# סטטיסטיקה – פתרון תרגיל 6

.1

a. פונקצית הנראות היא:

$$f(t_i;\lambda)=\begin{cases}\lambda e^{-\lambda t_i} & x\ge 0\\ 0 & x<0\end{cases}\Rightarrow L(\lambda)=\prod \lambda e^{-\lambda t_i}=\lambda^n e^{-\lambda\sum t_i}$$

נמצא את אומד הנראות המקסימאלי. נוכל להפעיל על המשוואה את הפונקציה log כיוון שהיא פונקציה מונוטונית עולה.

$$L(\lambda)=\lambda^n e^{-\lambda\sum t_i}\Rightarrow \log(L(\lambda))=\log(\lambda^n e^{-\lambda\sum t_i})\Rightarrow \log(L(\lambda))=n\log(\lambda)-\lambda\sum t_i$$

נגזור את הפונקציה ונקבל

$$\frac{d\log(L(\lambda))}{d\lambda}=\frac{n}{\lambda}-\sum t_i=0\Rightarrow \hat{\lambda}=\frac{n}{\sum t_i}=\frac{1}{\bar{t}}$$

נגזור פעם שנייה כדי לוודא שקיבלנו מקסימום:

$$\frac{d^2\log(L(\lambda))}{d^2\lambda}=-\frac{n}{\lambda^2}<0\qquad (n>0;\lambda^2>0)$$

b. במידה וערך הפרמטר היה ידוע לנו, הסיכוי שנמתין פחות מעשר דקות הוא $\Pr(T\le 10)=1-e^{-10\lambda}$. נשתמש באנ"מ שמצאנו בסעיף הקודם ונקבל $\hat{P}_{10}=1-e^{-\frac{10}{\bar{t}}}$.

.2 נניח שזמני ההמתנה לאוטובוס היו 7,4,4,4,1,5,12. נחשב את האומדן לסיכוי מהשאלה הקודמת:

$$\bar{t}=\frac{7+4+4+4+1+5+12}{7}=\frac{37}{7}=5.28$$

נציב באומד ונקבל:

$$\hat{P}_{10}=1-e^{-\frac{10}{5.28}}=1-0.1507863=0.8492137$$

a. (*) פונקצית הנראות של n תצפיות בלתי תלויות היא:

$$L(\sigma)=\frac{1}{(\sqrt{2\pi\sigma^2})^n}\cdot\exp(-\frac{\sum(X_i-\mu)^2}{2\sigma^2})$$

נעבור לסקאלת לוג ונקבל

$$\log(L(\sigma))=-n\log(\sqrt{2\pi\sigma^2})-\frac{\sum(X_i-\mu)^2}{2\sigma^2}$$

נגזור ונקבל

$$\frac{d\log(L(\sigma))}{d\sigma}=-n\frac{1}{\sqrt{2\pi\sigma^2}}\cdot\frac{1}{2}\cdot\frac{4\pi\sigma}{\sqrt{2\pi\sigma^2}}+[\sum(X_i-\mu)^2]\cdot\frac{4\sigma}{(2\sigma^2)^2}$$

$$=-\frac{n}{\sigma}+[\sum(X_i-\mu)^2]\cdot\frac{1}{\sigma^3}$$

נשווה ל-0

$$-\frac{n}{\sigma}+[\sum(X_i-\mu)^2]\cdot\frac{1}{\sigma^3}=0\Rightarrow\sigma^2=\frac{\sum(X_i-\mu)^2}{n}$$

נגזור פעם שנייה

$$\frac{d^2\log(L(\sigma))}{d^2\sigma}=\frac{n}{\sigma^2}-3[\sum(X_i-\mu)^2]\cdot\frac{1}{\sigma^4}$$

נציב את הנקודה כדי לראות שזו אכן נקודת מקסימום ונקבל

$$\frac{d^2\log(L(\sqrt{\frac{\sum(X_i-\mu)^2}{n}}))}{d^2\sigma}=\frac{n^2}{\sum(X_i-\mu)^2}-3\cdot\frac{n^2}{\sum(X_i-\mu)^2}=-2\cdot\frac{n^2}{\sum(X_i-\mu)^2}<0$$

כלומר זו אכן נקודת מקסימום.

.b

i. **לא נכון**. זה אומד מוטה כיוון שנוסחת השונות תיאורטית היא $\frac{\sum(X_i-\mu)^2}{N-1}$. כלומר, בנוסחה התיאורטית אנו מחלקים ב-N-1. לעומת זאת, באומד שלפנינו יש חלוקה ב-N. לכן, האומד לבטח מוטה.

ii. **לא נכון**. כפי שציינו בסעיף הקודם, החלוקה בשונות תיאורטית היא ב-N-1. כלומר אנו מחלקים במספר קטן יותר ולכן הערכים בפועל הם גדולים יותר. מכאן, שהאומד אינו מוטה כלפי מעלה (אלא כלפי מטה).

iii. **נכון**. האומד מוטה כלפי מטה כפי שהוסבר בסעיף הקודם.

iv. **לא נכון**. כאמור, בחישוב של שונות אנו מחלקים ב-N-1 ולכן אומד זה והשונות יחזירו ערכים שונים.

v. **נכון**. בעצם בעת חישוב האומדן זה למעשה מה שאנחנו עושים: אנו מחשבים את המרחק של כל איבר מהממוצע, מעלים בריבוע ועושים לסכום כל הסטיות ממוצע.

vi. **לא נכון**. אין הכרח שהממוצע והתוחלת יהיו שווים אלא אם המדגם מכיל את כלל האוכלוסייה. לכן, הערך שנקבל אינו הסטייה הריבועית הממוצעת מהתוחלת.

c. נפתר בסעיף a.

.3

a. **כן.** בשאלה 1 ראינו שאומד נראות מקסימאלית ל- $\lambda$ הוא $\frac{1}{\bar{t}}$. תוחלת של מ"מ המתפלג מעריכית היא $\frac{1}{\lambda}$. נשתמש בתכונת האינווריאנטיות ונקבל כי $\bar{t}$ הוא אכן אומד נראות מקסימאלית לתוחלת.

b. **כן.** כפי שראינו בכיתה ממוצע הוא תמיד אומד חסר הטיה לתוחלת.

c. **כן.** כפי שראינו בכיתה ממוצע הוא תמיד אומד עקיב לתוחלת.

d. (*) ע"פ משפט הגבול המרכזי, זמן ההמתנה הממוצע מתפלג בקירוב להתפלגות נורמאלית. סכום זמני ההמתנה מתפלג ככל הנראה התפלגות גמא (שעל פי ויקיפדיה ההתפלגות המעריכית היא מקרה פרטי של התפלגות זו).

e. נחשב את תוחלת האומד

$$E(\overline{T}) = E(\frac{\sum Xi}{n}) = \frac{1}{n}E(\sum Xi) = \frac{1}{n}\sum E(Xi) = \frac{n}{n}E(Xi) = \frac{1}{\lambda}$$

קיבלנו שתוחלת האומד שווה לתוחלת המשתנה המקרי (הפרמטר) ולכן האומד הוא אומד חסר הטיה.

f.

$$V(\overline{T}) = V(\frac{\sum Xi}{n}) = \frac{1}{n^2}V(\sum Xi) = \frac{1}{n^2}\sum V(Xi) = \frac{1}{n}V(Xi) = \frac{1}{n}\cdot\frac{1}{\lambda^2}$$

לכן שונות האומד כפונקציה של זמן ההגעה הטיפוסי היא:

$$f(x) = \frac{1}{n}x^2 \Rightarrow f[E(T)] = \frac{1}{n}\cdot[E(T)]^2$$

4.

a. ע"פ המדגם, אבנר גלש ל-4 אתרים שתמכו בפיירפוקס ולאתר אחד שלא תמך בדפדפן. לכן, פונקצית הנראות ע"פ המדגם היא $L(p) = (1-p)^4 p$ .

b. נגזור את פונקצית הנראות שמצאנו כדי למצוא מקסימום:

$$\frac{dL(p)}{dp} = -4(1-p)^3 p + (1-p)^4 = 0 \Rightarrow (1-p)^3(1-5p) = 0 \Rightarrow \begin{matrix} p_1 = 1 \\ p_2 = \frac{1}{5} \end{matrix}$$

נמצא נגזרת שנייה כדי לבדוק האם אחד מהערכים שמצאנו הוא המקסימום:

$$\frac{d^2L(p)}{d^2p} = -3(1-p)^2(1-5p) - 5(1-p)^3 = (1-p)^2(-8+20p)$$

$$\frac{d^2L(1)}{d^2p} = 0, \qquad \frac{d^2L(\frac{1}{5})}{d^2p} = (1-\frac{1}{5})^2(-4) = -\frac{16}{5} < 0$$

לכן, ערך p הסביר ביותר הוא כמובן **0.25**.

c. באופן כללי, פונקצית הנראות ע"פ המדגם בו האתר ה-k הוא הראשון שאינו תומך בדפדפן תהיה $L(p) = (1-p)^{k-1} p$ . נגזור את הפונקציה כדי לקבל נקודות חשודות כמקסימום:

$$\frac{dL(p)}{dp} = -(k-1)(1-p)^{k-2} p + (1-p)^{k-1} = 0 \Rightarrow (1-p)^{k-2}(1-kp) = 0 \Rightarrow \begin{matrix} p_1 = 1 \\ p_2 = \frac{1}{k} \end{matrix}$$

נגזר פעם שנייה כדי לבדוק מי מהנקודות היא נק' מקסימום (אם בכלל):

$$\frac{d^2L(p)}{d^2p} = -(k-2)(1-p)^{k-3}(1-kp) - k(1-p)^{k-2} = (1-p)^{k-3}(k^2p - (2+p)k + 2)$$

$$\frac{d^2L(1)}{d^2p} = 0, \qquad \frac{d^2L(\frac{1}{k})}{d^2p} = (1-\frac{1}{k})^{k-3}(k-2k-1+2) = (\frac{k-1}{k})^{k-3}(1-k) < 0$$

(תחת ההנחה ש-1<k)

כלומר אומד הנראות המקסימאלית לשיעור האתרים שאינם תומכים בדפדפן עבור המקרה המתואר הוא $\hat{p} = \frac{1}{k}$ .

.5

a. אנו רוצים רמת ביטחון של 90% ולכן אנו צריכים את האחוזון ה- $1-\alpha = 0.9 \Rightarrow \alpha = 0.1$, נשתמש בנוסחה $\bar{y} \pm \frac{\sigma}{\sqrt{n}} Z_{1-\frac{\alpha}{2}} = \bar{y} \pm \frac{12}{\sqrt{n}} Z_{0.95}$ .

b. תחילה נחשב את הממוצע:

$$\bar{y} = \frac{175+190+159+164+182+177}{6} = 174.5$$

ולכן הטווח הוא:

$$174.5 \pm \frac{12}{\sqrt{6}} 1.645 = 174.5 \pm 8.059 \Rightarrow [166.44, 182.56]$$

c. לא. אבנר היה יכול להגיד שאם היה עושה מספר מדגמים, אז ב-90% מהמדגמים היו כוללים את התוחלת האמיתית.

d. על אבנר לעמוד את השונות באמצעות הנוסחה $\sigma^2 = \frac{\sum (x_i - \bar{x})^2}{n}$ ולכן לעדכן את הנוסחה לטווח ל- $\bar{y} \pm \frac{\hat{\sigma}}{\sqrt{n}} t_{(n-1, 1-\frac{\alpha}{2})}$ .

e. נחשב תחילה את השונות:

$$\hat{\sigma}^2 = \frac{0.25+240.25+240.25+110.25+56.25+6.25}{6} = \frac{653.5}{6} = 130.7$$

לכן

$$\hat{\sigma} = \sqrt{130.7} = 11.43$$

ומכאן שהטווח החדש יהיה

$$174.5 \pm \frac{11.43}{\sqrt{6}} 2.015 = 174.5 \pm 9.40 \Rightarrow [165.09, 183.9]$$

f. בעקבות העדר מידע על שונות הגבהים הטווח גדל. אין זה מפתיע כיוון שהגדלנו את חוסר הוודאות שלנו בנתונים. כמו כן, ההבדל בגודל סטיית התקן היה קטן יחסית (השונות קטנה מעט לעומת חוסר הוודאות שנוסף). במידה וסטיית התקן הייתה קטנה משמעותית, ייתכן והטווח היה מצטמצם.

g. אבנר יכול לדרוש זאת, אך דרישה הוא יקבל במקרה זה טווח של $(-\infty, \infty)$ . טווח זה אינו אינפורמטיבי כלל ולכן אין טעם לדרוש זאת.

h. אבנר מעוניין לגלות P(X>200). אנו יודעים שגבהים מתפלגים נורמאלית ולכן נחפש את ההסתברות לכך ע"י נרמול באמצעות התוחלת (שאינה ידוע לנו) והשונות. נוכל להפעיל פונקציה זו על רווח הסמך המתאים בהמשך כדי לקבל טווח שיתפוס את הפרופורציה האמיתי ב-90% מהמדגמים (ע"פ אינווריאנטיות של רווח סמך).

הפונקציה אותה נפעיל על רווח הסמך שחושב בסעיף b תהיה (שונות ידועה):

$$P(X > 200) = P(\frac{X-\mu}{12} > \frac{200-\mu}{12}) = 1 - \phi(\frac{200-\mu}{12})$$

ואם נשתמש באומד לשונות נפעיל את הפונקציה $1-\phi(\frac{200-\mu}{11.43})$ על רווח הסמך שחושב בסעיף e.

i. נפעיל את הפונקציה הראשונה על רווח הסמך שחישבנו בסעיף b ונקבל

$$[1-\phi(\frac{200-166.44}{12}),1-\phi(\frac{200-182.56}{12})] \Rightarrow [0.00258,0.07306]$$

ואם נפעיל את הפונקציה השנייה על רווח הסמך שחישבנו בסעיף e נקבל

$$[1-\phi(\frac{200-165.09}{11.43}),1-\phi(\frac{200-183.9}{11.43})] \Rightarrow [0.001128,0.079480]$$

j. על מנת שהטווח יהיה קטן מ-2 ס"מ, על הביטוי $\frac{\sigma}{\sqrt{n}}Z_{1-\frac{\alpha}{2}}$ להיות קטן מ-1 (כי אנו מוסיפים ומחסרים אותו ממוצע המדגם). לכן נקבל:

$$\frac{\sigma}{\sqrt{n}}Z_{1-\frac{\alpha}{2}} < 1 \Rightarrow \frac{12}{\sqrt{n}}1.645 < 1 \Rightarrow 19.74 < \sqrt{\mathrm{n}} \Rightarrow n > 390$$

עליו לדגום למעלה מ-390 אנשים.

k. נחשב שוב את 2 הסעיפים. אבנר רוצה ביטחון של 95% ולכן $1-\alpha = 0.95 \Rightarrow \alpha = 0.05$ לכן הטווח יהיה

$$\bar{y} \pm \frac{\sigma}{\sqrt{n}}Z_{1-\frac{\alpha}{2}} = \bar{y} \pm \frac{12}{\sqrt{n}}Z_{0.975} = 174.5 \pm \frac{12}{\sqrt{6}}1.960 \Rightarrow [164.9,184.1]$$

$$\bar{y} \pm \frac{\hat{\sigma}}{\sqrt{n}}t_{(n-1,1-\frac{\alpha}{2})} = \bar{y} \pm \frac{\hat{\sigma}}{\sqrt{n}}t_{(5,0.975)} = 174.5 \pm \frac{11.43}{\sqrt{6}}2.571 \Rightarrow [162.5,186.5]$$

הטווחים גדלו ואין זה מפתיע כיוון שרצינו בטחון גדול יותר שהפרמטר ייפול בתוך רווח הסמך. לכן, גודל הטווח גדל וכתוצאה מכך גם הסיכוי שהפרמטר ייפול בטווח של 95% מהמדגמים.

6. $X_i \sim N(175,144)$

a.

$$P(X > 190) = P(\frac{X-175}{12} > \frac{190-175}{12}) = P(\frac{X-175}{12} > 1.25) = 1-\phi(1.25)$$
$$= 1-0.8944 = 0.1056$$

b.

i.

```
rsMidgam = matrix(0,1000,20)
rsMean = vector(length=1000)
for (i in 1:1000)
{
        midgam = rnorm(20,175,12)
        for (j in 1:20)
        {
          rsMidgam[i,j]=midgam[j]
        }
        rsMean[i] = mean(midgam)
}
```

.ii

```
rsP = rsMean + 12/sqrt(20)*1.645
rsN = rsMean - 12/sqrt(20)*1.645
```

.iii

```
rs190P = vector(length=1000)
rs190N = vector(length=1000)

for (i in 1:1000)
{
	rs190P[i] = 1 - pnorm((190-rsP[i])/12)
	rs190N[i] = 1 - pnorm((190-rsN[i])/12)
}
```

.iv

```
> isThere = vector(length=1000)
> for (i in 1:1000)
+ {
+ isThere[i]=rs190P[i]>=0.1056 && rs190N[i]<=0.1056
+ }
>
> sum(isThere)/1000
[1] 0.903
```

.v אם היו בידנו אינסוף מדגמים, אחוז המדגמים שהיה מכיל את הסיכוי האמיתי היה כמובן 90% על פי העיקרון של רווח סמך.

.c

.i

```
rsPro = vector(length=1000)
for (i in 1:1000)
{
	rsPro[i] = sum(rsMidgam[i,]>190)/20
}
rsProP = rsPro + sqrt(rsPro *(1-rsPro)/20)*1.645
rsProN = rsPro - sqrt(rsPro *(1-rsPro)/20)*1.645
```

.ii

```
> rsProP = rsPro + sqrt(rsPro *(1-rsPro)/20)*1.645
> rsProN = rsPro - sqrt(rsPro *(1-rsPro)/20)*1.645
>
> isTherePro = vector(length=1000)
> for (i in 1:1000)
+ {
+ isTherePro[i]=rsProP[i]>=0.1056 && rsProN[i]<=0.1056
+ }
> sum(isTherePro)/1000
[1] 0.873
```

אם היו בידנו אינסוף מדגמים, אחוז המדגמים שהיה מכיל את p האמיתי היה בערך 90% על פי העיקרון של רווח סמך.

.d

i.

```
> mean(rs190P-rs190N)
[1] 0.1373283
> mean(rsProP-rsProN)
[1] 0.2020025
```

ע"פ התוצאות, רווחי הסמך בסעיף B היו קטנים משמעותית (כמעט בחצי) מרווחי הסמך בדרך שחושבה בסעיף C.

ii. בחישוב שביצענו, עבור 1000 מדגמים בני 20, היה יתרון לשיטה שחושבה בסעיף B לעומת הגרסה של סעיף C. הסיבה לכך היא כיוון שבסעיף C קיים יותר חוסר וודאות (בשל הקירוב לנורמלי ושימוש באומד לשונות).

iii. במצב זה כמובן שנעדיף את השיטה שחושבה בסעיף B כיוון שבאמצעותה יש לנו ביטחון רב יותר שהפרמטר אכן ייפול במדגם וכן גודל רווח הסמך הוא קטן יותר (ויעידו על כך תוצאות הסימולציה).

e. (*) את החישוב בסעיף B חישבנו תחת הידיעה שהתפלגות המדגם היא נורמאלית ושונות ההתפלגות הייתה ידועה. בשל סיבה זו, קיבלנו שבאחוז גדול יותר של רווחי הסמך, נפל הפרמטר האמיתי. לכן, נעדיף שיטה זו באופן אוטומטי רק במידה וידוע שהתפלגות הנתונים היא התפלגות נורמאלית.

7.

a. רווח הסמך אינו בהכרח יחיד. לדוגמה, ראינו בכיתה שרווח סמך לפרופורציה p ניתן לחשב באמצעות 2 גרסאות. בגרסה הראשונה, כדי לאמוד את השונות (הכוללת את הפרמטר) אנו משתמשים באומד ל-p גם בשונות. בגרסה השנייה, אנו מחליפים את השונות בערך המקסימאלי האפשרי. כלומר, ראינו 2 דרכים לבחירת האומדים בגבולות רווח הסמך ולכן רווח הסמך אינו בהכרח יחיד.

b. נשים לב שלמרות שהדוגמאות שלמדנו עד כה היו סימטריות סביב האומדן הנקודתי, אך ע"פ ההגדרה אין הכרח שרווחי האפסילון סביב האומדן יהיו סימטריים. כמו כן, נזכור שההבטחה שב-90% מרווחי הסמך ייפול הפרמטר האמיתי מתייחס למצב שלפני חישוב של רווח סמך ספציפי. כלומר, לאחר חישוב רווח הסמך **לא** ניתן להגיד שיש וודאות של 90% שהפרמטר ייפול בתוכו. לכן, נעדיף בכל מקרה לדווח גם את אומד הנקודה וגם את רווח הסמך.