<u>סטטיסטיקה / תרגיל 6#</u>

אריאל סטולרמן
קבוצה 03

אמידה נקודתית

(1)

זמן הגעתו של אוטובוס מתפלג מעריכית $T_i \sim \exp(\lambda)$, $f(t_i;\lambda) = \begin{cases} \lambda e^{-\lambda t_i}, x \geq 0 \\ 0, x < 0 \end{cases}$.

כיוון שההתפלגות מעריכית, ידוע כי: $E(t_i) = \frac{1}{\lambda}$, $P(T \leq t_i) = 1 - e^{-\lambda t_i}$

(a) פונקציית הנראות ואומד הנראות המקסימלית ל-λ במקרה כללי של זמני המתנה $(t_1, \dots, t_n)$:

$$given\ \lambda: f(t_1, \dots, t_n) = \prod_{i=1}^{n} f(t_i) = \prod_{i=1}^{n} \lambda e^{-\lambda t_i} = \lambda^n e^{-\lambda \sum_{i=1}^{n} t_i}$$

$$\Rightarrow \boldsymbol{L(\lambda; t_1, \dots, t_n) = \lambda^n e^{-\lambda \sum_{i=1}^{n} t_i}}, \qquad l(\lambda) = \ln\big(L(\lambda)\big) = n \cdot \ln(\lambda) - \lambda \sum_{i=1}^{n} t_i$$

$$\frac{\partial l(\lambda)}{\partial \lambda} = \frac{n}{\lambda} - \sum_{i=0}^{n} t_i = 0 \Leftrightarrow \lambda = \frac{n}{\sum_{i=0}^{n} t_i} = \frac{1}{\bar{t}} \Rightarrow \boldsymbol{\hat{\lambda} = \frac{1}{\bar{t}}}$$

כיוון שמתקיים $\widehat{f(\theta)} = f(\hat{\theta})$, אז אין צורך לגזור פעם שניה ולהוכיח שזהו אכן מקס', כי הקצב הוא פשוט פונ' של התוחלת: $\frac{1}{E(t_i)}$.

(b) אמינות האוטובוס: הסיכוי שיגיע תוך פחות מ-10 דקות. להלן אנ"מ לאמינות של קו 25 בתחנת האונ' עבור מדגם כללי כקודם $(t_1, \dots, t_n)$:

$$\widehat{P(t_i \leq 10)} = 1 - e^{-\hat{\lambda} \cdot 10} = \boldsymbol{1 - e^{-\frac{10}{\bar{t}}}}$$

(2)

האומדן לסיכוי מהסעיף הקודם בהינתן שזמני ההמתנה היו 12,5,1,4,4,4,7 הוא:

$$\bar{t} = \frac{(12 + 5 + 1 + 4 + 4 + 4 + 7)}{7} = 5.285 \Rightarrow P(t_i \leq 10) = 1 - e^{-\frac{10}{5.285}} = \boldsymbol{0.849}$$

(i)

יהיו $X_i \sim N(\mu, \sigma^2)$, $1 \leq i \leq n$. נראה את אומד הנראות המקסימלית לשונות σ^2:

$$f_X(x) = \frac{1}{\sqrt{2\pi\sigma^2}} \cdot exp\left(-\frac{(x-\mu)^2}{2\sigma^2}\right)$$

$$\Rightarrow L(\sigma^2;\ X_1, \dots, X_n) = \prod_{i=1}^{n} f_X(X_i) = (2\pi\sigma^2)^{-\frac{n}{2}} \cdot exp\left(-\frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (X_i - \mu)^2\right)$$

$$\Rightarrow l(\sigma^2) = -\frac{n}{2}\ln(2\pi\sigma^2) - \frac{1}{2\sigma^2} \cdot \sum_{i=1}^{n} (X_i - \mu)^2 = -\frac{1}{2}\left[n \cdot \ln(2\pi\sigma^2) + \frac{1}{\sigma^2} \cdot \sum_{i=1}^{n} (X_i - \mu)^2\right]$$

$$\Rightarrow \frac{\partial l(\sigma^2)}{\partial \sigma^2} = -\frac{1}{2}\left[2\pi n \cdot \frac{1}{2\pi\sigma^2} - \frac{1}{\sigma^4}\cdot\sum_{i=1}^{n}(X_i-\mu)^2\right]$$

נשווה לאפס למציאת קיצון:

$$2\pi n \cdot \frac{1}{2\pi\sigma^2} - \frac{1}{\sigma^4}\cdot\sum_{i=1}^{n}(X_i-\mu)^2 = 0 \quad \Leftrightarrow \quad \frac{n}{\sigma^2} = \frac{1}{\sigma^4}\cdot\sum_{i=1}^{n}(X_i-\mu)^2 \quad \Leftrightarrow \quad /\cdot\sigma^2 \neq 0$$

$$n = \frac{1}{\sigma^2}\cdot\sum_{i=1}^{n}(X_i-\mu)^2 \quad \Leftrightarrow \quad \sigma^2 = \frac{1}{n}\cdot\sum_{i=1}^{n}(X_i-\mu)^2$$

נגזור פעם שניה לבדוק שזו אכן נק' מקסימום (לכל μ):

$$\frac{\partial^2 l(\sigma^2)}{\partial \sigma^4} = \frac{\partial}{\partial \sigma^2}\cdot\left[-\frac{1}{2}\left(\frac{n}{\sigma^2} - \frac{1}{\sigma^4}\cdot\sum_{i=1}^{n}(X_i-\mu)^2\right)\right] = \frac{n}{2\sigma^4} - \frac{1}{\sigma^6}\cdot\sum_{i=1}^{n}(X_i-\mu)^2 < 0 \Leftrightarrow \ /\cdot 2\sigma^4 > 0$$

$$n - \frac{2}{\sigma^2}\cdot\sum_{i=1}^{n}(X_i-\mu)^2$$

נציב את σ^2 שקיבלנו:

$$n - 2\cdot\frac{\sum_{i=1}^{n}(X_i-\mu)^2}{\frac{1}{n}\cdot\sum_{i=1}^{n}(X_i-\mu)^2} = n - 2n = -n < 0 \Rightarrow$$ זוהי אכן נקודת מקסימום

כעת נציב את $\hat{\mu}_{MLE}$ כדי למקסם את האומד שקיבלנו, והוא הרי $\bar{X}$, מכאן קיבלנו כנדרש כי:

$$\widehat{\sigma^2}_{MLE} = \sum_{i=1}^{n}(X_i-\bar{X})^2$$

(ii)

(1) האומד חסר הטיה: **לא**

$$E\left(\frac{1}{n}\cdot\sum_{i=1}^{n}(X_i-\bar{X})^2\right) = \frac{1}{n}\cdot E\left[\sum_{i=1}^{n}(X_i-\bar{X})^2\right] = \frac{1}{n}\cdot E\left[\sum_{i=1}^{n}X_i^2 - 2\sum_{i=1}^{n}X_i\bar{X} + \sum_{i=1}^{n}\bar{X}^2\right] =$$

$$\frac{1}{n}\cdot E\left[\sum_{i=1}^{n}X_i^2 - nE(\bar{X})^2\right] = \begin{bmatrix} E(X_i^2) = Var(X_i) + E(X_i)^2 = \sigma^2 + \mu^2 \\ E(\bar{X}^2) = Var(\bar{X}) + E(\bar{X})^2 = \frac{\sigma^2}{n} + \mu^2 \end{bmatrix} = \frac{1}{n}\cdot\left[n\sigma^2 + n\mu^2 - \frac{n\sigma^2}{n} - n\mu^2\right] =$$

$\boldsymbol{\frac{n-1}{n}\cdot\sigma^2 \neq \sigma^2}$ ⇒ האומד מוטה

(2) האומד מוטה כלפי מעלה, כלומר נוטה להחזיר ערכים **גדולים** מהערך האמיתי: **לא**, כפי שניתן לראות, האומד מחזיר ערך קטן מ-σ^2 כיוון שערכו הוא שבר (תלוי ב-n) כפול σ^2 (הולך ונהיה זניח כאשר $n \to \infty$).

(3) האומד מוטה כלפי מטה, כלומר נוטה להחזיר ערכים **קטנים** מהערך האמיתי: **כן**, בדיוק מהסיבה בסעיף הקודם.

(4) ניתן לראות את האומד כחישוב של שונות רק אם נסתכל על המדגם כאוכלוסיה: **כן**, אם אנו מסתכלים על המדגם כאוכלוסיה, נוסחת האומד הנ"ל היא בדיוק נוסחת חישוב השונות.

(5) ניתן להסתכל על האומד כ"הסטייה הריבועית הממוצעת מהממוצע": **כן**, זה בדיוק מה שמתארת הנוסחה.

(6) ניתן להסתכל על האומד כ"הסטייה הריבועית הממוצעת מהתוחלת": **לא**, כיוון שזוהי הגדרת השונות, וזהו רק אומד לשונות (ניתן היה להגיד כן כמו בסעיף (4) אם המדגם היה האוכלוסיה שלנו).

(3)

(a) הממוצע הוא אכן אומד נראות מקסימלית לתוחלת.

(b) הממוצע אכן חסר הטייה.

(c) הממוצע הוא עקיב.

(e) הוכחה שהאומד חסר הטייה:

$$E(\bar{T}) = E\left(\frac{1}{n}\cdot\sum_{i=1}^{n} T_i\right) = \underset{T_i\sim \exp(\lambda)\Rightarrow E(T_i)=\frac{1}{\lambda}}{\frac{1}{n}\cdot\sum_{i=1}^{n} E(T_i)} = \frac{1}{n}\cdot n\cdot\frac{1}{\lambda} = \frac{\mathbf{1}}{\boldsymbol{\lambda}} = \boldsymbol{E(T_i)}$$

(f) שונות האומד לתוחלת כפונקציה של זמן ההגעה הטיפוסי (נזכור כי $Var(T_i) = \frac{1}{\lambda^2}$):

$$Var(\bar{T}) = Var\left(\frac{1}{n}\cdot\sum_{\mathrm{i=1}}^{\mathrm{n}} T_i\right) = \frac{1}{\mathrm{n}^2}\cdot\sum_{i=1}^{n} Var(T_i) = \frac{1}{n^2}\cdot\frac{n}{\lambda^2} = \boldsymbol{\frac{1}{n\lambda^2}}$$

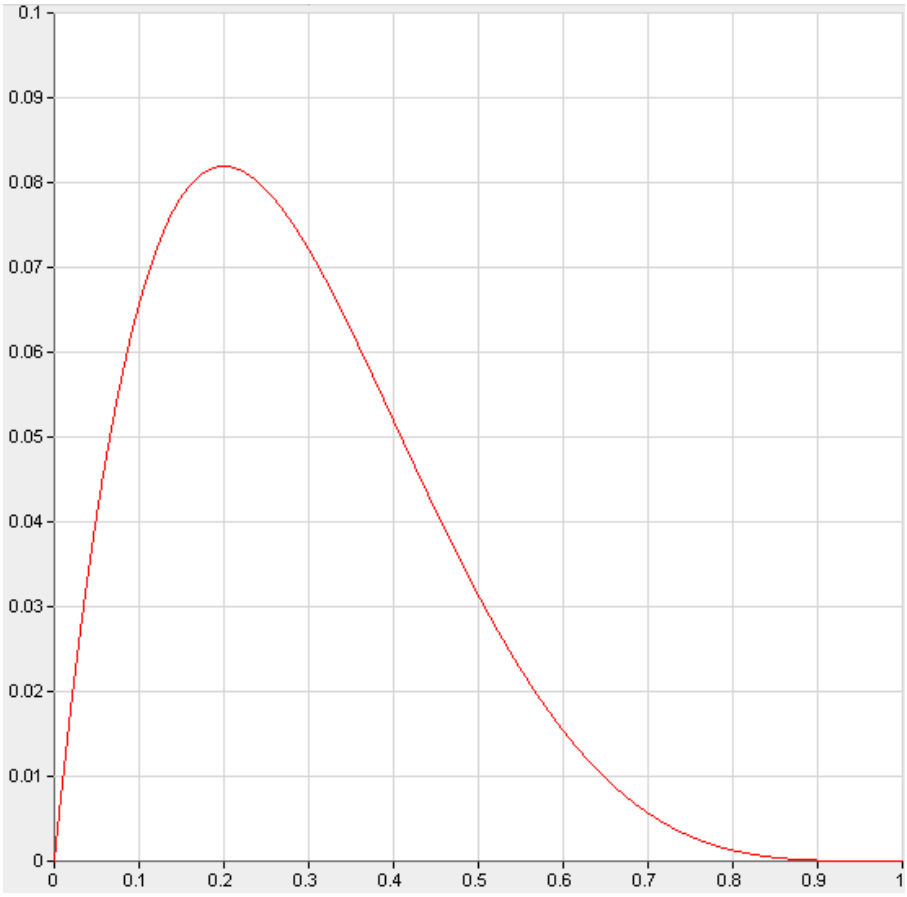
(4) שיעור אתרי האינטרנט שאינם תומכים בפיירפוקס הוא p.

$$X_i \sim Bernoulli(p)$$

(a) להלן חישוב פונ' הנראות הגיאומטרית:

$$L(p; x_1 = 0, x_2 = 0, \dots, x_5 = 1) = \prod_{i=1}^{5} p^{II\{x_i=1\}}\cdot(1-p)^{II\{x_i=0\}} = (1-p)^4\cdot p$$

להלן שרטוט של $f(p) = (1-p)^4\cdot p$:



(*b*) נחשב את אומד הנראות המקסימלית:

$$\frac{\partial}{\partial p}[(1-p)^4 \cdot p] = -4p(1-p)^3 + (1-p)^4 = 0 \quad \Leftrightarrow \quad p = 1, \boldsymbol{p = 0.2}$$

ומהגרף קל לראות שאומד הנראות המקסימלית הוא $\boldsymbol{p = 0.2}$.

(*c*) עבור המקרה ה-k:

$$L(p; x_1 = 0, \dots, x_k = 1) = (1-p)^{k-1} \cdot p$$

$$\frac{\partial}{\partial p}[(1-p)^{k-1} \cdot p] = -(k-1)(1-p)^{k-2} \cdot p + (1-p)^{k-1} = (1-p)^{k-2} \cdot (1-kp) = 0 \quad \Leftrightarrow$$

$$p = 1, \boldsymbol{p = \frac{1}{k}}$$

קל לראות (ללא גזירה שניה) כי אומד הנראות המקסימלית למקרה ה-k הוא $\boldsymbol{p = \frac{1}{k}}$.

רווח בר סמך

(5) $X_i \sim N(\mu, \sigma^2 = 144)$,$\sigma = 12$

(*a*) אנו רוצים רמת סמך של 90% ולכן נקח $\alpha = 0.1 \Leftarrow$ ניקח את $Z_{1-\frac{\alpha}{2}} = Z_{0.95}$. במקום התוחלת ניקח את $\hat{\mu}_{MLE} = \bar{X}$,

ולכן גבולות הר״ס שעל אבנר לקחת: $\bar{X} \pm \frac{\sigma}{\sqrt{n}} \cdot Z_{0.95} = \bar{X} \pm \frac{12 \cdot 1.645}{\sqrt{n}} = \boldsymbol{\bar{X} \pm \frac{19.74}{\sqrt{n}}}$

(*b*) נחשב את $\bar{X}$ עבור המדגם:

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

רווח הסמך יהיה:

$$\Rightarrow \left[174.5 - \frac{19.74}{\sqrt{6}}, 174.5 + \frac{19.74}{\sqrt{6}}\right] = [\boldsymbol{166.44, 182.55}]$$

(*c*) אבנר לא יכול לטעון שהסיכוי שהתוחלת האמיתית נמצאת בטווח שחישב הוא 90%, כיוון שברגע שחישב את הטווח התשובה היא או כן או לא, כלומר או 0% או 100%.

(*d,e*) נאמוד את השונות באומד חסר הטיה ונשתמש בהתפלגות t של סטודנט המתאימה:

$$X_i \sim N(\mu, \sigma^2) \Rightarrow \hat{\sigma}_{MLE} = \sqrt{\frac{\sum_{i=1}^{n}(X_i - \bar{X})^2}{n-1}} \Rightarrow$$

במקרה שלנו:

$$\hat{\sigma}_{MLE} = \sqrt{\frac{(175-174.5)^2 + \dots + (177-174.5)^2}{6-1}} = \sqrt{130.7} = 11.43, \qquad t_{n-1,1-\frac{\alpha}{2}} = t_{5,0.95} = 2.015$$

$$\Rightarrow \bar{X} \pm \frac{\hat{\sigma}_{MLE}}{\sqrt{n}} \cdot t_{n-1,1-\frac{\alpha}{2}} = 174.5 \pm \frac{11.43}{\sqrt{6}} \cdot 2.015 \Rightarrow [\boldsymbol{165.09, 183.9}]$$: רווח הסמך החדש

(*f*) הטווח **גדל** כתוצאה מהיעדר מידע על שונות הגבהים. מצד אחד זה לא מפתיע כיוון שנוספה אי ודאות למערכת, אך מצד שני זה כן כיוון שישנה החטאה כלשהי בחישוב האומד של σ, במקרה זה קיבלנו אומד הקטן ממהנתונים מאתר האו״ם, ואולי היינו מצפים להקטנת הטווח.

(*g*) טווח זה יהיה כמובן כל הישר הממשי, כיוון ש-$\lim_{x\to 1}\Phi(x)=\infty$, ולכן בחישוב הטווח נוסיף ונחסר מ-$\bar{X}$ $\infty\pm$. היות וזהו מידע ריק כיוון שברור ש-$\mu\in\mathbb{R}$, אז התשובה היא לא (או כן: הטווח יהיה $(-\infty,+\infty)$).

(*h*) במקרה זה אנו צריכים למדוד את:

$$P(X_i>200)=1-\Phi\left(\frac{200-\mu}{\sigma}\right)$$

כיוון שזוהי פונקציה מונוטונית והפיכה (שכן Φ פונקציה מונוטונית והפיכה), אז ניתן פשוט להפעיל את הפונקציה על רווח הסמך שקיבלנו בסעיף (*b*):

$$\left[1-\Phi\left(\frac{200-166.44}{12}\right),1-\Phi\left(\frac{200-182.55}{12}\right)\right]=[\mathbf{0.0026},\mathbf{0.0735}]$$

ועבור שימוש באומד שלו לשונות נשתמש ברווח הסמך שקיבלנו בסעיף (*e*):

$$\left[1-\Phi\left(\frac{200-165.09}{11.43}\right),1-\Phi\left(\frac{200-183.9}{11.43}\right)\right]=[\mathbf{0.0011},\mathbf{0.0808}]$$

(*i*) האומדן לטווח במדגם שלו יהיה:

(*j*) נחשב את גודל המדגם הרצוי לפי הנתונים מסעיף (*b*), שהם: $\sigma=12, Z_{0.95}=1.645$:

$$\frac{2\cdot 12}{\sqrt{n}}\cdot 1.645<2\Leftrightarrow n>389.66\Rightarrow \boldsymbol{n=390}$$ נצטרך לקחת:

אבנר יצטרך לקחת מדגם בן 390 תצפיות לפחות ע"י לקבל טווח שלא יעלה על 2 ס"מ.

(*k*)

עבור רמת סמך של 95%, ניקח $\alpha=0.05\Leftarrow$ ניקח את $Z_{1-\frac{\alpha}{2}}=Z_{0.975}=1.960$ ואת $t_{5,0.975}=2.571$ ונחשב את סעיפים (*b*) ו-(*e*) מחדש:

חישוב סעיף (*b*) עבור רמת סמך של 95%:

$$\left[174.5-\frac{12}{\sqrt{6}}\cdot 1.96,174.5+\frac{12}{\sqrt{6}}\cdot 1.96\right]=[\mathbf{164.89},\mathbf{184.1}]$$

חישוב סעיף (*e*) עבור רמת סמך של 95%:

$$\left[174.5-\frac{11.43}{\sqrt{6}}\cdot 2.571,174.5+\frac{11.43}{\sqrt{6}}\cdot 2.571\right]=[\mathbf{162.5},\mathbf{186.49}]$$

הרווח **גדל** ואין זה מפתיע כיוון שאנו דורשים רמת סמך גבוהה יותר, כלומר דורשים תנאי חזק יותר – טווח שיהיה סיכוי גבוה יותר שיכיל את התוחלת מאשר הטווח הקודם.

(6)

התפלגות הגבהים במדינה: $X_i\sim N(\mu=175,\sigma=12)$

(*a*)

$$P(X_i>190)=1-\Phi\left(\frac{190-175}{12}\right)=\mathbf{0.1056}$$

כלומר ישנו סיכוי של 10.56% שגובהו של האדם הבא שנפגוש ברחוב יהיה מעל 190 ס"מ.

(*b*)

i. הגרלת 1000 מדגמים מקריים בגודל 20:

```
data=matrix(nrow=1000, ncol=20)
for(i in 1:1000){
        data[i,]=rnorm(20,175,12)
}
```

ii. חישוב רב"ס לתוחלת בכל אחד מהמדגמים ברמת סמך של 90%:

```
data.ci=matrix(nrow=1000, ncol=2)
for(i in 1:1000){
        data.ci[i,1]=(mean(data[i,])-12/sqrt(20)*qnorm(0.95))
        data.ci[i,2]=(mean(data[i,])+12/sqrt(20)*qnorm(0.95))
}
```

iii. חישוב רב"ס לסיכוי שאדם יהיה גבוה מ-190 ס"מ בכל אחד מהמדגמים:

```
data.bt190ci=matrix(nrow=1000, ncol=2)
for(i in 1:1000){
        data.bt190ci[i,1]=1-pnorm((190-data.ci[i,1])/12)
        data.bt190ci[i,2]=1-pnorm((190-data.ci[i,2])/12)
}
```

iv. האחוז מתוך 1000 המדגמים בו הרב"ס מכילים את הערך האמיתי של הפרמטר (0.1056) הוא **89.8%** (בדגימות אלה):

```
count=0
for(i in 1:1000){
        if((0.1056 >= data.bt190ci[i,1])&&(0.1056 <= data.bt190ci[i,2]))(count=count+1)
}
count/1000                (מדפיס 0.898)
```

v. אם היו בידינו ∞ מדגמים בגודל 20, אחוז המדגמים שהיו מכילים את הערך האמיתי היה **90%**.

(*c*)

i. חישוב רב"ס בביטחון 90% לפרופורציית התושבים מעל 190 ס"מ:

```
data.bt190cip=matrix(nrow=1000, ncol=2)
for(i in 1:1000){
        cnt=0
        for(j in 1:20){
                if(data[i,j]>190)(cnt=cnt+1)}
        p_hat=cnt/20
```

```
    data.bt190cip[i,1]=(p_hat - sqrt((p_hat*(1-p_hat))/20)*qnorm(0.95))
    data.bt190cip[i,2]=(p_hat + sqrt((p_hat*(1-p_hat))/20)*qnorm(0.95))
}
```

ii. האחוז מתוך 1000 המדגמים בו הרב"ס מכילים את הערך האמיתי של הפרמטר (0.1056) הוא **89.4%** (בדגימות אלה):

```
count=0
for(i in 1:1000){
    if((0.1056 >= data.bt190cip[i,1])&&(0.1056 <= data.bt190cip[i,2]))(count=count+1)
}
count/1000          (מדפיס 0.894)
```

אם בידינו ∞ מדגמים בגודל 20, אחוז המדגמים שהיו מכילים את הערך האמיתי היה **90%**.

(*d*)

i. השיטה בסעיף (*b*) עדיפה מבחינת אורך הרב"ס, שכן בממוצע הרווחים קטנים יותר מאשר השיטה בסעיף (*c*):

```
cilen=c(1:1000)
for(i in 1:1000){
    cilen[i]=data.bt190ci[i,2]-data.bt190ci[i,1]
}
cilen_mean=mean(cilen)

ciplen=c(1:1000)
for(i in 1:1000){
    ciplen[i]=data.bt190cip[i,2]-data.bt190cip[i,1]
}
ciplen_mean=mean(ciplen)

cilen_mean          (ממוצע גודל הרווח עבור השיטה הראשונה: 0.1371188)
ciplen_mean         (ממוצע גודל הרווח עבור השיטה השניה: 0.2072902)
```

ii. מבחינת אחוז המדגמים בהם הרב"ס תופס את הפרמטר האמיתי נעדיף גם את השיטה מסעיף (*b*), כיוון שרמת הביטחון לפי השיטה השניה נמוכה יותר, בשל בשימוש בקירוב נורמלי ובאומד לשונות.

iii. אני אעדיף את השיטה הראשונה, מסעיף (*b*).

(7)

(*a*)

הרב״ס אינו יחיד, להלן דוגמא (מויקיפדיה):

נניח $X\sim U(0,\theta)$, ו-θ אינו ידוע. לכל $0<t<0.05$ מתקיים: $P(t\theta<X<(0.95+t)\theta)=0.95$ ולפיכך $P\left(\theta\in\left[\frac{X}{0.95+t},\frac{X}{t}\right]\right)=0.95$, וזה מתקיים לכל t בתחום הנ״ל, וכל רווחי הסמך שיתקבלו יהיו בעלי אותה רמת סמך. כיוון שכל t מגדיר לנו אומד אחר, הרי לנו אינסוף אומדים המבטיחים את אותה ר״ס ל-θ.

(*b*)

לדעתי ישנו שימוש שונה לאומדן טווח ברמת סמך אל מול אומד נקודתי. אם נרצה לבדוק השערה כלשהי (אומדן לפרמטר) לפי רב״ס, נוכל להגיד על אותה השערה שהיא נכונה/לא נכונה באחוז רמת הסמך של הרב״ס, וכלי זה שימושי לחוקר, בייחוד אם הנתונים אינם שלמים. לעומת זאת, אם נרצה להגיד משהו מדוייק יותר (בגבולות דיוק האומדן), נשתמש באומדן נקודתי – כך למשל לאמר על השערה מסויימת אם היא מתקיימת או לא מתקיימת לפי אותו אומדן. לפיכך, לא נראה כי יש ערך מוסף כבד לציון גם רמת סמך לפי רב״ס כלשהו וגם אומדן נקודתי (ונעדיף אחד על פני השני לעתים בהתאם למסר שרוצים להעביר).