בס"ד

מבוא להסתברות וסטטיסטיקה 88-165

תרגילי בית 10 - אומדנים, רווחי סמך - פתרונות

1. למשתנה מקרי X יש פונקציית הצטברות

$$F(x) = x^a \ , \qquad 0 < x < 1$$

כאשר $a > -1$ הוא קבוע, אבל לא ידוע.

(א) הוכח ש- $\mathbf{E}[X] = \frac{a}{1+a}$.

(ב) מצא את אומדן הנראות המירבית לפרמטר a. האם אומדן זה הוא חסר הטייה ?

(ג) לאור התוצאה של סעיף (א) מציעים $\frac{\overline{X}}{1-\overline{X}}$ כאומדן ל- a. האם אומדן זה הוא חסר הטייה ?

(ד) במקרים $a = 2, 4$, העזר במחשב לייצור מדגמים מגודל 10 מההתפלגות הרלוונטית, ובדוק (על ידי הסטוגרמה) את ההתפלגות של שני האומדנים ל-a. איזה מהם עדיף ?
הערה: ניתן לייצור מדגמים של ההתפלגות על ידי הפקודה rand(10,1).^ (1/a)

(א) הצפיפות היא $f(x) = F'(x) = ax^{a-1}$. ולכן

$$\mathbf{E}[X] = \int_{-\infty}^{\infty} xf(x)dx = \int_0^1 ax^a \ dx = \frac{a}{a+1}$$

(ב) אומדן הנראות המירבית הוא הבחירה של a כך ש

$$\sum_{i=1}^n \log(f(X_i)) = \sum_{i=1}^n (\log a + (a-1)\log X_i)$$

הוא מקסימלי. על ידי נגזרת ביחס ל- a מקבלים

$$\frac{n}{a} + \sum_{i=1}^n \log X_i = 0$$

ולכן אומדן הנראות המירבית הוא

$$a = -\frac{1}{\frac{1}{n}\sum_{i=1}^n \log X_i} \ .$$

לא נראה לי שזה חסר הטייה. כאשר $n = 1$ יש לנו

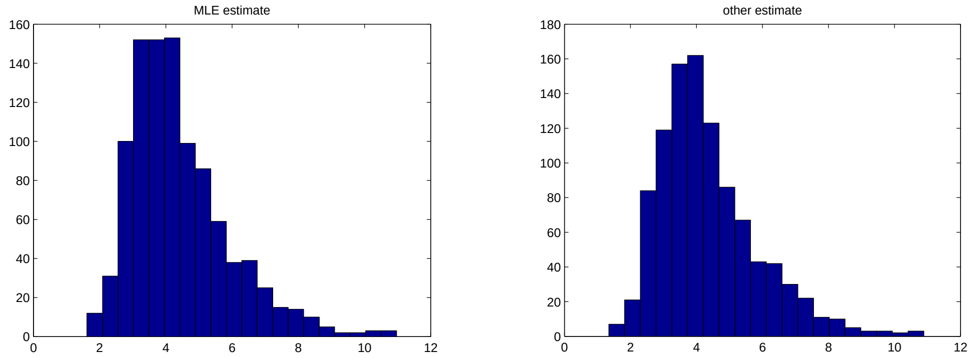
$$\mathbf{E}\left[-\frac{1}{\log X_1}\right] = -\int_0^1 \frac{ax^{a-1}}{\log x}dx = \int_0^\infty \frac{e^{ay}}{y} \ dy$$

(על ידי הצבה $x = e^y$) ואנטגרל זה אינו מתכנס.

(ג) כאשר $n = 1$ יש לנו

$$\mathbf{E}\left[\frac{\overline{X}}{1-\overline{X}}\right] = \int_0^1 \frac{ax^a}{1-x} \ dx$$

שגם זה לא מתכנס !



,

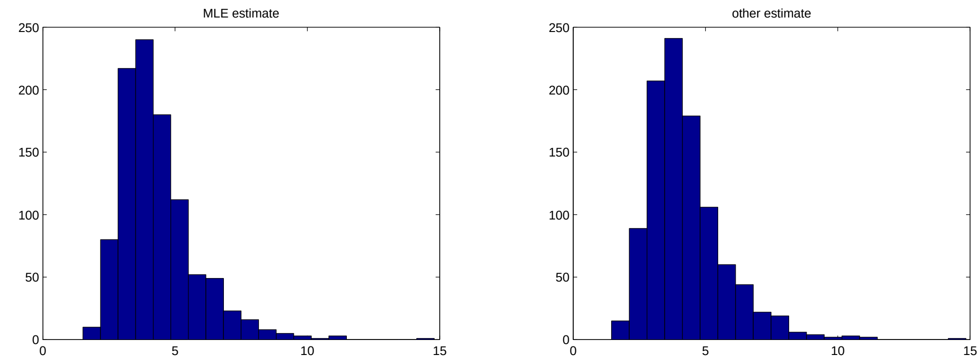
איור 1: המקרה $a = 4$. התפלגות ה- MLE ל- a (שמאל) והתפלגות האומדן השני ל- a (ימין)

(ד) הנה הפקודות לעשות את הסימולציה של שני האומדנים למקרה $a = 4$: ולצייר את ההסתוגרמות

```
N=1000;                          % number of samples
n=10;                            % sample size
a=2;                             % true value of a
es1=zeros(N,1);                  % first estimate values (MLE)
es2=zeros(N,1);                  % second estimate values
for i=1:N
   X=rand(n,1).^(1/a);           % make the sample
   es1(i) = -n/sum(log(X));      % compute first estimate (MLE)
   es2(i) = mean(X)/(1-mean(X)); % compute second estimate
end
figure(1)
hist(es1,20)                     % plot histogram of first estimate
title('MLE estimate')
figure(2)
hist(es2,20)                     % plot histogram of first estimate
title('other estimate')
mean(es1)                        % find average values of the 2 estimates
mean(es2)
```

לשני האומדנים התוחלת היא כנראה גבוה מדי - לאמדן הראשון בסימולציה זו התוחלת היא 4.41 לאומדן השני 4.35. באופן כללי התוחלת של ה- MLE יוצא יותר גבוה מהתוחלת של האומדן השני.

למקרה $a = 2$ יש תוצאות דומות: הסטוגרמות בעמוד הבא, הממוצעים: 2.34 (MLE), 2.18 (אומדן שני).



,

איור 2: המקרה $a = 2$. התפלגות ה- MLE ל- a (שמאל) והתפלגות האומדן השני ל- a (ימין)

---

2. $X_1, \ldots, X_n$ הוא מדגם מהתפלגות גאומטרית עם פרמטר p. מצא את אומדן הנראות המרבית ל- p. האם הוא חסר הטייה ?

ההתפלגות גאומטרית

$$\mathbf{P}(X = r) = p(1-p)^{r-1} \ , \qquad r = 1, 2, \ldots$$

לכן ה- MLE נבחר לעשות את

$$\sum_i \log \mathbf{P}(X = X_i) = n \log p + \left(\sum X_i - 1\right) \log(1-p)$$

מקסימלי: נגזור ביחס ל- p לקבל

$$\begin{aligned} \frac{n}{p} - \frac{(\sum X_i - 1)}{1-p} &= 0 \\ \Leftrightarrow \quad (1-p) - (\overline{X} - 1)p &= 0 \\ \Leftrightarrow \quad p &= \frac{1}{\overline{X}} \end{aligned}$$

אין שום סיבה מיוחדת שזה יהיה חסר הטייה. במקרה $n = 1$ יש לנו

$$\mathbf{E}\left[\frac{1}{\overline{X}}\right] = \mathbf{E}\left[\frac{1}{X}\right] = \sum_{r=1}^{\infty} \frac{p(1-p)^{r-1}}{r} = -\frac{p}{1-p} \log(1-p)$$

(טור טיילור מוכר של $-\log(1-p)$.) וזה לא שווה ל- p, למרות שכמובן $\mathbf{E}[X] = \frac{1}{p}$.

---

3. בוחרים, ללא חזרות, k מספרים מהשלמים $1, 2, \ldots, n$, כאשר n אינו ידוע. הוכח ששני הסטטיסטיים

$$\frac{k+1}{k}\max_i(X_i) - 1 \;, \quad 2\overline{X} - 1$$

הם אומדנים חסרי הטייה ל- n. על ידי ניסוי במחשב (או אחרת) מצא לאיזה משני אומדנים אלה יש שונות נמוכה יותר.

נסמן על ידי $X_1, \ldots, X_k$ המספרים שנבחרו. יש תלות ביניהם; אבל בכל זאת לכל i יש לנו

$$\mathbf{E}\left[X_i\right] = \frac{n+1}{2}$$

ולכן על ידי הליניאריות של תוחלת (אפילו במצב של תלות) יש לנו

$$\mathbf{E}\left[2\overline{X} - 1\right] = \mathbf{E}\left[\frac{2}{k}\sum_{I=1}^{K} X_i - 1\right] = \frac{2}{k} \cdot k \cdot \frac{n+1}{2} - 1 = n$$

יש לנו גם ש-

$$\mathbf{P}(\max_i(X_i) \le r) = \mathbf{P}(X_1, \ldots, X_k \le r) = \frac{\begin{pmatrix} r \\ k \end{pmatrix}}{\begin{pmatrix} n \\ k \end{pmatrix}} \;, \qquad r = k, \ldots, n$$

ולכן

$$\mathbf{P}(\max_i(X_i) = r) = \frac{\begin{pmatrix} r \\ k \end{pmatrix} - \begin{pmatrix} r-1 \\ k \end{pmatrix}}{\begin{pmatrix} n \\ k \end{pmatrix}} \;, \qquad r = k, \ldots, n$$

(כאן מובן מאליו ש- $\begin{pmatrix} k-1 \\ k \end{pmatrix} = 0$). ולכן:

$$\begin{aligned}
\mathbf{E}[\max_i(X_i)] &= \sum_{r=k}^{n} r \frac{\begin{pmatrix} r \\ k \end{pmatrix} - \begin{pmatrix} r-1 \\ k \end{pmatrix}}{\begin{pmatrix} n \\ k \end{pmatrix}} \\
&= \sum_{r=k}^{n} \frac{r\begin{pmatrix} r \\ k \end{pmatrix} - (r-1)\begin{pmatrix} r-1 \\ k \end{pmatrix} - \begin{pmatrix} r-1 \\ k \end{pmatrix}}{\begin{pmatrix} n \\ k \end{pmatrix}} \\
&= \frac{1}{\begin{pmatrix} n \\ k \end{pmatrix}}\left(n\begin{pmatrix} n \\ k \end{pmatrix} - \sum_{r=k+1}^{n}\begin{pmatrix} r-1 \\ k \end{pmatrix}\right)
\end{aligned}$$

כאן השתמשנו בסכום של טור טלסקופי. הסכום שנשאר הוא חלק של עמוד של משולש פסקל וניתן להוכיח על ידי אנדוקציה שהוא שווה $\begin{pmatrix} n \\ k+1 \end{pmatrix}$. ולכן

$$\mathbf{E}[\max_i(X_i)] = n - \frac{\begin{pmatrix} n \\ k+1 \end{pmatrix}}{\begin{pmatrix} n \\ k \end{pmatrix}} = n - \frac{n-k}{k+1} = \frac{k(n+1)}{k+1}$$

וכמו כן

$$\mathbf{E}\left[\frac{k+1}{k}\max_i(X_i) - 1\right] = \frac{k+1}{k}\frac{k(n+1)}{k+1} - 1 = n$$

סיכום: $\frac{k+1}{k}\max_i(X_i) - 1$ ו- $2\overline{X} - 1$ הם אומדנים חסרי הטייה ל n.

לדוגמה: $n = 5, k = 2$. יש 10 דרכים לבחור:

$$12, 13, 14, 15, 23, 24, 25, 34, 35, 45$$

ההתפלגות של $2\overline{X} - 1$:

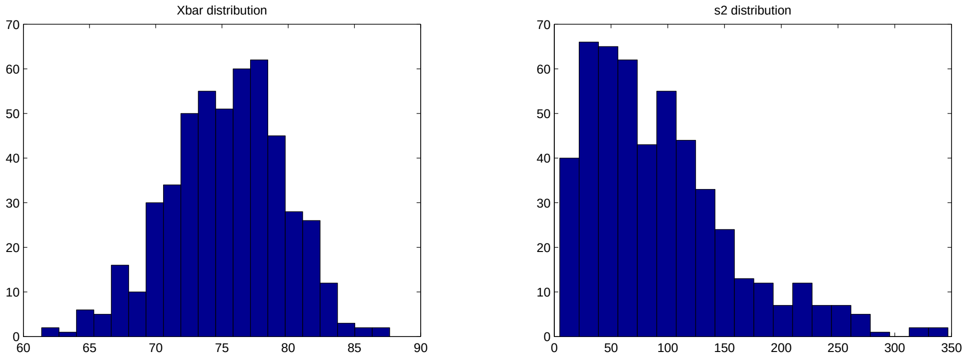
ערך	8	7	6	5	4	3	2	
הסת'	1	1	2	2	2	1	1	$\times\frac{1}{10}$

עם תוחלת 5 ושונות 3.

ההתפלגות של $\frac{3}{2}\max_i(X_i) - 1$:

ערך	$\frac{13}{2}$	5	$\frac{7}{2}$	2	
הסת'	4	3	2	1	$\times\frac{1}{10}$

עם תוחלת 5 ושונות 2.25.



,

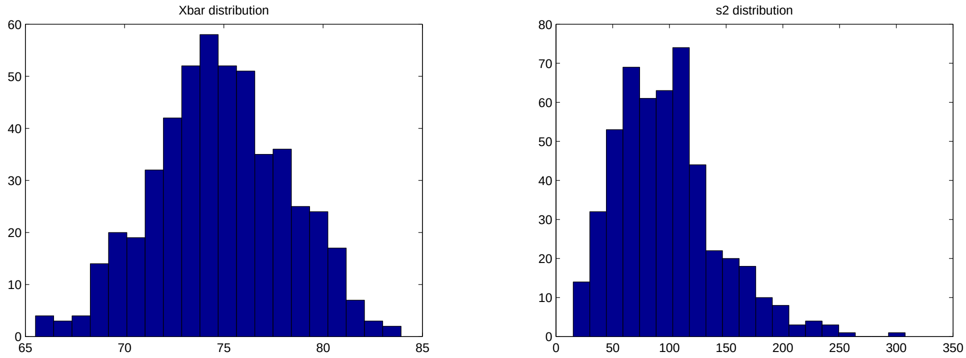
איור 3: מדגמים בגודל 5. התפלגות של $\overline{X}$ (שמאל) והתפלגות של s^2 (ימין)

4. במטלב בחר 500 מדגמים מגודל 5 מתוך ההתפלגות $N(75,10)$ ולכל מדגם חשב את הערכים של $\overline{X}$ ושל s^2. תאר את ההתפלגויות של $\overline{X}$ ושל s^2 (מומלץ הסטוגרמות). חזור על השאלה למדגמים מגודל 10. מה היא ההשפעה של גודל המדגם על ההתלפגות ?

קוד מטלב:

```
N=500;                          % number of samples
n=5;                            % sample size
mu=75;
sigma=10;
Xb=zeros(N,1);                  % values of Xbar
s2=zeros(N,1);                  % values of s2
for i=1:N
   X=mu+sigma*randn(n,1);       % make the sample
   Xb(i) = mean(X);             % make value of Xbar
   s2(i) = var(X);              % make value of s2
end
figure(1)
hist(Xb,20)                     % plot histogram of Xbar
title('Xbar distribution')
figure(2)
hist(s2,20)                     % plot histogram of s2
title('s2 distribution')
```

תוצאות $n=5$ בעמוד זה למעלה, תוצאות $n=10$ בעמוד הבא. עבור $n=10$ ההתפלגויות יותר מרוכזות - וכמובן כל שהמדגם יהיה יותר גדול ההתפלגויות יהיו עוד יותר מרוכזות. חשוב להבין למה!



,

איור 4: מדגמים בגודל 10. התפלגות של $\overline{X}$ (שמאל) והתפלגות של s^2 (ימין)

---

5. למדגם X_1, X_2, X_3 מגודל 3 מההתפלגות $N(\mu,\sigma)$ עם σ ידוע, מציעים 3 אומדנים ל- μ:

$$W_1 = \frac{1}{3}(X_1+X_2+X_3)$$
$$W_2 = \text{median}(X_1,X_2,X_3)$$
$$W_3 = \frac{1}{2}\left(\max_i(X_i)+\min_i(X_i)\right)$$

הוכח שכולם חסרי הטייה. על ידי ניסוי במחשב או אחרת, מצא את האומדן עם שונות נמוכה ביותר

חזור על השאלה כאשר עכשיו המדגם הוא מההתפלגות $U\left(a-\frac{1}{2},a+\frac{1}{2}\right)$ עם a לא ידוע, והאומדנים הם ל- a.

W_1, W_2, W_3 הם חסרי הטייה (כלומר בעלי תוחלת μ) בגלל שברור שלכולם יש התלפגויות סימטריות מסביב ל μ.

קוד מטלב:

```
N=10000;
n=3;
mu=75;
sigma=10;
W1=zeros(N,1);
W2=zeros(N,1);
W2=zeros(N,1);
for i=1:N
   X=mu+sigma*randn(n,1);
   W1(i) = mean(X);
   W2(i) = median(X);
   W3(i) = (max(X)+min(X))/2;
end
```

```
var(W1)
var(W2)
var(W3)
```

קיבלתי תשובות לשונויות 33.5 (W_1), 45.0 (W_2), 36.4 (W_3). במקרה זה הממוצע הוא ה-MVUE. התשובה המדוייקת לשונות של הממוצע הוא כמובן $33\frac{1}{3}$.

אותו דבר להתפלגות אחידה: קוד מטלב

```
N=10000;
n=3;
a=12;
W1=zeros(N,1);
W2=zeros(N,1);
W3=zeros(N,1);
for i=1:N
   X=a-0.5+rand(n,1);
   W1(i) = mean(X);
   W2(i) = median(X);
   W3(i) = (max(X)+min(X))/2;
end
var(W1)
var(W2)
var(W3)
```

קיבלתי תשובות לשונויות 0.0277 (W_1), 0.0498 (W_2), 0.0250 (W_3). במקרה זה כנראה ש-W_3 עדיף! (ניתן לחשב תשובות מדוייקות - הרבה עבודה - והם $\frac{1}{40}, \frac{1}{20}, \frac{1}{36}$.)

---

6. העזר במטלב לייצור 500 מדגמים מגודל 10 מההתפלגות הנורמלית עם $\mu = 15$, $\sigma = 5$. מכל מדגם מצא את הממוצע $\overline{X}$ ואת סטיית תקן של המדגם s.

על ידי בדיקת 500 המדגמים וודא ש-

$$\mathbf{P}\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < 1.645\right) \approx 0.90$$

ו-

$$\mathbf{P}\left(\frac{\overline{X}-\mu}{\frac{\sigma}{\sqrt{n}}} < 1.96\right) \approx 0.95$$

למה

$$\mathbf{P}\left(\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} < 1.645\right)$$

ו-

$$\mathbf{P}\left(\frac{\overline{X}-\mu}{\frac{s}{\sqrt{n}}} < 1.96\right)$$

הם יותר נמוכים ? הסבר איך ניתן לשנות את המספרים 1.645 ו- 1.96 כך שהסתברויות אלה יעלו שוב ל- 0.90 ו- 0.95.

קוד להתחיל את הסימולציה:

```
N=500;
n=10;
mu=15;
sigma=5;
Xbar=zeros(N,1);
s2=zeros(N,1);
for i=1:N
   X=mu+sigma*randn(n,1);
   Xbar(i) = mean(X);
   s2(i) = std(X);
end
```

כאשר כתבתי את הפקודות

```
sum( abs(Xbar-mu)/(sigma/sqrt(n)) < 1.645 )/N
sum( abs(Xbar-mu)/(sigma/sqrt(n)) < 1.96 )/N
```

קבלתי תשובות 0.896 ו- 0.948 - בערך 0.9 ו- 0.95.

אבל כאשר כתבתי

```
sum( abs(Xbar-mu)./(s/sqrt(n)) < 1.645 )/N
sum( abs(Xbar-mu)./(s/sqrt(n)) < 1.96 )/N
```

קבלתי רק 0.868 ו- 0.920.

כאשר תקנתי להשתמש באחוזונים של התפלגות t_9 :

```
sum( abs(Xbar-mu)./(s/sqrt(n)) < 1.83 )/N
sum( abs(Xbar-mu)./(s/sqrt(n)) < 2.26 )/N
```

אזי קבלתי 0.902 ו- 0.950.

הכלל: כאשר יש צורך להשמתש ב- s בגלל שלא יודעים σ מתקנים את רווח הסמך לתוחלת על ידי שימוש בהתפלגות t_{n-1} במקום $N(0,1)$.

---

7. לוקחים מדגם של 8 רכיבים המיוצרים במכונה האמורה לייצר עם אורך 1.500 ס"מ. מוצאים שהאורכים של הרכיבים הם

$$1.502, 1.501, 1.504, 1.498, 1.503, 1.499, 1.505, 1.504$$

(א) מצא אומדנים לתוחלת ולשונות של האורך של רכיב.

(ב) מצא רווח סמך סימטרי לתוחלת של האורך עם רמת מובהקות 90%

(ג) מצא חסם עליון לסטיית התקן של האורך עם רמת מובהקות 95%.

(א)

$$\overline{x} = \frac{1}{8}\sum_{i=1}^{8} x_i = 1.5020$$

$$s^2 = \frac{1}{7}\sum_{i=1}^{8}(x_i - \overline{x})^2 = 6.3 \times 10^{-6}$$

(ב) הסטטיסטי

$$\frac{\overline{x}-\mu}{\frac{s}{\sqrt{8}}}$$

מתפלג t_7. ולכן עם הסתברות 90%

$$\left|\frac{\overline{x}-\mu}{\frac{s}{\sqrt{8}}}\right| \leq 1.89$$

ולכן עם רמת מובהקות 90% $1.5003 \leq \mu \leq 1.5037$.

(ג) $\frac{7s^2}{\sigma^2}$ מתפלג $\chi^2(7)$ ולכן עם הסתברות 95%

$$\frac{7s^2}{\sigma^2} \geq 2.17$$

כלומר

$$\sigma \leq \sqrt{\frac{7}{2.17}}s \approx 4.5 \times 10^{-3}$$

---

8. במדגם של גברים נשואים מוצאים שהגיל בזמן החתונה מתפלג כדלהלן:

מספר גברים	גיל (אמצע קטע)
28	17.5
68	22.5
43	27.5
18	32.5
9	37.5
4	42.5
2	47.5
1	52.5
0	57.5
2	62.5

מצא אומדנים לתוחלת ולסטיית התקן של הגיל של גברים בזמן החתונה, ומצא רווח סמך עם רמת מובהקות 90% לתוחלת.

חישוב הסטטיסטיים:

$$\overline{x} = \frac{1}{175}\sum x_i = 26.1$$

$$s^2 = \frac{1}{174}\sum (x_i - \overline{x})^2 = 59.1$$

$\overline{x}$ הוא האומדן לתוחלת והאומדן לסטיית התקן הוא $\sqrt{s^2} \approx 7.7$.

המדגם גדול לכן ניתן להשתמש ברווח הסמך של ההתפלגות הנורמלית

$$\overline{x} - 1.645\frac{s}{\sqrt{175}} \le \mu \le \overline{x} + 1.645\frac{s}{\sqrt{175}}$$

כלומר $25.1 \le \mu \le 27.1$. יש מקום לחשוש שהחוסר דיוק בנתונים (טווחים גדולים של 5 שנים) משפיע על אמינות התשובה.
