תרגיל 6- אמידת נראות מקסימלית ורב"סים

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

אמידה נקודתית

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

(a)מהי פונקציית הנראות ומהו אומד הנראות המקסימלית ל λ במקרה כללי של זמני המתנה $(t_1,\ldots,t_n)$?

(b)נגדיר את *אמינות האוטובוס* בתור ה"סיכוי שיגיע תוך פחות מעשר דקות".
מצא אומד נראות מקסימלית לאמינות של קו 25 בתחנת האוניברסיטה עבור מדגם כללי $(t_{1,}..,t_n)$.

2.מהו **האומדן** לסיכוי מסעיף קודם אם זמני ההמתנה השבוע היו: 12,5,1,4,4,4,7אומדי נראות מקסימלית באוכלוסייה נורמלית:

i.(*) הראו ש $\frac{1}{n}\sum_{i=1}^{n}\left(X_i-\bar{X}\right)^2$ הוא אומד נראות מקסימלית **לשונות** של אוכלוסייה נורמלית המבוסס על מדגם של n תצפיות בלתי תלויות?
המתכון לפתרון:

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

(2)כיתבו את פונקציית הנראות של n תצפיות בלתי תלויות מאוכלוסייה זו.

(3)שימו לב שזו פונקצייה של **שני** משתנים- השונות והתוחלת.
מיצאו את השונות (σ^2) שממקסמת את הנראות על ידי מעבר לסקאלה לוגריתמית, גזירה חלקית לפי השונות והשוואה לאפס.

ii.האם האמירות הבאות נכונות לגבי אומד הנראות המקסימלית **לשונות** באוכלוסייה נורמלית:

(1)האומד חסר הטייה.

(2)האומד מוטה כלפי מעלה כלומר נוטה להחזיר ערכים **גדולים** מהערך האמיתי.

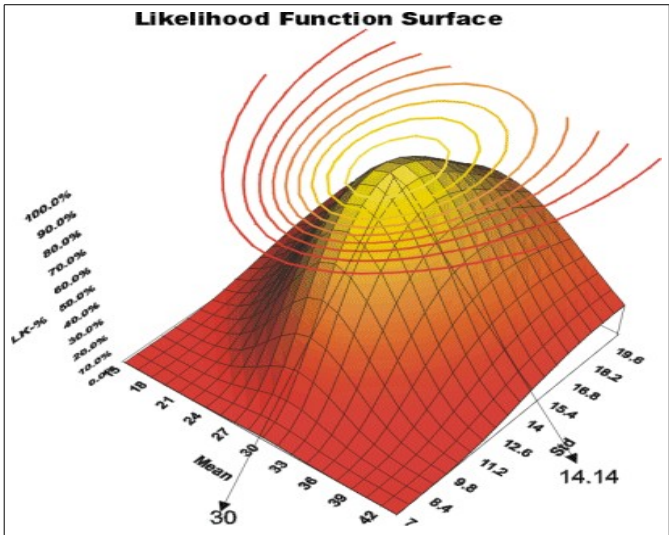
(3)האומד מוטה כלפי מטה כלומר נוטה להחזיר ערכים **קטנים** מהערך האמיתי.

(4)ניתן לראות את האומד כחישוב של שונות אם רק נסתכל על המדגם כאוכלוסייה.

(5)ניתן להסתכל על האומד כ"הסטייה הריבועית הממוצעת מהממוצע".

(6)ניתן להסתכל על האומד כ"הסטייה הריבועית הממוצעת מהתוחלת".

iii.(*) ודאו שהשונות והתוחלת שהתקבלו הן אכן נקודות מקסימום של (לוג) פונקציית הנראות.

מצורף להנאתכם תרשים יפה שמצאתי באינטרנט של פונקציית נראות נורמלית של מדגם כלשהו כתלות בתוחלת ובשונות.	

3. נמשיך להניח שזמן המתנה לאוטובוס מתפלג מעריכית על פני הגעות $(T_i \sim \exp(\lambda))$ אבל הפעם אני מעוניין לאמוד את **תוחלת** זמן ההמתנה שלי על בסיס n המתנות בלתי תלויות לאוטובוס. אומד מתבקש הוא כמובן זמן ההמתנה **הממוצע**.

(a) הם זהו אומד נראות מקסימלית? (אפשר ללא חישוב)

(b) האם הוא חסר הטייה? (ללא חישוב)

(c) האם הוא עקיב? (ללא חישוב)

(d)(*) כיצד מתפלג **סכום** זמני ההמתנה? כיצד מתפלג זמן ההמתנה הממוצע?

(e) הוכח שהאומד חסר הטייה, כלומר ש $E(\bar{T})=E(T_i)=1/\lambda$.

(f) מהי שונות האומד לתוחלת כפונקציה של זמן ההגעה הטיפוסי? $Var(\bar{T})=f[E(T)]$.

4. שיעור אתרי האינטרנט שאינם תומכים בפיירפוקס הוא p (באוכלוסייה).
אבנר גלש כל פעם לאתר מקרי, ועצר באתר החמישי כי האתר לא תמך בדפדפן שלו (פיירפוקס כמובן).

(a) חשבו את פונקציית הנראות (הגאומטרית) וציירו גרף שלה כפונקציה של הפרמטר $p \in [0,1]$. $L(p; x_1=0, x_2=0, \ldots, x_5=1)$

(b) מהו ערך הפרמטר p הכי סביר לאור התוצאה (כלומר מהו אומד הנראות המקסימלית ל p)?

(c) מצא אומד נראות מקסימלית לשיעור האתרים שאינם תומכים בפיירפוקס עבור מקרה כללי שבו בשיטוט מקרי באינטרנט האתר הראשון שאינו תומך הוא האתר ה k?

(d)(*) מצאו את התוחלת של האומד ל p בשביל להוכיח שהוא מוטה.

רווח בר סמך (Confidence Interval)

5. גובה האוכלוסיה בישראל ניתן לקירוב נורמלי. נניח שההבדל בין ישראל והעולם הוא רק בגובה הטיפוסי (התוחלת) ופיזור הגבהים במדינות שונות הוא שווה (השונות).
כמו כן נניח שניתן לגלוש לאתר האו"ם באינטרנט ולמצוא אומדן לשונות של הגבהים במדינות השונות $(\sigma=12)$.

(a) אבנר מעוניין לבנות טווח שיבטיח לו שב 90% מהמדגמים ייתפוס את התוחלת אותה אין הוא יודע. כיצד כדאי לו לאמוד את גבולות הטווח?

(b) אבנר יוצא לרחוב ושואל מדגם מקרי של אנשים לגובהים. הגבהים שאסף:
$(x_1=175, x_2=190, x_3=159, x_4=164, x_5=182, x_6=177)$.
מה יהיה הטווח של אבנר?

(c) האם אבנר יכול לאמר ש"הסיכוי שהתוחלת האמיתית נמצאת בטווח שחישב הוא 90%"?

(d) לאחר שקרא מאמר בנושא, החליט אבנר שאולי שונות הגבהים בישראל שונה משאר העולם ולכן אינו יכול להישען על נתוני האו"ם ועליו לאמוד גם את השונות. כיצד יאמוד את גבולות הטווח הפעם?

(e) מה יהיה הטווח של אבנר הפעם?

(f) האם הטווח גדל או קטן כתוצאה מהיעדר מידע על השונות של הגבהים? האם זה מפתיע?

(g) האם אבנר יכול לדרוש טווח שיתפוס את התוחלת האמיתית ב 100% מהמדגמים?

(h) אבנר מעוניין להעריך את פרופורציית תושבי ישראל שגבוהים משני מטרים. למעשה הוא רוצה לבנות טווח שיתפוס את הפרופורציה האמיתית (שהיא כמובן פונקציה של הפרמטרים) ב 90% מהמדגמים.
באילו אומדים ישתמש לגבולות הטווח אם ישתמש בשונות ששלף מאתר האו"ם (כלומר השונות ידועה)?
ואם ישתמש באומד משלו לשונות (כלומר השונות לא ידועה)?

(i) מה יהיה האומדן לטווח במדגם שלו?

(j) אחרי שראה כי הטווח מסעיף b רחב מידי לטעמו, כלומר אינו מספיק אינפורמטיבי לגבי התוחלת אותה הוא מנסה לאמוד, החליט לדרוש שאורך הטווח **לא יעלה** על שני סנטימטרים. כמה אנשים יצטרך לדגום בשביל שטווח קצר משני סנטימרים ייתפוס את התוחלת האמיתית ב 90% מהמדגמים?

(k) חזור על סעיפים (b),(e) כאשר אבנר דורש 95% בטחון במקום 90%. האם הרווחים גדלו או קטנו? האם זה מפתיע? נמק.

6. מטרת שאלה זו היא להדגים את מושג הרב"ס באמצעות סימולציה:
בהמשך לבעייה מתרגיל הכיתה... נניח כי הגבהים במדינה ניתנים לקירוב על ידי $X_i \sim N(\mu=175, \sigma=12)$.

(a) מהו הסיכוי שהאדם הבא שנפגוש ברחוב יהיה גבוה ממטר תשעים?

(חשבו אנליטית).

(b)בצעו סימולציה לבדיקת אחוז המדגמים בהם הרב"ס תופס את הפרמטר האמיתי (הסיכוי להיות גבוה משני מטרים במקרה זה):

i.הגרילו אלף מדגמים מקריים בגודל 20 מהתפלגות זו באמצעות הפקודה $rnorm(20,175,12)$.

ii.חשבו רב"ס לתוחלת (תניחו שהשונות ידועה) בכל אחד מהמדגמים ברמת סמך של 90%.

iii.על בסיס הרב"ס לתוחלת, חשבו רב"ס לסיכוי שאדם יהיה גבוה ממטר תשעים בכל אחד מהמדגמים.

iv.חשבו באיזה אחוז מאלף המדגמים הרב"ס מכיל את הערך האמיתי של הפרמטר (כלומר את הסיכוי שחישבתם בסעיף (a)).

v.מה היה אחוז המדגמים שמכילים את הסיכוי האמיתי אם היו בידינו אינסוף מדגמים בגודל עשרים ולא רק אלף.

(c)בסעיף זה נשווה בין רב"ס לפרופורציה באוכלוסייה (p) המבוסס על הנחת הנורמליות של הנתונים[1] אל מול רב"ס ללא הנחה זו.

i.חיזרו למדגמים מהסעיף הקודם ובכל אחד בנו רב"ס בבטחון של 90% לפרופורציית התושבים מעל מטר תשעים. הנוסחה היא כזכור $\hat{p}\pm\sqrt{\frac{\hat{p}(1-\hat{p})}{n}}\cdot Z_{1-\alpha/2}$.

ii.חשבו באיזה אחוז מהמדגמים הרב"ס מכיל את הערך האמיתי של הפרמטר? מה היה אחוז זה אם היו בידינו אינסוף מדגמים?

(d)בסעיפים (b) ו (c) בניתם רב"סים לאותו הפרמטר בדרכים שונות. האם ישנה עדיפות לשיטה אחת על פני השנייה מבחינת:

i.אורך הרב"ס (הרי רווחים קטנים יותר הם יותר אינפורמטיביים)?
ii.אחוז המדגמים בהם הרב"ס יתפוס את הפרמטר האמיתי?
iii.איזה מהשיטות תעדיפו?

(e)(*) האם תעדיפו שיטה זו בכל קובץ נתונים או שהעדפתכם תלוייה באחת ההנחות?

7.קצת עבודה על האינטואיציה שמאחורי רב"ס...

(a)רב"ס לפרמטר θ ברמת סמך של $1-\alpha$ מוגדר בתור רווח מקרי $\left[T_1(\vec{X})\ ,\ T_2(\vec{X})\right]$ שמקיים: $P\left(\theta\in\left[T_1(\vec{X})\ ,\ T_2(\vec{X})\right]\right)=1-\alpha$.

i.האם הרב"ס יחיד? כלומר האם ישנה יותר מדרך אחת לבחור את האומדים בגבולות הרווח כך שיקיימו את תכונות רווח הסמך?

(b)על פניו, אמידה ברווח (רב"ס) מוסיפה מידע על מידת הבטחון שלנו באומדן, שחסרה באומדי נקודה (Point Estimates) כמו א.נ.מ. ואומד מומנטים. אבל האם אנו מאבדים מידע על ידי אמידה ברווח? כלומר האם כחוקר מוטב לדווח רק על רווח הסמך או גם

1 הנחה זו מן הסתם נכונה במקרה זה, כי הסימולציה בנוייה כך שאנו דוגמים מהתפלגות נורמלית.

מבוא לסטטיסטיקה למדמ"ח- תשס"ט

את הרווח וגם את האומדן הנקודתי?