך השאלה, הניחו כי אלו העוגות היחידות שיצרו אי פעם במאפייה.

(b) מהי תוחלת מספר הצימוקים בעוגה? מהי שונות מספר הצימוקים?

(c) ידיעת מספר הצימוקים בכל עוגה היא מצב לא סביר במציאות. ספירת הצימוקים היא תהליך ארוך שמשחית את העוגות ולכן יותר סביר שבשביל לדעת את תוחלת מספר הצימוקים נצטרך **לאמוד** אותו...

i. בכמה דרכים שונות אפשר לבחור שתי עוגות **שונות**? האם דגימה שכזו מקיימת

1 זאת הפעם האחרונה בה נתעכב על ההבדל בין "מדגם" ל"מדגם מקרי". אזכיר שממדגם **מקרי** אנו דורשים שכל אחת ואחת מהדגימות מתפלגת כמו האוכלוסייה אותה היא נועדה לייצג. דרישה זו מכתובה כמובן אי תלות (אחרת הדגימה השנייה כבר לא תתפלג כמו הראשונה). בעתיד, המילה "מדגם" תתייחס רק **למדגם מקרי** אלא אם נאמר אחרת.

את הגדרות הדגימה המקרית? (בפרט את השערת אי-התלות בין התצפיות).

ii. בכמה דרכים שונות אפשר לבחור שתי עוגות שאינן בהכרח שונות (כלומר מותר לבחור את אותה העוגה פעמיים)? האם דגימה שכזו מקיימת את הגדרות הדגימה המקרית?

iii. ענו על השאלות הבאות אם החלטנו לאמוד את מספר הצימוקים הטיפוסי (התוחלת) על ידי ממוצע של שתי עוגות שנדגמו מקרית:

A. אילו ערכים אפשריים יש לאומד? מה הסיכוי לכל ערך?

B. על בסיס כל התוצאות האפשריות והסיכויים שלהן מהסעיף הקודם, חשבו את תוחלת האומד? האם הוא חסר הטייה?

C. מהי השונות של האומד?

D. מהו ה MSE של האומד?

E. (\*) הציעו מדד אלטרנטיבי (שאינו MSE) לאיכות של אומד וחשבו אותו על הממוצע כאומד לתוחלת באותם הנתונים.

i. (\*) אולי בכלל עדיף להשתמש בממוצע גיאומטרי כאומד ולא בממוצע חשבוני. חשבו את ה MSE של הממוצע הגיאומטרי המבוסס על שתי תצפיות והשוו אותו לממוצע החשבוני. איזה אומד עדיף?

תזכורת- ממוצע גיאומטרי:  $geomean(\vec{x}) = (x_1 \cdot \dots \cdot x_p)^{1/p}$

2. עבור  $X_i \sim N(\mu, \sigma); i.i.d, i \in 1, \dots, n$  השלם את הטבלה:

עקיב? (כלומר כיצד יתנהג במדגם אינסופי?)	שונות	תוחלת	האם משתנה מקרי? כיצד מתפלג?
			$\mu$
			$\bar{X}$
			$\sigma^2$
			$\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$
			$\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$
			$\sqrt{\frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2}$

3. סדרה של n נסיונות מתוארת על ידי n משתנים:  $X_i \sim B(1, p=1/2); i.i.d, i \in 1, \dots, n$  נניח שאיננו יודעים ש  $p=1/2$ , ונניח שבעשרה נסיונות התקבלו התוצאות הבאות: 1,0,0,1,1,0,1,0,0,0

$$16.28 + \frac{44.89 \times 2}{89.78} + \frac{2.1904 \times 2}{4.38} + \frac{22.75 \times 2}{45.5} + 95.64 + \frac{34 \times 2}{2.69} + \frac{59.9 \times 2}{119.8} + 20.25$$

$$0.25 + \frac{1}{2} \times 2 + 20.25 \times 2 + \frac{1}{2} \times 2 + 42.25 + 2.25 \times 2 + 20.25 \times 2 + 90.25 + \dots$$

(a) האם האומדים הבאים הם חסרי הטייה עבור  $p=1/2$ ? ועבור  $p$  כללי?  
 האם הם עקיבים?  
 מהי שונותם?

i. פרופורציית ההצלחות.

ii. הממוצע.

$$iii. \frac{\sum_{i=1}^n X_i + 1}{n+2}$$

iv. תוצאת הניסוי הראשון.

(b) מהו האומדן ל  $p$  בשיטות האמידה השונות במדגם שהתקבל? האם האומדן קרוב לפרופורציה האמיתית?

4. אורך החיים של סוללות מסוג "דוֹרְזוֹל" הוא אי שם בין אפס ל  $\theta$  (בסיכוי שווה כלומר מתפלג אחיד). אבנר מעוניין לאמוד את  $\theta$  בהסתמך על מדגם של סוללות ולהן אורכי החיים  $x_1, \dots, x_n$ .

(a) האם האומדים הבאים הם חסרי הטייה? האם הם עקיבים (ללא חישוב)?  
 מהי שונותם?

i. פעמיים התצפית הראשונה:  $T_1 = 2 \cdot X_1$

ii. פעמיים ממוצע המדגם:  $T_2 = \frac{2}{n} \sum_{i=1}^n X_i$

iii. ריבוע התצפית האחרונה:  $T_3 = X_n^2$

הערה: אם אתם מתקשים לחשב את השונות של אומדן זה, אל תתעכבו על האינטרגל יותר מידי. זו אינה מטרת התרגיל.

(b) איזה מהאומדים בסעיף הקודם עדיף? נמק.

(c) נפשט את הבעייה המקורית ונניח שבידי אבנר מדגם בגודל 1 דהיינו, אורך החיים של סוללה בודדת שנבחרה באקראי  $(x_1)$  על בסיסה הוא ינסה לאמוד את  $\theta$ .

i. האם ייתכן ש  $\theta$ , שכאמור אינו ידוע לנו, הוא למעשה קטן מאורך החיים של הסוללה שבדיז?

ii. ואם בידיו שתי סוללות. האם ייתכן ש  $\theta$  קטן מאורך החיים של סוללה כלשהי מהשתיים?

iii. צייר בתרשים את הצפיפות המשותפת של מדגם בגודל 1 מהתפלגות  $Unif \sim [0, \theta]$  כפונקציה של הפרמטר הלא ידוע  $\theta$ .

iv. צייר בתרשים את הצפיפות המשותפת של מדגם בגודל 2  $Unif \sim [0, \theta]$  כפונקציה של הפרמטר הלא ידוע  $\theta$ .

(d) אחרי שקרא את הסעיף הקודם, אבנר החליט להשתמש באורך החיים **המקסימלי** כאומד  $T_4 = \max\{x_1, \dots, x_n\}$ . האם הוא עקיב (ללא חישוב)?  
 (\*) האם הוא חסר הטייה?

הצעה לאופן חישוב התוחלת של האומד לשם בדיקת חוסר הטייה:  
 מצאו את ההתפלגות **המצטברת** של המקסימום של מדגם מהתפלגות  $Unif \sim [0, \theta]$ . בשביל שהמקסימום יהיה קטן מערך כלשהו, כל התצפיות צריכות להיות קטנות מאותו ערך. גזרו ומצאו או הצפיפות של המקסימום של מדגם מהתפלגות זו. עכשיו אפשר לחשב את התוחלת של המקסימום.

(e) הראה ש  $T_4$  הוא למעשה הערך **הכי סביר** של  $\theta$  לאור הנחת ההתפלגות האחידה של אורך חיי סוללה והמדגם שקיבל אבנר (כלומר שממקסם את הצפיפות המשותפת).  
 כדאי להעמיק בשני הסעיפים הקודמים לצורך הפתרון.  
 בהמשך הקורס תראו ש  $T_4$  נקרא למעשה "אומד נראות מקסימלית" (Maximum likelihood).

5. נניח שמספר השאלות בדקה בשרתי האוניברסיטה מתפלג פואסונית עם קצב  $\lambda$ . במציאות איננו יודעים מהו הפרמטר האמיתי של ההתפלגות הפואסונית ולכן לא נותר לנו אלא לאמוד אותו. לשם כך נאסוף נתונים על מספר השאלות בדקה, בדקות שנבחר באקראי.

(a) האם ממוצע המדגם הוא אומד חסר הטייה לתוחלת של מספר השאלות בשנייה?  
 האם הוא אומד חסר הטייה לשונות?

(b) האם **סכום** התצפיות שייך למשפחת התפלגויות שלמדתם?

(c) האם **ממוצע** המדגם שייך למשפחת התפלגויות שלמדתם?

(d) מהי שונות ממוצע המדגם (כפונקציה של הפרמטר הלא ידוע של ההתפלגות הפואסונית)?

(e) הסעיפים הבאים מתבססים על מדגם ספציפי. כל מדידה מייצגת את מספר השאלות בשנייה כלשהי שנבחרה באקראי על פני היממה.  
 1,1,4,10,2,3,7,3,5,6,6

(f) השתמש בממוצע המדגם בשביל לאמוד את:

i. הפרמטר של ההתפלגות הפואסונית ( $\lambda$ ).

ii. מספר השאלות בשנייה טיפוסית:  $E(X_i)$ .

iii. השונות של מספר השאלות בשנייה:  $Var(X_i)$ .

iv. השונות של האומד שבו השתמשנו:  $Var(\hat{\lambda})$  (ראה סעיף d לעיל).

(g) **תן אומדן** לסיכוי שבדקה כלשהי יגיעו יותר מ-200 שאלות. חשבו את הסיכוי על ידי הקירוב הנורמלי והשוו לחישוב מדויק ב  $R$  ( $ppois$ ).  
 (בשלב זה לא נתעכב על השאלה האם האומד לסיכוי זה הוא חסר הטייה, עקיב וכו' אם כי אפשר ורצוי לשאול את השאלות הללו).



(h) (\*) הציעו אומדן למספר השאלות המירבי ב-90% מהדקות. כלומר הציעו פונקציה של מדגם/וקטור מקרי של  $n$  משתנים פואסוניים בלתי תלויים, שמחזירה את האחוזן התשעים של התפלגות מספר השאלות **בדקה**.  
 המז: השתמשו במשפט הגדול המרכזי.

6. (\*) מטרת שאלה זו היא להדגים את מושג ה-MSE באמצעות סימולציה. ונשווה את הסטייה הריבועית הטיפוסית (MSE) של האומדים לגבול העליון של ההתפלגות האחידה  $Unif[0, \theta]$  :

שני האומדים לבדיקה הם פעמיים ממוצע המדגם  $\left(T_2 = \frac{2}{n} \sum_{i=1}^n T_i\right)$  אל מול המקסימום של המדגם  $\left(T_4 = \max\{T_i\}\right)$ .

(a) חשב את הסטייה הריבועית הטיפוסית עבור התפלגות אחידה בקטע  $[0,1]$  כלומר  $\theta=1$ . איזה מהאומדים למקסימום של ההתפלגות עדיף לפי קריטריון זה?  
**הערה:** זו כמובן שאלה לא מציאותית כי אם היינו יודעים שהמקסימום הוא אחד, כרגיל לא היינו צריכים לאמוד אותו.

(b) כעת בצעו סימולציה שתאמוד את הסטייה הריבועית הטיפוסית על ידי מיצוע על פני מדגמים. המתכון הוא הבא:

i. הגרילו עשר תצפיות מהתפלגות אחידה בקטע  $[0,1]$ .

ii. חזרו על הסעיף הקודם עשר פעמים כך שיהיו לכם עשרה מדגמים בגודל עשר.

iii. חשבו את הממוצע של כל מדגם כך שיהיו בידיכם עשרה ממוצעים מעשר תצפיות כל אחד. אלו כמובן אומדנים לתוחלת של ההתפלגות וכנראה שתקבלו ערכים מסביב לחצי (התוחלת האמיתית).

iv. הכפילו את הממוצע של כל מדגם פי שניים בשביל לקבל אומדן למקסימום  $(T_2)$ .

v. חשבו את המקסימום של כל עשר תצפיות בשביל לקבל אומדן נוסף למקסימום  $(T_4)$ .

vi. חשבו את הסטייה הריבועית של כל אומדן (פעמיים הממוצע או המקסימום של המדגם) שקיבלתם מהערך  $I$  שהוא כאמור המקסימום האמיתי.

vii. אימדו את הסטייה הריבועית הממוצעת על פני עשרת המדגמים על ידי מיצוע של עשר הסטיות הריבועית שחישבתם בסעיף הקודם.

viii. מהי הסטייה הריבועית הממוצעת שחישבתם עבור  $T_1$  ועבור  $T_4$  ?

ix. כעת שיש בידיכם שני ערכים ל-MSE, האם הם דומים ל-MSE התיאורטי שחישבתם בסעיף (a)? איזה מהאומדים עדיף? האם הסימולציה והפיתוח האנליטי הביאו אתכם לבחירת אותו האומדן?

x. הגדילו את מספר המדגמים ל-100 ואימדו מחדש את ה-MSE של כל אחד מהאומדים. האם הם יותר קרובים לערך התיאורטי שחישבתם בסעיף (1)? מתוקף איזו תכונה האומדנים קרובים יותר לגודל האמיתי- חוסר הטייה או עקיבות?

נכון לשנת 2007 היו בוטיקן 821 תושבים. הגבהים שלהם נמצאים בקובץ heights שבאתר הקורס.

- (a) האם נתונים אלו הם "אוכלוסייה", או "מדגם (מקרי)" ביחס למדינת הוטיקן?
- (b) האם נתונים אלו הם "אוכלוסייה" או "מדגם (מקרי)" ביחס לכלל העולם?
- (c) מהי תוחלת הגבהים בוטיקן? מהי שונותם?
- (d) השתמשו בתרשים צפיפות בשביל להעריך האם התפלגותם דומה לנורמלית?
- (e) כמה מדגמים שונים של 50 אנשים **שונים** אפשר לבחור במדינת הוטיקן? הערה: אפשר להיעזר בפקודה choose ב R לצורך חישובים קומבינטוריים.
- (f) האם המדגמים מסעיף e מקיימים את ההנחות של **מדגם מקרי**?
- (g) כמה מדגמים שונים של אנשים (לא בהכרח שונים) אפשר לבחור במדינת הוטיקן?
- (h) מהי השונות התיאורתית של ממוצע המבוסס על 50 תצפיות בלתי תלויות מאוכלוסייה שאת שונותה חיבתם בסעיף (c)?
- (i) הישתמשו בפונקציה sample בשביל "לדגום **מקרי**" 50 מתושבי הוטיקן ולחשב את הגובה הממוצע שלהם. מהו ממוצע זה? האם הוא קרוב לתוחלת הגבהים במדינה מסעיף (c)?  
הערה: שימו לב למתג replace של הפקודה !sample
- (j) חזרו על הסעיף הקודם 1000 פעמים ושימרו את התוצאות. אלו למעשה 1000 מדגמים מקריים בגודל 50 מאוכלוסיית הוטיקן. חשבו לכל מדגם בגודל 50 את הממוצע ואז חשבו את השונות של אלף הממוצעים.
- (k) האם השונות שחיבתם על בסיס המדגמים דומה לשונות התיאורתית מסעיף (h)?

8. התפלגות הממוצע:

סעיף זה נועד לבדוק את ההתפלגות של הממוצע על פני מדגמים כאשר דגמתם (באופן מקרי) באוכלוסיות שונות. עבור כל אחת ממשפחת ההתפלגויות שבהמשך ענו על השאלות הבאות:

כיצד יתפלג **סכום** התצפיות על פני מדגמים? האם זו התפלגות מוכרת? האם ניתן לקרב אותה על ידי התפלגות מוכרת?

כיצד יתפלג **ממוצע** התצפיות על פני מדגמים? האם זו התפלגות מוכרת? האם ניתן לקרב אותה על ידי התפלגות מוכרת?

(a) כאשר האוכלוסייה ניתנת לתיאור על ידי ההתפלגות הבינומית:  $X_i \sim B(1, p)$  (לדוגמה, קנה/לא קנה יוגורט השבוע בסופר).

(b) כאשר האוכלוסייה ניתנת לתיאור על ידי התפלגות פואסונית:  $X_i \sim \text{poiss}(\lambda)$  (לדוגמה, מספר גביעי היוגורט שנקנים בשבוע בסופר).

(c) כאשר האוכלוסייה ניתנת לתיאור על ידי התפלגות גיאומטרית:  $X_i \sim G(p)$   
(לדוגמה, מספר האנשים שנכנסים לסופר עד שמישהו מהם קונה יוגורט).

(d) כאשר האוכלוסייה ניתנת לתיאור על ידי ההתפלגות הנורמלית  $X_i \sim N(\mu, \sigma)$   
(לדוגמה, המשקל של גביע יוגורט).

