

מבוא להסתברות וסטטיסטיקה

אריאל חורי ויונתן סמידוברסקי (תשפ"ב)

3 באוגוסט 2022

מבוסס על מערכי תרגול של מבוא להסתברות וסטטיסטיקה (88165) ותרגילים מהקורס.

מבוא לסטטיסטיקה

בהסתברות, ניסינו לחזות תכונות של דגימה מקרית על סמך התפלגות ומודל תיאורטי. לעומת זאת, בסטטיסטיקה נרצה להיעזר במדגם נתון שהתקבל מניסוי על מנת להעריך או לאמוד את ההתפלגות שיצרה אותו.

הפרדת השערות

הגדרה 1. H_0 - השערת האפס (לרוב מייצגת את ברירת המחדל או המצב הקיים)

הגדרה 2. H_1 - השערה אלטרנטיבית

מטרתנו היא לדעת האם לקבל את H_0 או לדחות אותה ביחס ל H_1 .

הגדרה 3. השערה פשוטה היא השערה שקובעת את התפלגותו של המשתנה המקרי בניגוד להשערה מורכבת שלפיה התפלגות המשתנה המקרי שייכת למשפחה של התפלגויות.

הגדרה 4. מבחן הוא הגדרה של מאורע $A \subseteq \Omega$ כך שאם $\omega \in A$, מקבלים את H_0 , ואחרת דוחים את H_0 ומקבלים את H_1 .

סוגי טעויות

נאמר כי טעינו

הגדרה 5. טעות מסוג ראשון - דחייה של H_0 , בהינתן ש H_0 הוא הנכון. נסמן את ההסתברות לטעות שכזו (רמת המובהקות)

$$\alpha := P_{H_0}(\neg H_0)$$

הגדרה 6. טעות מסוג שני - קבלה של H_0 בהינתן שהיא אינה נכונה. נסמן את ההסתברות לטעות שכזו:

$$\beta := P_{H_1}(H_0)$$

נגדיר גם את $1 - \beta$ להיות העוצמה של המבחן.

הלמה של ניימן-פירסון

הגדרה 7. נניח שאנחנו מפרידים בין שתי השערות פשוטות

$$H_0 : P = P_0, H_1 : P = P_1$$

יחס הנראות מוגדר ע"י

$$\lambda(x_1, \dots, x_n) = \frac{P_1(x_1, \dots, x_n)}{P_0(x_1, \dots, x_n)}$$

כאשר x_1, \dots, x_n הן תוצאות הניסוי, כאשר מדובר על התפלגויות בדידות, P_1, P_0 יהיו הסיכויים לקבלת סדרת התוצאות המדויקות. בהתפלגויות רציפות, נשתמש בפונקציית הצפיפות (המשותפת). ככל שיחס הנראות גדול יותר כך הסיכוי שהתוצאה התקבלה מ- H_1 גדול יותר מאשר מ- H_0 .

הגדרה 8. מבחן ניימן פירסון הוא מבחן מהצורה $\lambda(x_1, \dots, x_n) \leq K$ עבור K קבוע. כלומר, משווים את יחס הנראות לקבוע K ; אם הוא נמוך יותר, מקבלים את H_0 , ואחרת דוחים את H_0 . על ידי שינוי הערך של K ניתן לקבל רמות מובהקות שונות.

משפט 9. הלמה של ניימן-פירסון

בהפרדה בין השערות פשוטות עם רמת מובהקות נתונה α , מבחן ניימן-פירסון הוא המבחן בעל העוצמה המקסימלית. איך נתכנן ניסוי המבוסס על מבחן ניימן-פירסון?

• נבחר את רמת המובהקות שבה אנו רוצים לעבוד

• נחשב את הסף K_α שעבורו מבחן ניימן-פירסון המתאים יהיה בעל רמת מובהקות α באופן תיאורטי (על סמך H_0)

• נחשב את יחס הנראות של הדגימות הנתונות, ונשווה אותו ל- K_α .

אמידה

נניח שנתון מדגם שיוצר מתוך התפלגות ידועה אך הפרמטרים אינם ידועים, נרצה להעריך אותם על סמך המדגם.

הגדרה 10. סטטיסטי - פונקציה של המדגם שאינה תלויה בפרמטר.

הגדרה 11. אומדן - סטטיסטי שערכו אמור להיות קרוב לפרמטר המבוקש.

הגדרה 12. אומדן - ערך של האומדן (לאחר הצבת ערכי המדגם).

הגדרה 13. אומדן חסר הטייה - נניח $X_1, \dots, X_n \sim F_\theta$, כאשר θ הוא פרמטר הקובע את ההתפלגות.

אומרים שאומדן $T = t(X_1, \dots, X_n)$ הוא אומדן חסר הטייה של θ , אם לכל θ מתקיים $\mathbb{E}[T | X_1, \dots, X_n \sim F_\theta] = \theta$ בכל התפלגות,

X_1 הוא חסר הטייה לתוחלת.

גם $\bar{X} = \frac{1}{n} \sum_{i=1}^n X_i$ הוא חסר הטייה לתוחלת.

תרגילים

1. בהינתן מטבע, נרצה לקבוע האם הוא מטבע הוגן או מוטה (הסיכוי ליפול על עץ הוא 0.6). נגדיר מבחן לדוגמה: נטיל את המטבע 1,000 פעמים, ואם נקבל פחות מ-550 עץ, נקבע שהוא הוגן. אחרת, נקבע שהוא מוטה. חשב את רמת המובהקות והעוצמה של המבחן.
2. בחנות נמצאת קובייה שהמוכר טוען שהיא הוגנת. אתם חושבים שהטענה שלו אינה נכונה, והסיכוי שלה ליפול על 6 הוא $\frac{1}{4}$ (ושאר התוצאות בסיכוי זהה). לשם כך, החלטתם להטיל את הקובייה 300 פעמים, ואם תקבלו יותר מ-65 פעמים את התוצאה 6 - תגיעו למסקנה שהוא טועה. מהי רמת המובהקות של המבחן שהוצע, ומהי העוצמה שלו?
3. מגרילים משתנה מקרי נורמלי X . דנית טוענת כי $X \sim N(0, 100)$, ואילו דן טוען כי $X \sim N(0, 10000)$. הם בחרו לבדוק את השערתם על ידי מבחן מהצורה $X^2 \leq C$, כאשר C קבוע כלשהו. אם $X^2 \leq C$, יקבלו את השערה של דנית, ואחרת ידחו אותה. מצאו קבוע C כך שרמת המובהקות של המבחן תהיה 5%. מהי העוצמה של המבחן?
4. במפעל לייצור נורות טוענים שמשך החיים של נורות שהם מייצרים היא מעריכית עם תוחלת 200 שעות. עיתונאי חוקר שמע מהמקורות שלו שתוחלת החיים היא דווקא 140 שעות. הוא נזכר בקורס בסטטיסטיקה שלמד, והחליט לקנות 100 נורות (בלתי-תלויות זו בזו) ולמדוד את משך החיים שלהן. הוא קיבל שבממוצע הן שרדו 160 שעות. כתבו מבחן ניימן-פירסון שיבדוק האם טענת החברה נכונה, האם ניתן לדחות את טענתה ברמת מובהקות 5%? מהי העוצמה של המבחן שכתבתם?
5. ברצונכם לבדוק האם בחודש האחרון משך זמן הנסיעה לאוניברסיטה התארך. מנתוני העבר ידוע שמשך הזמן מתפלג נורמלית, עם תוחלת שעה וסטיית תקן של 10 דקות. אתם חושדים כי משך הזמן כעת מתפלג נורמלית עם תוחלת של שעה ו-5 דקות וסטיית תקן של 10 דקות. אספתם נתונים מ-10 נסיעות בלתי-תלויות לאוניברסיטה, וגיליתם שבממוצע נסעתם שעה ו-2 דקות. כתבו את מבחן ניימן-פירסון המתאים לרמת מובהקות 5%.
6. יהיו X_1, \dots, X_n משתנים מקריים בלתי תלויים שווי-התפלגות, בעלי תוחלת μ ושוונות σ^2 . הראו כי $T := \frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2$ אינו אומדן חסר הטייה לשונות. כיצד ניתן לתקן את האומדן T כך שיהיה חסר הטייה?
7. ידוע X_1, \dots, X_n הם משתנים מקריים בלתי-תלויים המתפלגים $Ber(p)$. לכל אחד מהבאים, קבעו האם הוא אומדן חסר הטייה עבור p :

$$X_1, X_1 X_2, \sqrt{X_1 X_2}, \frac{1}{n} \sum_{i=1}^n X_i$$

8. יהיו $X_1, \dots, X_n \sim U[0, \theta]$ משתנים מקריים בלתי תלויים מהתפלגות אחידה רציפה על $[0, \theta]$. הוכיחו כי $\frac{2}{n} \sum_{i=1}^n X_i$ הוא אומדן חסר הטייה ל- θ . מצאו קבוע a שעבורו $a \cdot \max\{X_1, \dots, X_n\}$ הוא אומדן חסר הטייה ל- θ .
9. נניח $X, Y \sim Geo(p)$. תנו דוגמה לאומדן חסר הטייה עבור $\frac{1}{p}$, $\frac{1}{p^2}$. הראו כי $\frac{1}{X}$ אינו אומדן חסר הטייה עבור p , ואילו $\frac{1}{X+Y-1}$ הוא כן אומדן חסר הטייה עבור p .

פתרונות

1. נחשב את רמת המובהקות $\alpha := P_{H_0}(-H_0)$, כלומר נניח ש H_0 נתון ונבדוק מה הסיכוי לדחות אותו. נגדיר משתנה מקרי X הסופר את כמות העצים שיצאו. $X|H_0 \sim Bin(1000, 1/2)$, ונרצה לדעת את ההסתברות $P(X \geq 550)$, ניעזר במשפט הגבול המרכזי ובתיקון רציפות, נגדיר $Z \sim N(500, 250)$

$$P(X \geq 550) \approx P(Z \geq 549.5) = P\left(\frac{Z - 500}{\sqrt{250}} \geq \frac{549.5 - 500}{\sqrt{250}}\right) = 1 - \Phi(3.13) = 0.00087$$

קיבלו $\alpha = 0.00087$

בדומה, מחשבים את $\beta := P_{H_1}(H_0)$, כעת $X|H_1 \sim Bin(1000, 0.6)$, ניעזר במשפט הגבול המרכזי ונעזרים בתיקון רציפות, נגדיר $Y \sim N(600, 240)$

$$\beta = P(X < 550) \approx P(Y \leq 549.5) = P\left(N(0, 1) \leq \frac{549.5 - 600}{\sqrt{240}}\right) = \Phi(-3.26) = 0.00056$$

ולכן העוצמה של המבחן היא $1 - \beta = 0.99944$

2. נגדיר את השערת האפס להיות שהקוביה הוגנת וההשערה האלטרנטיבית להיות שהקובייה מוטה (עם הסיכויים המתוארים)
נגדיר X_6 להיות כמות הפעמים שיצא 6 בקובייה, נחשב:

חישוב α

$X|H_0 \sim Bin(300, 1/6)$ וניעזר במשפט הגבול המרכזי לקרב את ההסתברות

$$\begin{aligned}\alpha &= P(X > 65) \approx P(N(50, 250/6) \geq 65.5) = P\left(\frac{N(50, 250/6) - 50}{\sqrt{250/6}} \geq \frac{65.5 - 50}{\sqrt{250/6}}\right) \\ &= P(N(0, 1) > 2.4) = 1 - \Phi(2.4)\end{aligned}$$

$$\boxed{\alpha = 0.00798}$$

חישוב β

$X|H_1 \sim Bin(300, 1/4)$ וניעזר במשפט הגבול המרכזי לקרב את ההסתברות

$$\begin{aligned}\beta &= P(X \leq 65) \approx P(N(75, 225/4) \leq 64.5) = P\left(\frac{N(75, 225/4) - 75}{\sqrt{225/4}} \leq \frac{64.5 - 75}{\sqrt{225/4}}\right) \\ &= P(N(0, 1) \leq -7/5) = \Phi(-1.4) = 0.08076\end{aligned}$$

$$\boxed{1 - \beta = 0.91924}$$

3. נסמן את ההשערה של דנית להיות השערת האפס ואת של דן האלטרנטיבית. עבור המבחן $X^2 \leq C$, באופן שקול,

$$-\sqrt{C} \leq X \leq \sqrt{C}$$

נרצה רמת מובהקות $\alpha = 0.05$, כלומר

$$P_{H_0}(\neg H_0) = 1 - P(-\sqrt{C} \leq X \leq \sqrt{C})$$

ונרצה

$$0.95 = P(-\sqrt{C} \leq X \leq \sqrt{C})$$

$$0.95 = P\left(\frac{-\sqrt{C}}{10} \leq X \leq \frac{\sqrt{C}}{10}\right)$$

$$0.95 = \Phi\left(\frac{\sqrt{C}}{10}\right) - \Phi\left(\frac{-\sqrt{C}}{10}\right)$$

מסימטריות נקבל

$$0.95 = 2\left(\Phi\left(\frac{\sqrt{C}}{10}\right) - \Phi(0)\right)$$

כלומר

$$0.975 = \Phi\left(\frac{\sqrt{C}}{10}\right)$$

$$\frac{\sqrt{C}}{10} = 1.96$$

ונקבל

$$\boxed{C = 384.16}$$

כעת, נחשב את עוצמת המבחן. $X|H_1 \sim N(0, 10000)$.

$$\begin{aligned} 1 - \beta &= 1 - P_{H_1}(H_0) = 1 - P\left(-\sqrt{384.16} \leq X \leq \sqrt{384.16}\right) \\ &= 1 - P\left(-\frac{\sqrt{384.16}}{100} \leq \frac{X}{100} \leq \frac{\sqrt{384.16}}{100}\right) = 1 - \left(\Phi\left(\frac{\sqrt{384.16}}{100}\right) - \Phi\left(-\frac{\sqrt{384.16}}{100}\right)\right) \end{aligned}$$

$$\boxed{1 - \beta = 0.841}$$

4. נסמן X_1, \dots, X_{100} את תוחלת החיים של הנורות שהעיתונאי רכש, השערת האפס $X_i \sim \text{Exp}(\frac{1}{200})$ ואילו ההשערה האלטרנטיבית היא $X_i \sim \text{Exp}(\frac{1}{140})$. נחשב את יחס הנראות (בעזרת אי-תלות)

$$\begin{aligned} \lambda(x_1, \dots, x_{100}) &= \frac{P_1(x_1, \dots, x_{100})}{P_0(x_1, \dots, x_{100})} = \frac{\prod_{i=1}^{100} f_{X_i|H_1}(x_i)}{\prod_{i=1}^{100} f_{X_i|H_0}(x_i)} = \frac{\prod_{i=1}^{100} 1/140 e^{-x_i/140}}{\prod_{i=1}^{100} 1/200 e^{-x_i/200}} \\ &= \left(\frac{200}{140}\right)^{100} e^{(\frac{1}{140} - \frac{1}{200}) \sum_{i=1}^{100} x_i} = \left(\frac{10}{7}\right)^{100} e^{\frac{-3}{1400} \sum_{i=1}^{100} x_i} \end{aligned}$$

מבחן ניימן פירסון המתאים יהיה מהצורה $\left(\frac{10}{7}\right)^{100} e^{\frac{-3}{1400} \sum_{i=1}^{100} x_i} \leq C$ לקבוע C כלשהו. נקול ללקחת $\frac{1}{100} \sum_{i=1}^{100} x_i \leq C$ נמצא את המתאים שעבור

$$\alpha = P_{H_0} \left(\frac{1}{100} \sum_{i=1}^{100} x_i < C \right) \approx P_{H_0} (N(200, 400) < C) = \Phi \left(\frac{C - 200}{20} \right) = 0.05$$

נקבל $\frac{C-200}{20} = -1.645$, ולכן $C = 167.1$ ולכן נוכל לדחות את טענת החברה ברמת מובהקות 5%. נחשב את עוצמת המבחן

$$1 - \beta = P_{H_1}(H_0) = P_{H_1} \left(\frac{1}{100} \sum_{i=1}^{100} X_i \leq 167.1 \right)$$

תחת ההשערה האלטרנטיבית, כלומר $X_i \sim \text{Exp}(\frac{1}{140})$, וממשפט הגבול המרכזי

$$1 - \beta \approx P(N(140, 196) \leq 167.1) = \Phi \left(\frac{167.1 - 140}{14} \right) = 0.9735$$

$$H_0 : X_i \sim N(60, 10^2)$$

$$H_1 : X_i \sim N(65, 10^2)$$

אספנו x_1, \dots, x_{100} והממוצע 62 דקות,
נכתוב את ניימן-פירסון המתאים ל $\alpha = 0.05$

$$\begin{aligned} \lambda(x_1, \dots, x_n) &= \frac{P_1(x_1, \dots, x_n)}{P_0(x_1, \dots, x_n)} = \frac{\prod_{i=1}^{100} \frac{1}{\sqrt{2\pi} \cdot 10} e^{-\frac{(x_i-65)^2}{200}}}{\prod_{i=1}^{100} \frac{1}{\sqrt{2\pi} \cdot 10} e^{-\frac{(x_i-60)^2}{200}}} \\ &= e^{\sum_{i=1}^{100} \frac{(x_i-65)^2 - (x_i-60)^2}{200}} \end{aligned}$$

נרצה

$$e^{\sum_{i=1}^{100} \frac{(x_i-65)^2 - (x_i-60)^2}{200}} \leq K_\alpha$$

זה שקול

$$\sum_{i=1}^{100} x_i \leq K'_\alpha$$

נרצה $\alpha = 0.05$,

$$P_{H_0}(\neg H_0) = 0.05$$

$$P_{H_0} \left(\sum_{i=1}^{100} x_i > K'_\alpha \right) = 0.05$$

בהינתן $X_i|H_0 \sim N(60, 100)$,

$$P(N(60 \cdot 10, 100 \cdot 10) > K_{\alpha'}) = 0.05$$

$$P(N(600, 1000) > K_{\alpha'}) = 0.05$$

$$P\left(N(0,1) > \frac{K_{\alpha'} - 600}{\sqrt{1000}}\right) = 0.05$$

$$K'_{\alpha} = 600 + \sqrt{1000} \cdot 1.645$$

6. ראשית, נראה כי $\mathbb{E}[T] \neq \sigma^2$,

$$\begin{aligned} \sum_{i=1}^n (X_i - \bar{X})^2 &= \sum_{i=1}^n ((X_i - \mu) - (\bar{X} - \mu))^2 \\ &= \sum_{i=1}^n (X_i - \mu)^2 + 2(X_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2 \\ &= \sum_{i=1}^n (X_i - \mu)^2 + 2(X_i - \mu)(\bar{X} - \mu) + (\bar{X} - \mu)^2 \\ &= \dots = \sum_{i=1}^n (X_i - \mu)^2 - n(\bar{X} - \mu)^2 \end{aligned}$$

לכן

$$\begin{aligned} \mathbb{E} \left[\frac{1}{n} \sum_{i=1}^n (X_i - \bar{X})^2 \right] &= \frac{1}{n} \sum_{i=1}^n \mathbb{E} [(X_i - \mu)^2] - \mathbb{E} [(\bar{X} - \mu)^2] \\ &= \frac{1}{n} \sum_{i=1}^n \text{Var}(X_i) - \text{Var}(\bar{X}) = \sigma^2 - \frac{\sigma^2}{n} = \frac{n-1}{n} \sigma^2 \end{aligned}$$

אם נחשב את $S^2 = \frac{1}{n-1} \sum_{i=1}^n (X_i - \bar{X})^2$ במקום T , נקבל אומדן חסר הטייה לשונות.

• האומד X_1 הינו חסר הטייה:

$$\mathbb{E}[X_1 | X_1, \dots, X_n \sim \text{Ber}(p)] = \mathbb{E}[X_1] = p$$

• האומד $X_1 X_2$ אינו חסר הטייה:

$$\mathbb{E}[X_1 X_2 | X_1, \dots, X_n \sim \text{Ber}(p)] = p^2$$

• האומד $\sqrt{X_1 X_2}$ אינו חסר הטייה:

$$\begin{aligned} \mathbb{E}[\sqrt{X_1 X_2} | X_1, \dots, X_n \sim \text{Ber}(p)] &= \sum_{i=0}^1 \sqrt{i} \cdot P(X_1 X_2 = i) \\ &= 0 \cdot p^2 + 1 \cdot p^2 = p^2 \end{aligned}$$

• האומד $T = \frac{1}{n} \sum_{i=1}^n X_i$ הינו חסר הטייה:

$$\begin{aligned} \mathbb{E}[T | X_1, \dots, X_n \sim \text{Ber}(p)] &= \frac{1}{n} \cdot \sum_{i=1}^n \mathbb{E}[X_i | X_1, \dots, X_n \sim \text{Ber}(p)] \\ &= \frac{1}{n} \cdot \sum_{i=1}^n p = \frac{1}{n} \cdot np = p \end{aligned}$$

$$\mathbb{E}\left[\frac{2}{n} \sum_{i=1}^n X_i\right] = \frac{2}{n} \cdot \sum_{i=1}^n \mathbb{E}[X_i] = \frac{2}{n} \sum_{i=1}^n \frac{\theta}{2} = \frac{2 \cdot n\theta}{2n} = \theta$$

ובנוסף, נחשב את ההתפלגות $M = \max\{X_1, \dots, X_n\}$ לכל $0 \leq m \leq \theta$ מתקיים

$$P(M \leq m) = P(X_1 \leq m, \dots, X_n \leq m) = \prod_{i=1}^n P(X_i \leq m)$$

(מאי תלות ותכונות הסתברות אחידה)

$$= \prod_{i=1}^n \frac{m}{\theta} = \left(\frac{m}{\theta}\right)^n$$

פונקציית הצפיפות של M הינה:

$$f_M(m) = \begin{cases} \frac{n \cdot m^{n-1}}{\theta^n} & 0 < m < \theta \\ 0 & otherwise \end{cases}$$

ולכן התוחלת של M היא

$$\begin{aligned} \mathbb{E}[M] &= \int_0^\theta m \cdot \frac{n \cdot m^{n-1}}{\theta^n} dm = \frac{n}{\theta^n} \int_0^\theta m^n dm \\ &= \frac{n}{\theta^n} \left[\frac{m^{n+1}}{n+1} \right]_0^\theta = \frac{n}{\theta^n} \cdot \frac{\theta^{n+1}}{n+1} = \frac{n}{n+1} \cdot \theta \end{aligned}$$

ולכן ניקח $a = \frac{n+1}{n}$ ונקבל את הדרוש.

דוגמאות עבור $X, Y, \frac{X+Y}{2} : \frac{1}{p}$
 דוגמאות עבור $XY : \frac{1}{p^2}$ (מאית-לות), דוגמה נוספת היא $(\mathbb{E}[X^2] = \frac{2-p}{p^2}) \frac{X^2+X}{2}$
 עבור $\frac{1}{X}$ מתקיים $\mathbb{E}[\frac{1}{X}] = \sum_{n=1}^{\infty} \frac{1}{n} p \cdot (1-p)^{n-1}$ נשים לב כי $Z := X + Y \sim NB(2, p)$

$$\mathbb{E} \left[\frac{1}{Z-1} \right] = \sum_{n=2}^{\infty} \frac{1}{n-1} \binom{n-1}{1} p^2 (1-p)^{n-2} = \sum_{n=2}^{\infty} p^2 (1-p)^{n-2} = p$$