

סטטיסטיקה למדעי המחשב – פתרון תרגיל 7

1.

א. זוהי פונקציה חיובית בכל תחומה. לכן, כדי לבדוק האם זו פונקציית צפיפות לגיטימית, נבדוק באמצעות אינטגרל שהשטח מתחת לגרף אכן שווה ל-1.

$$\int_0^1 f_x(x, \theta) dx = \int_0^1 2\theta x + 1 - \theta dx = \theta x^2 + 1 - \theta \Big|_0^1 = \theta + 1 - \theta - 0 = 1$$

השטח מתחת לגרף אכן שווה ל-1 ולכן זו פונקציית צפיפות לגיטימית.

ב. יהיו ההשערות $H_0: \theta = 0$; $H_1: \theta = 1$. ע"פ הגרף של פונקציית הצפיפות בהצבת הפרמטר של כל השערה, נראה כי ערכים גדולים (קרובים ל-1) יגרמו לשלילת H_0 וערכים קטנים (קרובים ל-0) יגרמו לחיזוק התמיכה בנכונות H_0 . לכן, סטטיסטי המבחן יהיה תוצאת המדגם בהתפלגות הנתונה (x) . אזור הדחייה יהיה מהתבנית $R = [c, 1]$ כאשר c הוא מספר כלשהו בקטע $[0, 1]$ בהתאם למובהקות שנרצה.

ג. נמצא את סטטיסטי המבחן ברמת מובהקות α :

$$P_{H_0}(x \in R) = \alpha \Rightarrow \int_c^1 2\theta x + 1 - \theta dx = \int_c^1 1 dx = x \Big|_c^1 = 1 - c = \alpha \Rightarrow c = 1 - \alpha$$

סה"כ קיבלנו כי $R = [1 - \alpha, 1]$

ד.

i. במקרה זה גרף פונקציית הצפיפות תחת H_0 לא ישתנה, אך הגרף של פונקציית הצפיפות תחת H_1 יהפוך תלול יותר. לכן, אזור הדחייה יגדל (כי ההבדלים בין ההנחות יגדלו ולכן יהיו יותר ערכים שפחות סבירים עבור H_0).

ii. במקרה זה גרף פונקציית הצפיפות תחת H_0 יהפוך משופע בדומה לגרף של פונקציית הצפיפות תחת H_1 . H_1 יישאר תלול יותר, אך ההבדלים בין השניים יצטמצמו. לכן, אזור הדחייה יקטן (כי יהיו פחות ערכים שפחות סבירים עבור H_0).

2. $X_i \sim \exp(\lambda)$ – זמן הגעה של טכנאי יחיד. $T_j = \sum_{i=1}^{30} X_i \sim N\left(\frac{30}{\lambda}, \frac{30}{\lambda^2}\right)$ – סיכוי שבעיה תיפתר.

א. נציב $\lambda = \frac{1}{8}$ ונחשב:

$$P(T_j < 200) \approx \Phi\left(\frac{200 - \frac{30}{\lambda}}{\sqrt{\frac{30}{\lambda^2}}}\right) = \Phi\left(\frac{200 - 240}{\sqrt{1920}}\right) = \Phi(-0.91287) = \mathbf{0.18}$$

ב. ההנחה אינה ריאלית כיוון שלדוגמה, זמן ההגעה של הטכנאים צפוי להיות זהה (כתלות במיקום של התקלה אל מול המרכז הלוגיסטי) ולכן אין אי תלות אמיתית בין המ"מ.

ג. נתון מדגם זמני תיקון בלתי תלויים. מכאן שזמן התיקון הממוצע הוא

$\bar{t} \sim N\left(\frac{30}{\lambda}, \frac{30}{m\lambda^2}\right)$. לכן, רווח הסמך לזמן התיקון הטיפוסי הוא $\bar{t} \pm \sqrt{\frac{30}{m\lambda^2}} Z_{1-\frac{\alpha}{2}}$. לכן,

רווח הסמך לזמן ההגעה הטיפוסי הוא $\frac{1}{30}(\bar{t} \pm \sqrt{\frac{30}{m\lambda^2}} Z_{1-\frac{\alpha}{2}})$. כיוון שהפרמטר λ

אינו ידוע לנו, אך נוכל לאמוד אותו באמצעות \bar{t} כך $\bar{t} = \frac{30}{\lambda} \Leftrightarrow \frac{\bar{t}}{30} = \frac{1}{\lambda}$ ולכן רווח

הסמך יהיה $\frac{1}{30}(\bar{t} \pm \sqrt{\frac{\bar{t}^2}{30m}} Z_{1-\frac{\alpha}{2}})$

ד. אנו רוצים למצוא את רווח הסמך בביטחון של 90% בהינתן שהזמן הממוצע לתיקון 20 תקלות היה 220 שעות. ע"פ הנתונים: $1 - \alpha = 0.9 \Rightarrow \alpha = 0.1$. נציב את כל הנתונים ברווח הסמך ונקבל:

$$\left[\frac{1}{30} \left(220 - \sqrt{\frac{220^2}{30 \cdot 20}} Z_{0.95} \right), \frac{1}{30} \left(220 + \sqrt{\frac{220^2}{30 \cdot 20}} Z_{0.95} \right) \right] = [6.84, 7.82]$$

ה. נרצה לבחון את ההשערות $H_0: \frac{1}{\lambda} = 8$; $H_1: \frac{1}{\lambda} = 7$

- i. היות ו- $\frac{1}{\lambda}$ היא תוחלת ההתפלגות, נוכל לבחור כסטטיסטי המבחן להיות ממוצע המדגם \bar{t} כיוון שכידוע הממוצע מתכנס לתוחלת.
- ii. ככל שזמן התיקון יהיה יותר קטן, הסיכוי להשערה האלטרנטיבית גדל. לכן כיוון הדחייה יהיה כלפי ערכים קטנים. מכאן שאזור הדחייה יהיה $R = [-\infty, c]$

iii. נחשב:

$$P_{H_0}(\bar{t} \in R) = \alpha = P_{H_0}(\bar{t} < c) = \Phi\left(\frac{c - \frac{30}{\lambda}}{\sqrt{\frac{30}{m\lambda^2}}}\right) \Rightarrow Z\alpha = \frac{c - \frac{30}{\lambda}}{\sqrt{\frac{30}{m\lambda^2}}} \Rightarrow c = Z\alpha \sqrt{\frac{30}{m\lambda^2}} + \frac{30}{\lambda}$$

אנו נמצאים תחת השערת H_0 ולכן נוכל להציב $\frac{1}{\lambda} = 8$ ונקבל כי סף אזור הדחייה הוא $R = [-\infty, Z\alpha\sqrt{1920m} + 240]$

3.

- א. $X \sim Pois(\lambda)$ - מספר השאלות בשנייה. נגדיר את ההשערות כך:
 $H_0: \lambda = 2$; $H_1: \lambda = 1$
- ב. אם נחשוב על מקרה קצה בו לא בוצעו כלל שאלות - מקרים אלו יותר סבירים תחת ההנחה H_1 . לכן, נדחה עבור ערכים קטנים.
- ג. אם מרכז החישובים צודק, $X \sim Pois(2)$. היות ומספר השאלות בכל שנייה הוא ב"ת, סכום השאלות ב-4 שניות יתפלג פואסונית עם פרמטר 8 (נסמן $X4 \sim Pois(8)$). אם החשב צודק, באופן דומה סכום השאלות ב-4 שניות יתפלג פואסונית עם פרמטר 4.
- ד. אזור הדחייה מהתבנית $R = [0, c]$. אנחנו בעצם מחפשים מצב במו מתקיים:
 $P(X4 < c) < \alpha \Rightarrow \sum_{i=0}^c P(X4 = i) = \sum_{i=0}^c \frac{e^{-8} 8^i}{i!} < 0.05$
נבצע הצבות לערכי c אפשריים ונקבל כי עבור $c=4$ הסכום עובר לראשונה את המגבלה שהצבנו. לכן, ערך ה- c המבוקש הוא $c=3$. כלומר, $R = [0, 3]$.
- ה. בהמשך לסעיף הקודם, במקרה זה, אזור הדחייה יהיה $R = [0, \frac{3}{4}]$.

1.

- i. איזור הדחייה יהיה עבור ערכים קטנים $R = [0, c]$
- ii. איזור הדחייה יהיה עבור ערכים גדולים $R = [c, \infty]$
- iii. איזור הדחייה יהיה עבור ערכים קטנים $R = [0, c]$

- iv. איזור הדחייה יהיה עבור ערכים גדולים מאוד או קטנים מאוד (ביחס ל-2)
 $R = [0, c_1] \cup [c_2, \infty]$
 v. איזור הדחייה יהיה עבור ערכים קטנים $R = [0, c]$
 vi. איזור הדחייה יהיה עבור ערכים קטנים $R = [0, c]$

ז. סכום של n מ"מ מקריים המתפלגים פואסונית מתפלג פואסונית כך שהפרמטר מוכפל ב- n . לכן תוחלת ושונות הסכום יהיו $n\lambda$. מכאן, שאם ערך n גדול מספיק, נוכל להפעיל את משפט הגבול המרכזי ולקרב את הסכום $T \sim N(n\lambda, n\lambda)$

ח.

$$P(T < c) = 0.05 \Rightarrow P\left(\frac{T - 200}{\sqrt{200}} < \frac{c - 200}{\sqrt{200}}\right) = \Phi\left(\frac{c - 200}{\sqrt{200}}\right) = 0.05$$

$$\Rightarrow \frac{c - 200}{\sqrt{200}} = Z_{0.05} \Rightarrow c = \sqrt{200} \cdot Z_{0.05} + 200 = \mathbf{176.736}$$

ט. בהנחה שהמדגם מספיק גדול, נוכל לקרב את ההתפלגות להתפלגות נורמאלית $\bar{x} \sim N\left(\lambda, \frac{\lambda}{n}\right)$ ולכן רווח הסמך יהיה: $\bar{x} \pm \sqrt{\frac{\lambda}{n}} \cdot Z_{1-\frac{\alpha}{2}}$. היות והממוצע מתכנס לתוחלת, נוכל להשתמש ב- \bar{x} גם עבור חישוב השונות ולכן רווח הסמך יהיה $\bar{x} \pm \sqrt{\frac{\bar{x}}{n}} \cdot Z_{1-\frac{\alpha}{2}}$

י. נציב את הנתונים ונקבל

$$1.54 \pm \sqrt{\frac{1.54}{100}} \cdot Z_{0.975} = 1.54 \pm 0.124 \cdot 1.96 = 1.54 \pm 0.243 \Rightarrow \mathbf{[1.296, 1.783]}$$

4.

א. נשתמש ברווח סמך שמרני לפרופורציה P ונקבל $\hat{p} \pm Z_{0.975} \frac{0.5}{\sqrt{n}}$. כדי שטעות של למעלה מאחוז תקרה רק ב-5% מהמדגמים, התנאי $Z_{0.975} \frac{0.5}{\sqrt{n}} < 0.01$ צריך להתקיים. לכן, $Z_{0.975} \cdot \frac{0.5}{0.01} < \sqrt{n} \Rightarrow n > \mathbf{9604}$

ב. במצב כזה, נוכל לחסוך את הערך של $p(1-p)$ ע"י 0.16. לכן נקבל $Z_{0.975} \cdot \frac{0.4}{0.01} < \sqrt{n} \Rightarrow n \geq \mathbf{6147}$. גודל המדגם קטן. אין זה מפתיע כיוון שהורדנו חלק מחוסר הוודאות ולכן מספיק מדגם קטן יותר.

ג. זהו למעשה אותה שאלה כמו הסעיף הראשון כיוון שרווח הסמך ברוב 2 מבטיח לנו שלא "נחטיא" את הפרמטר האמיתי ביותר מאחוז עבור לפחות 95% מהמדגמים. כלומר, אחוז המדגמים אשר יחטיא את הפרמטר האמיתי לא יעלה על 5%.

ד. כבר עשיתי את זה בסעיף הקודם.

ה. נבצע את החישובים באמצעות הנתונים:

$$\text{רב"ס רגיל: } \hat{p} \pm Z_{0.975} \frac{\sqrt{\hat{p}(1-\hat{p})}}{\sqrt{n}} = 0.09 \pm \frac{\sqrt{0.0819}}{10} \cdot 1.96 \Rightarrow \mathbf{[0.034, 0.146]}$$

$$\text{רב"ס שמרני (ללא הנחה): } \hat{p} \pm Z_{0.975} \frac{0.5}{\sqrt{n}} = 0.09 \pm \frac{0.5}{10} \cdot 1.96 \Rightarrow \mathbf{[0, 0.188]}$$

$$\text{רב"ס שמרני (עם הנחה): } \hat{p} \pm Z_{0.975} \frac{0.4}{\sqrt{n}} = 0.09 \pm \frac{0.4}{10} \cdot 1.96 \Rightarrow \mathbf{[0.0116, 0.1684]}$$

5. בקטע "אמת וסקר" מתוארים מספר סוגי הטיות בעת עריכת סקרים. ההטיות המצוינות בקטע הן:

- א. **הטיית הברירה למדגם (selection bias)** – הטייה זו נוצרת בעת בחירת המשיבים לסקר. במקרה המתואר בקטע, ניתנות דוגמאות בהן דרכי בחירת המשיבים היא מוטה. בין הדוגמאות: מערכת העיתון Literary Digest אשר ביצע משאל בקרב 2 מליון קוראי העיתון ו"מדגם" מעריב שביצע תשאל בקלפיות שהוצבו בתחנות דלק. הבעיה בדוגמאות שצוינו היא בכך שהנשאלים במדגם נלקחו מקבוצה בעלת מאפיינים ייחודיים (רפובליקנים מבוססים בדוגמה הראשונה, ובעלי מכוניות אמידים בדוגמה השנייה). הגבלת הקבוצה אותה דוגמים בצורה כזו עשויה לגרום לתוצאות שגויות. הגדלת כמות הנדגמים במקרה כזה אינה פותרת את הבעיה.
- ב. **הטיית בעת עירוב שיקול דעת** – הטייה זו נוצרת כאשר אין אקראיות מכוונת בעת עריכת הסקר. בדוגמה שהובאה בקטע ניסו עורכי הסקר לייצג באופן אמין את הרכב האוכלוסייה. עם זאת, למרות ההגבלות הרבות שחלו על הסוקרים בעת בחירת המשיבים, עדיין היה קיים לכל סוקר שיקול דעת בבחירה. התוצאה שהתקבלה היא שהסוקרים העדיפו משיבים אשר קל לתשאל ובכך נגרמו תוצאות שגויות בסקר. לכן, בעת דגימה יש להכניס מרכיב הסתברותי המאפשר להעריך את הגודל האופייני של השגיאה.
- ג. **הטייה מצורת השאלה** – בקטע מודגם כיצד ניסוח שונה של אותה שאלה יכול לגרום לתוצאות שונות. בדוגמה שהוצגה, נערכו 3 סקרים אשר בדקו את אותה השאלה. למרות שבבסיס השאלה הייתה זהה, רק סקר אחד הצליח לחזות באופן אמין את תוצאות האמת. הסיבה לכך הייתה שבשניים מתוך שלושת הסקרים נוסחה השאלה בצורה שנתנה יתרון לצד אחד על פני אחר. בשל כך, חשוב לתת את המידע על צורת עריכת הסקר והשאלות עם תוצאותיו.
- ד. **אי הבנת תוצאות הסקר (או הצגתם בצורה שגויה)** – במקרה כזה מוצגות תוצאות הסקר בצורה מטעה או מוצגות מסקנות שגויות מהסקר כתוצאה מאי הבנת מושגים ודרכי עריכת סקרים. מצב כזה יכול לגרום להליך קבלת החלטות שגוי כאשר חוסר ההבנה הוא גדול וגורם להבנה הפוכה (או חלקית) של הנתונים.