תאוריה סטטיסטית
88-775

עוזי וישנה

יוני 2017

מהדורה 1.067

סטטיסטיקה היא מערכת של כלים ושיטות הנמצאת בשימוש בכל תחומי החיים: במדע, בכלכלה, במדעי החברה, ועוד. הכלים הסטטיסטיים הם תמצית השיטה האינדוקטיבית המדעית, בכך שהם מאפשרים להסיק מסדרת ניסויים, כלומר נתוני מדגם, על האוכלוסיה כולה.

חלקו הראשון של הקורס הזה מציג שני נושאים בתאוריה סטטיסטית:

1. התאוריה של אמידה נקודתית: כיצד אומדים פרמטרים של האוכלוסיה מתוך המדגם, ואלו מגבלות יש על אמידה כזו. לצורך כך נציג את הכלים הנחוצים מתורת ההסתברות, במיוחד בכל הנוגע להתפלגויות של כמה משתנים, שאינן מקבלות טיפול ראוי בקורס ראשון בהסתברות.
2. העקרונות של תורת המבחנים הסטטיסטיים ובדיקת השערות.

בחלק השני נציג מגוון שיטות סטטיסטיות, בעיקר רגרסיה ומבחנים אי־פרמטריים.

תוכן עניינים

1 תורת ההסתברות 7
1.1 רקע נדרש . . . 7
1.1.1 מרחב הסתברות . . . 7
1.1.2 משתנים מקריים . . . 7
1.1.3 התוחלת . . . 8
1.1.4 שונות ושונות משותפת . . . 9
1.1.5 פונקציה יוצרת מומנטים . . . 10
1.1.6 חוקי המספרים הגדולים ומשפט הגבול המרכזי . . . 11
1.2 ההתפלגות הנורמלית . . . 12
1.2.1 ההתפלגות הנורמלית . . . 12
1.2.2 התפלגות רב־ממדית . . . 13
1.2.3 ההתפלגות הרב־נורמלית . . . 16
1.2.4 התפלגויות נלוות . . . 21
1.2.5 האומד ה־r־סימטרי . . . 26

2 אמידה 29
2.1 מודל, אוכלוסיה ומדגם . . . 29
2.2 אמידה נקודתית . . . 30
2.2.1 שיטת המומנטים . . . 31
2.2.2 אומדים חסרי הטיה . . . 33
2.2.3 השוואת אומדים . . . 36
2.2.4 אומד נראות מקסימלית . . . 37
2.2.5 סטטיסטיים מספיקים ומספיקים במשותף . . . 40
2.2.6 אינפורמציית פישר . . . 42
2.2.7 אי־שוויון קרמר־ראו . . . 44
2.2.8 אומדים חסרי הטיה בעלי שונות מינימלית במידה שווה . . . 46
2.2.9 סטטיסטיים שלמים . . . 48
2.2.10 משפחות מעריכיות . . . 49

2.3 רווחי סמך 50
2.3.1 שיטת הכמות הצירית 51
2.3.2 רווחי סמך עבור ההתפלגות הנורמלית 51

3 בדיקת השערות 53
3.1 השערות, הכרעות, והליך הבדיקה 53
3.1.1 השערת האפס וההשערה האלטרנטיבית 53
3.1.2 הכרעות ושגיאות 54
3.1.3 הליך הבדיקה 55
3.1.4 פונקציית העוצמה 56
3.2 השערות פשוטות 57
3.3 בדיקת השערות כללית 59
3.3.1 מבחן יחס הנראות המוכלל 59
3.3.2 בדיקת השערות למבחנים חד־צדדיים 60
3.3.3 בדיקת השערות באמצעות רווחי סמך 60

4 רגרסיה לינארית 63
4.1 רגרסיה דו־ממדית 63
4.1.1 אומדים לקו הרגרסיה 64
4.1.2 פירוק השונות 66
4.1.3 אמידת ערך חדש 68
4.1.4 בדיקת השערות על קו הרגרסיה 70
4.2 מבוא לרגרסיה רב־ממדית 70
4.2.1 בדיקת השערות 71
4.3 מבוא לניתוח שונות 72
4.3.1 ניתוח שונות חד־ממדי 72
4.3.2 ניתוח שונות דו־ממדית ואינטרקציה 73
4.4 מבוא לניתוח גורמים 73

5 מבחנים לא פרמטריים 75
5.1 ההתפלגות המולטינומית 75
5.2 מבחני χ^2 76
5.2.1 הכרעה בין שתי התפלגויות 76
5.2.2 מבחן לטיב ההתאמה 78
5.2.3 תלות בין משתנים בינאריים 80
5.2.4 האם משתנים בלתי תלויים הם שווי התפלגות 82
5.3 ניתוח אשכולות 82
5.3.1 מה אי־אפשר לעשות 82

5.3.2 ממוצעי־k . . . 83
5.3.3 שיטת GMM . . . 84
5.3.4 אישכול הירכי . . . 84
5.3.5 עץ פורש מינימלי . . . 85
5.3.6 אלגוריתמים מבוססי צפיפות . . . 85

תוכן עניינים

פרק 1

תורת ההסתברות

1.1 רקע נדרש

על אף שאנו מניחים שהקורא למד קורס ראשון בתורת ההסתברות, נחזור כאן על עיקרי הדברים. הקורא המבקש הרחבה מוזמן לעלעל בחוברת ההרצאות שלי לקורס "מבוא להסתברות וסטטיסטיקה", 88-165.

1.1.1 מרחב הסתברות

מרחב הסתברות הוא שלשה סדורה $(\Omega, \mathcal{F}, P)$ הכוללת את קבוצה Ω, סיגמא־אלגברה $\mathcal{F} \subseteq \mathcal{P}(\Omega)$, ופונקציה חיובית וסיגמא־אדיטיבית $P : \mathcal{F} \to \mathbb{R}$ המקיימת $P(\Omega) = 1$.

אם $A \in \mathcal{F}$, הערך $P(A)$ הוא ה**הסתברות** של **המאורע** A. לתת־קבוצות $A \subseteq \Omega$ שאינן שייכות ל־$\mathcal{F}$ אין ערך הסתברות. כאשר מרחב ההסתברות בדיד (סופי או בן־מניה), אפשר לקחת $\mathcal{F} = \mathcal{P}(\Omega)$. עבור מרחבים גדולים יותר, הבחירה ב־$\mathcal{F}$ במקום בקבוצת החזקה $\mathcal{P}(\Omega)$ כולה אינה נובעת מעצלות: לדוגמא, בשל אקסיומת הבחירה, לא ניתן להגדיר מידה אינווריאנטית להזזות על כל תת־הקבוצות של המעגל S^1.

הדוגמא החשובה ביותר היא **הסיגמא־אלגברה של בורל** על $\mathbb{R}$, הנוצרת על־ידי הקטעים הפתוחים, וכוללת את כל הקטעים, כל הקרניים, כל הקבוצות בנות־המניה, ועוד קבוצות רבות אחרות. את זאת אפשר להכליל בקלות לסיגמא־אלגברה של $\mathbb{R}^n$, הנוצרת על־ידי קוביות פתוחות.

1.1.2 משתנים מקריים

משתנה מקרי הוא פונקציה (מדידה) ממרחב ההסתברות אל המספרים הממשיים. זהו מושג יסודי ביותר, משום שהצמדת ערכים מספריים לנקודות של Ω היא ראשיתה של האנליזה על מרחב ההסתברות.

משתנה מקרי על מרחב בדיד מתואר באמצעות ה**התפלגות** שלו, שהיא הפונקציה מ־$a \in \Omega$ אל ההסתברות $P(X = a)$. דוגמאות חשובות: **התפלגות ברנולי**, ה**התפלגות הבינומית**, **התפלגות פואסון**, ה**התפלגות הגאומטרית**, ועוד רבות אחרות.

במקרה הכללי, שבו המרחב Ω אינו בן־מניה, שיטה זו אינה מועילה משום שבדרך כלל $P(X = a) = 0$ לכל a. אם כך, כיצד מתארים משתנה מקרי על $\Omega = \mathbb{R}$? הדרך הכללית ביותר היא באמצעות **פונקציית הצטברות**, $F_X(t) = P(X \leq t)$. פונקציית הצטברות היא מונוטונית, רציפה מימין, ושואפת בגבולות לאפס ולאחד. כל פונקציה כזו מתארת משתנה מקרי.

לעתים קרובות אפשר לתאר את המשתנה באמצעות **פונקציית צפיפות**, שהיא פונקציה חיובית ואינטגרבילית $f : \mathbb{R} \to \mathbb{R}$ כך ש־$\int_{-\infty}^{\infty} f(x)dx = 1$. פונקציית הצפיפות מגדירה פונצקיית הצטברות גזירה, לפי $F_X(x) = \int_{-\infty}^{x} f(t)dt$; ולהיפך כמובן, $f(x) = F'(x)$. היינו, $P(a \leq X \leq b) = \int_a^b f(x)dx$.

בין ההתפלגויות הרציפות שראוי להכיר נמנה את **ההתפלגות האחידה**, **ההתפלגות המעריכית** (המאופיינת בכך שהיא חסרת זכרון), ואת **ההתפלגות הנורמלית** שתשחק תפקיד חשוב לכל אורך הקורס.

התפלגות משותפת, תלות ואי־תלות

מכיוון שמשתנה מקרי הוא פונקציה ממרחב ההסתברות אל המספרים הממשיים, אפשר להתבונן בו־זמנית בכמה פונקציות, שלהן יש **התפלגות משותפת**. אינפורמציה (מהצורה $X \in A$, כאשר A קבוצת ערכים) מגדירה מאורע, ומאפשרת לטפל בכל משתנה מקרי אחר כ**משתנה מותנה**, $Y|(X \in A)$, קרי "Y בהנתן ש־$X \in A$". אם ההתפלגות של Y בהנתן $X = a$ היא תמיד אותה התפלגות, כלומר אינה משתנה עם a, אז X, Y **משתנים בלתי תלויים**, ואז מתקיים

$$P(X \in A, Y \in B) = P(X \in A) \cdot P(Y \in B).$$

1.1.3 התוחלת

המדד החשוב ביותר של משתנה מקרי הוא ה**תוחלת** $\mu = \mathbf{E}(X)$, השווה ל־$\sum P(X = a)a$ במקרה בדיד ול־$\int f(x)xdx$ במקרה הרציף.

התוחלת היא לינארית, כלומר הומוגנית ואדיטיבית: לכל משתנה מקרי X וסקלר a מתקיים $\mathbf{E}(aX) = a\mathbf{E}(X)$, ולכל שני משתנים מקריים X, Y מתקיים

$$\mathbf{E}(X + Y) = \mathbf{E}(X) + \mathbf{E}(Y).$$

תרגיל 1.1.1 אם X משתנה מקרי חיובי, אז $\mathbf{E}(X) = \int_0^\infty P(X \geq x)dx$.

אם X, Y משתנים מקריים, אז ידיעת X מגדירה משתנה מותנה $Y|X$, שההתפלגות שלו תלויה ב־X. לכן גם **התוחלת המותנית** $\mathbf{E}(Y|X)$ היא פונקציה של X. **חוק התוחלת החוזרת** קובע שהתוחלת של התוחלת המותנית של Y שווה לתוחלת של Y עצמו:

$$\mathbf{E}(Y) = \mathbf{E}(\mathbf{E}(Y|X)).$$

1.1.4 שונות ושונות משותפת

מן התוחלת מגדירים את **השונות** $\mathbf{V}(X) = \mathbf{E}((X - \mathbf{E}(X))^2) = \mathbf{E}(X^2) - \mathbf{E}(X)^2$.

בדומה לחוק התוחלת החוזרת, אפשר לחשב גם את השונות באמצעות פירוק המשתנה למשתנים מותנים:

$$\mathbf{V}(Y) = \mathbf{V}(\mathbf{E}(Y|X)) + \mathbf{E}(\mathbf{V}(Y|X)).$$

אם X, Y בלתי תלויים אז $\mathbf{E}(XY) = \mathbf{E}(X)\mathbf{E}(Y)$. משתנים המקיימים את התכונה החלשה הזו נקראים **בלתי מתואמים**. **השונות המשותפת** של X, Y היא

$$\mathrm{Cov}(X, Y) = \mathbf{E}(XY) - \mathbf{E}(X)\mathbf{E}(Y).$$

כלומר, X, Y בלתי מתואמים אם ורק אם $\mathrm{Cov}(X, Y) = 0$.

תרגיל 1.1.2 עבור משתנים מקריים $X_1, \dots, X_n$, נסמן $\mathrm{Cov}(\vec{X}) = (\mathrm{Cov}(X_i, X_j))_{ij}$. נגדיר משתנים חדשים לפי $\vec{Y} = P\vec{X}$, כאשר $P = (P_{ij})$ מטריצה; כלומר, $Y_i = \sum P_{ij}X_j$. הראה ש־

$$\mathrm{Cov}(P\vec{X}) = P\mathrm{Cov}(\vec{X})P^t.$$

הדרכה.

$$\begin{aligned}
\mathrm{Cov}(Y_i, Y_{i'}) &= \mathrm{Cov}(\sum_j P_{ij}X_j, \sum_{j'} P_{i'j'}X_{j'}) \\
&= \sum_{j,j'} P_{ij}\mathrm{Cov}(X_j, X_{j'})P_{i'j'} \\
&= (P\mathrm{Cov}(\vec{X})P^t)_{ii'}.
\end{aligned}$$

1.1.5 פונקציה יוצרת מומנטים

המומנט ה־k הוא התוחלת $\mathbf{E}(X^k)$. **המומנט המרכזי** ה־k הוא המומנט המתאים של $X-\mu$, כלומר $\mathbf{E}((X-\mu)^k)$. התוחלת היא המומנט הראשון, והשונות היא המומנט המרכזי השני. גם לשאר המומנטים יש תפקיד תאורטי חשוב. עם זאת, יש לציין כי המומנטים, ואפילו השונות או התוחלת, קיימים רק כאשר האינטגרל המגדיר אותם מתכנס, וזה לא בהכרח קורה.

תרגיל 1.1.3 אם למשתנה יש מומנט מסדר n, אז יש לו מומנטים מכל סדר קטן יותר.

יהי X משתנה מקרי. מגדירים את ה**פונקציה יוצרת המומנטים** של X לפי $M_X(t) = \mathbf{E}(e^{tX})$. מדוע טורחים להגדיר פונקציה כזו?

משפט 1.1.4 אם הפונקציה יוצרת המומנטים של משתנה מקרי X קיימת (וסופית) בקטע פתוח סביב 0, אז כל המומנטים קיימים, ויש פיתוח טיילור $M_X(t) = \sum_{n=0}^{\infty} \frac{\mathbf{E}(X^n)}{n!} t^n$; בפרט $\mathbf{E}(X^n) = M_X^{(n)}(0)$.

משפט 1.1.5 אם $M_X(t)$ קיימת לכל t, אז הפונקציה קובעת את ההתפלגות. היינו, אם $M_X(t) = M_Y(t)$ לכל t, אז X, Y שווי התפלגות.

(נכונה אפילו גרסה חזקה יותר: אם $M_{Y_n}(t) \rightarrow M_Y(t)$ לכל t, אז $Y_n \xrightarrow{D} Y$, כאשר $Y_n \xrightarrow{D} Y$ פירושו התכנסות בהתפלגות, היינו $P(Y_n \leq t) \rightarrow P(Y \leq t)$.)
פונקציות יוצרות מומנטים מאפשרות לחסר התפלגויות:

טענה 1.1.6 יהיו X, X', Y משתנים מקריים, כאשר Y אינו תלוי ב־X ואינו תלוי ב־X'. אם $X+Y$ ו־$X'+Y$ שווי־התפלגות, אז כך גם X, X'.

הוכחה. לפי ההנחה $M_X(t)M_Y(t) = M_{X+Y}(t) = M_{X'+Y}(t) = M_{X'}(t)M_Y(t)$, ולכן $M_X(t) = M_{X'}(t)$. כלומר, ל־X, X' יש אותה פונקציה יוצרת מומנטים, ולכן הם שווי־התפלגות. □

תרגיל 1.1.7 חשב את הפונקציה יוצרת המומנטים של ההתפלגות המעריכית. הוכח שההתפלגות הזו חסרת זכרון.

תרגיל 1.1.8 יהי X משתנה מקרי עם $\mu = \mathbf{E}(X)$ ו־$\sigma_k = \mathbf{E}((X-\mu)^k)$, ונסמן $\sigma^2 = \sigma_2$. יהי X' משתנה מקרי בעל אותה התפלגות, שאינו תלוי ב־X. הראה ש:

$$\begin{aligned}
\mathrm{Cov}(X,X^2) &= \sigma_3 + 2\mu\sigma^2;\\
\mathbf{E}(X^3) &= \sigma_3 + 3\mu\sigma^2 + \mu^3;\\
\mathbf{E}(X^4) &= \sigma_4 + 4\mu\sigma_3 + 6\mu^2\sigma^2 + \mu^4;\\
\mathbf{V}(X^2) &= \sigma_4 + 4\mu\sigma_3 + 4\mu^2\sigma^2 - \sigma^4;\\
\mathbf{V}(XX') &= \sigma^4 + 2\mu^2\sigma^2;\\
\mathbf{V}((X-\mu)^2) &= \sigma_4 - \sigma^4.
\end{aligned}$$

1.1.6 חוקי המספרים הגדולים ומשפט הגבול המרכזי

סדרת משתנים מקריים Y_n **מתכנסת בהסתברות** לקבוע μ (כותבים $Y_n \xrightarrow{P} \mu$), אם לכל $\varepsilon > 0$,

$$\lim_{n\to\infty} P(|Y_n - \mu| < \varepsilon) = 1.$$

במלים אחרות, לכל $\varepsilon > 0$, הסיכוי לכך ש־$|Y_n - \mu| > \varepsilon$ שואף לאפס.

משפט 1.1.9 (החוק החלש של המספרים הגדולים) תהי $X_1, X_2, \ldots$ סדרה של משתנים מקריים בלתי מתואמים, בעלי אותה תוחלת μ ושונות σ^2. אז $\bar{X}_n \xrightarrow{P} \mu$.

סדרת משתנים מקריים Y_n **מתכנסת כמעט תמיד** לקבוע μ (כותבים $Y_n \xrightarrow{a.s.} \mu$), אם $P(\lim Y_n = \mu) = 1$.

הערה 1.1.10 כאן צריך להוכיח שהדרישה $\lim Y_n = 0$ היא אכן מאורע (אחרת, הסתברות מניין). ואכן, המאורע הזה שווה ל־$\bigcap_d \bigcup_N \bigcap_{n\geq N}\{|Y_n| < 1/d\}$, ולכן שייך ל־$\sigma$-אלגברה שביחס אליה כל ה־$Y_n$ הם משתנים מקריים.

משפט 1.1.11 (החוק החזק של המספרים הגדולים) תהי $X_1, X_2, \ldots$ סדרת משתנים מקריים בלתי תלויים בעלי תוחלת μ ושונות σ^2. אז $\bar{X}_n \xrightarrow{a.s.} \mu$.

סדרת משתנים מקריים Y_n **מתכנסת בהתפלגות** למשתנה מקרי Y (כותבים $Y_n \xrightarrow{D} Y$), אם לכל a,

$$\lim_{n\to\infty} P(Y_n \leq a) = P(Y \leq a).$$

משפט 1.1.12 (משפט הגבול המרכזי) תהי $X_1, X_2, \ldots$ סדרה של משתנים מקריים בלתי תלויים בעלי אותה התפלגות, שיש לה תוחלת μ ושונות σ^2. אז $\frac{\bar{X}_n - \mu}{\sigma/\sqrt{n}} \xrightarrow{D} N(0,1)$.

1.2 ההתפלגות הנורמלית

1.2.1 ההתפלגות הנורמלית

ההתפלגות הנורמלית היא ההתפלגות השכיחה ביותר בתאוריה ובמעשה. לכך אחראי בעיקר משפט הגבול המרכזי, אבל גם כמה סיטואציות פשוטות שבהן מופיעה ההתפלגות הנורמלית כבמטה קסם, כפי שנדגים בסעיף הזה.

משתנה רציף שפונקציית הצפיפות שלו היא $\frac{1}{\sqrt{2\pi}}e^{-t^2/2}$ הוא **משתנה נורמלי סטנדרטי**; את ההתפלגות הזו מסמנים ב־$N(0,1)$. בטענה 1.2.12 נוכיח שזו אכן התפלגות (כלומר שהאינטגרל שווה ל־1), ואגב־כך נסביר את הקבוע המוזר שבמכנה.

תרגיל 1.2.1 הוכח בעזרת משפט הגבול המרכזי שאם p קבוע ו־$X \sim \mathrm{Bin}(n,p)$, אז, בקירוב, $X \sim N(np, npq)$.

תרגיל 1.2.2 חשב את הפונקציה יוצרת המומנטים של משתנה $Z \sim N(0,1)$, והוכח ש־$\mathbf{E}(Z) = 0$ ו־$\mathbf{V}(Z) = 1$. **הדרכה**.

$$\begin{aligned} M_Z(t) &= \mathbf{E}(e^{tZ}) = \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-z^2/2} e^{tz} dz \\ &= e^{t^2/2} \int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}} e^{-(t-z)^2/2} dz = e^{t^2/2}. \end{aligned}$$

לכן

$$\sum_{n=0}^{\infty} \frac{\mathbf{E}(Z^n)}{n!} t^n = M_Z(t) = e^{t^2/2} = \sum_{m=0}^{\infty} \frac{1}{2^m m!} t^{2m};$$

ומכאן ש־$\mathbf{E}(Z^{2m+1}) = 0$ ואילו $\mathbf{E}(Z^{2m}) = \frac{(2m)!}{2^m m!}$.

יהי $Z \sim N(0,1)$ משתנה נורמלי סטנדרטי. למשתנה $X = \mu + \sigma Z$ יש תוחלת μ ושונות σ^2; את ההתפלגות של X מסמנים ב־$N(\mu, \sigma^2)$. הצפיפות של $X \sim N(\mu, \sigma^2)$ היא $\frac{1}{\sqrt{2\pi\sigma^2}} e^{-(t-\mu)^2/(2\sigma^2)}$.

תרגיל 1.2.3 תהי F התפלגות כך שאם $X, Y \sim F$ בלתי תלויים, ו־$\alpha^2 + \beta^2 = 1$, אז גם $\alpha X + \beta Y \sim F$. הוכח ש־$F$ נורמלית.

הדרכה. ראשית, לפי ההנחה $\mathbf{V}(X) = \mathbf{V}(\alpha X + \beta Y) = (\alpha^2 + \beta^2)\mathbf{V}(X)$, ולכן התנאי על α, β הכרחי. נתבונן בפונקציה יוצרת המומנטים של F:

$$M_F(t) = M_{\alpha X + \beta Y}(t) = M_F(\alpha t) M_F(\beta t);$$

נסמן $f(t) = \log M_F(t)$, ניקח את הלוגריתם הטבעי, ונקבל

$$f(t) = f(\alpha t) + f(\beta t).$$

כעת נכתוב $\alpha = \cos\theta$ ו־$\beta = \sin\theta$, נציב, ונגזור לפי θ:

$$0 = -\sin\theta \cdot f'(\cos\theta \cdot t) + \cos\theta \cdot f(\sin\theta \cdot t).$$

על־ידי הצבת $\frac{1}{\cos\theta}t$ במקום t, נקבל

$$\tan\theta \cdot f'(t) = f'(\tan\theta \cdot t).$$

אבל מכיוון ש־$\tan\theta$ יכול לקבל כל ערך ממשי, נכתוב $a = \tan\theta$ ונקבל בנקודה $t = 1$ עבור קבוע C מתאים ש־$f'(a) = 2Ca$, כלומר $f(a) = Ca^2$. לכן $M_F(t) = e^{Ct^2}$, וזו הפונקציה יוצרת המומנטים של ההתפלגות $N(0, \sigma^2)$.

1.2.2 התפלגות רב־ממדית

בסעיף זה נעסוק בהתפלגות המשותפת של מספר משתנים רציפים, כהכללה של המקרה החד־ממדי, שבו מתוארת ההתפלגות על־ידי פונקציית צפיפות במשתנה ממשי אחד.

הגדרה 1.2.4 *פונקציית צפיפות n־ממדית היא פונקציה חיובית אינטגרבילית $f : \mathbb{R}^n \to \mathbb{R}$ כך ש־* $\int_{\mathbb{R}^n} f(\vec{x})d\vec{x} = 1$.
בניסוח מפורש יותר, נדרש ש־ $\int_{-\infty}^{\infty} \cdots \int_{-\infty}^{\infty} f(x_1, \ldots, x_n)dx_n \cdots dx_1 = 1$.

הצפיפות מגדירה את ההתפלגות המשותפת לפי הנוסחה

$$P((X_1, \ldots, X_n) \in A) = \int_A f(\vec{x})d\vec{x},$$

לכל קבוצה מדידה A.

אם $(\vec{X}, \vec{Y})$ הוא וקטור (באורך $n_x + n_y$) עם התפלגות $f_{\vec{X},\vec{Y}}$, אפשר לקבל את ההתפלגות של $\vec{X}$ על־ידי הטלה:

$$f_{\vec{X}}(\vec{x}) = \int f_{\vec{X},\vec{Y}}(\vec{x}, \vec{y})d\vec{y}.$$

זוהי **הצפיפות השולית** של $\vec{X}$, ובאותו אופן מוגדרת כמובן גם הצפיפות השולית של $\vec{Y}$. אפשר גם להגדיר את **הצפיפות המותנית**

$$f_{\vec{X}|\vec{Y}=y}(x) = \frac{f_{\vec{X},\vec{Y}}(x, y)}{f_{\vec{Y}}(y)},$$

וזו אכן פונקציית צפיפות (התלויה ב־y) עבור המשתנה $\vec{X}$. אם פונקציית הצפיפות המותנית אינה משתנה עם y, אז המשתנים **בלתי תלויים**. אפשר להכליל את המושג הזה ל**אי־תלות משותפת** של כמה משתנים (רב־ממדיים).

אם הוקטורים $\vec{X},\vec{Y}$ הם משתנים מקריים עם צפיפויות $f_{\vec{X}},f_{\vec{Y}}$ בהתאמה, אז הפונקציה $f_{\vec{X},\vec{Y}}(x,y)=f_{\vec{X}}(x)f_{\vec{Y}}(y)$ מגדירה פונקציית צפיפות חדשה. באופן כללי יותר, אם $f_{\vec{X},\vec{Y}}$ היא הצפיפות המשותפת ו־$f_{\vec{X}},f_{\vec{Y}}$ הצפיפויות השוליות, אז $\vec{X},\vec{Y}$ בלתי תלויים אם ורק אם $f_{\vec{X},\vec{Y}}(x,y)=f_{\vec{X}}(x)f_{\vec{Y}}(y)$.

תרגיל 1.2.5 $\vec{X},\vec{Y}$ בלתי תלויים אם ורק אם יש פירוק $f_{\vec{X},\vec{Y}}(x,y)=g(x)h(y)$ עבור פונקציות כלשהן g,h.

תרגיל 1.2.6 אם X,Y,Z בלתי תלויים במשותף, אז X בלתי תלויה ב־ $Y+Z$.

תרגיל 1.2.7 תן דוגמא שבה המשתנים X,Y,Z בלתי תלויים בזוגות, אבל X,Y,Z אינם בלתי תלויים במשותף.

אפשר להגדיר פונקציה יוצרת מומנטים לווקטור של משתנים, לפי

$$M_{\vec{X}}(\vec{t})=\mathbf{E}(e^{t_1X_1+\cdots+t_nX_n}).$$

הערה 1.2.8 המשתנים X,Y בלתי תלויים אם ורק אם $M_{X,Y}(t_1,t_2)=\mathbf{E}(e^{t_1X+t_2Y})$ שווה למכפלה $M_X(t_1)M_Y(t_2)$.

טרנספורמציה של צפיפויות

יהי $\vec{X}$ משתנה מקרי רב־ממדי, עם צפיפות $f_X(x)$. תהי $u\colon\mathbb{R}^n\to\mathbb{R}^n$ טרנספורמציה הפיכה. אפשר להגדיר משתנה מקרי חדש, $Y=u(X)$. במקרה הבדיד $P(Y=y)=P(X=u^{-1}(y))$, כך שהמעבר בין ההתפלגויות פשוט וקל. במקרה הרציף יש לקחת בחשבון את היעקוביאן של הטרנספורמציה.

הגדרה 1.2.9 ***היעקוביאן של הטרנספורמציה*** $(y_1,\dots,y_n)=u(x_1,\dots,x_n)$ *הוא המטריצה* $J(u)=(\frac{\partial y_i}{\partial x_j})_{ij}$.

תרגיל 1.2.10 היעקוביאן של טרנספורמציה לינארית $\vec{y}=P\vec{x}$, כאשר P מטריצה הפיכה, הוא $J(y)=P$.

טענה 1.2.11 יהי $\vec{X}$ משתנה מקרי רב־ממדי, עם צפיפות $f_X(x)$. תהי $u\colon\mathbb{R}^n\to\mathbb{R}^n$ טרנספורמציה הפיכה. אז הצפיפות של $Y=u(X)$ היא

$$f_Y(y)=f_X(x)\left|\det(J(u))\right|^{-1}.$$

הנה דוגמא חשובה:

טענה 1.2.12 $\int_{-\infty}^{\infty} \frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt = 1$, ולכן ההתפלגות הנורמלית $N(0,1)$ מוגדרת היטב.

הוכחה. **נתבונן במשתנים מקריים בלתי תלויים R, Θ, כאשר $\Theta \sim U[0,2\pi]$ ו-$f_R(r) = re^{-r^2/2}$** ($r > 0$). **קל לאשר שהפונקציה f_R היא אכן פונקציית צפיפות. מכיון ש- R, Θ בלתי תלויים, הצפיפות המשותפת שלהם היא**

$$f_{R,\Theta}(r,\theta) = \frac{1}{2\pi}re^{-r^2/2}.$$

נתבונן במשתנים X, Y, המוגדרים על־ידי הטרנספורמציה $(X,Y) = (R\cos\Theta, R\sin\Theta)$. **היעקוביאן הוא** $J = \begin{pmatrix} \cos\theta & \sin\theta \\ -r\sin\theta & r\cos\theta \end{pmatrix}$, **עם דטרמיננטה** $\det(J) = r$. **לכן הצפיפות המשותפת של X, Y היא**

$$f_{X,Y}(x,y) = \frac{1}{2\pi}e^{-r^2/2} = \frac{1}{2\pi}e^{-(x^2+y^2)/2} = \left(\frac{1}{\sqrt{2\pi}}e^{-x^2/2}\right)\left(\frac{1}{\sqrt{2\pi}}e^{-y^2/2}\right).$$

נחשב את ריבוע האינטגרל, על־ידי החלפת משתנים:

$$\left(\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi}}e^{-t^2/2}dt\right)^2 = \int_{-\infty}^{\infty}\int_{-\infty}^{\infty}\frac{1}{2\pi}e^{-(x^2+y^2)/2}dxdy,$$

כלומר $X, Y \sim N(0,1)$, ומכאן שהצפיפות הנטענת עבור ההתפלגות הנורמלית היא אכן פונקציית צפיפות. □

תרגיל 1.2.13 נניח ש-$U, V \sim U[0,1]$ משתנים בלתי תלויים. הראה ש-$X = \sqrt{-2\log U}\sin(2\pi V)$ ו-$Y = \sqrt{-2\log U}\cos(2\pi V)$ הם נורמליים ובלתי תלויים [Box-Muller, 1958].

תרגיל 1.2.14 אם $X, Y \sim N(0,1)$ הם משתנים בלתי תלויים, אז $\Theta = \arctan(Y/X)$ הוא משתנה מקרי אחיד בקטע $[0,2\pi]$. **הדרכה**. זו תוצאה מן ההוכחה של טענה 1.2.12.

הנוסחה שבטענה 1.2.11 מאפשרת לחשב את הצפיפות של סכום משתנים בלתי תלויים.

טענה 1.2.15 יהיו X, Y משתנים בלתי תלויים עם צפיפויות f_X, f_Y. הראה שהצפיפות של $X+Y$ היא

$$f_{X+Y}(z) = \int_{-\infty}^{\infty} f_X(x)f_Y(z-x)dx.$$

(זוהי **הקונוולוציה** של f_x, f_Y.)

הוכחה. המעבר מהזוג הסדור (x,y) לסכום $x+y$ אינו הפיך, ולכן צריך להוסיף משתנה עזר. נתבונן בטרנספורמציה $u(x,y)=(x+y,y)$. היעקוביאן הוא $J=\begin{pmatrix}1 & 1\\ 0 & 1\end{pmatrix}$, עם דטרמיננטה 1, ולכן $f_{X+Y,Y}(z,y)=f_X(z-y)f_Y(y)$. נשאר לחלץ את הצפיפות השולית של $X+Y$, על־ידי אינטגרציה על y. □

תרגיל 1.2.16 תן נוסחאות דומות לצפיפות של מכפלה ושל מנה של שני משתנים מקריים.

1.2.3 ההתפלגות הרב־נורמלית

מטריצות חיוביות לחלוטין

להתפלגות הרב־נורמלית יש תפקיד מיוחד בכל ניתוח רב־משתני. לפני שנגדיר את משפחת ההתפלגויות הזו, נזכיר מושג חשוב מאלגברה לינארית.

הגדרה 1.2.17 *מטריצה סימטרית $\Sigma\in \mathrm{M}_n(\mathbb{R})$* ***היא חיובית*** *אם לכל $x\in\mathbb{R}^n$ מתקיים $x^t\Sigma x\geq 0$,* ***וחיובית לחלוטין*** *אם לכל $0\neq x\in\mathbb{R}^n$ מתקיים $x^t\Sigma x>0$.*

(באנגלית נקראת מטריצה חיובית לחלוטין positive definite. הזהרו מן התרגום השגוי בתכלית "מטריצה מוגדרת חיובית"; שום דבר במטריצות האלה אינו מוגדר באופן היוצא מגדר הרגיל.)

תרגיל 1.2.18 לכל וקטור של משתנים $X=(X_1,\dots,X_n)^t$, המטריצה $\mathrm{Cov}(X)$ חיובית; והיא חיובית לחלוטין אלא אם המשתנים תלויים לינארית (בהסתברות 1). **הדרכה.** $a^t\mathrm{Cov}(X)a=\mathbf{V}(\sum a_iX_i)\geq 0$.

תרגיל 1.2.19 מטריצה סימטרית $A\in \mathrm{M}_n(\mathbb{R})$ היא חיובית לחלוטין אם ורק אם היא מהצורה $A=PP^t$ כאשר P מטריצה הפיכה.

מסקנה 1.2.20 המטריצה Σ חיובית לחלוטין אם ורק אם Σ^{-1} חיובית לחלוטין.

ההתפלגות הרב־נורמלית

הגדרה 1.2.21 *תהי Σ מטריצה חיובית לחלוטין בגודל n.* ***ההתפלגות הרב־נורמלית*** *$N(0,\Sigma)$ היא ההתפלגות בעלת פונקציית הצפיפות ה־n ממדית*

$$f(\vec{x})=\frac{1}{\sqrt{(2\pi)^n\det(\Sigma)}}e^{-\frac{1}{2}\vec{x}^t\Sigma^{-1}\vec{x}}.$$

מיד נראה שזו אכן פונקציית צפיפות. ראשית נבחן מה קורה כאשר Σ היא מטריצת היחידה.

טענה 1.2.22 נניח ש־$X = (X_1, \ldots, X_n)$ מתפלג לפי $X \sim N(0, \Delta)$, כאשר $\Delta = \mathrm{diag}(\delta_{11}, \ldots, \delta_{nn})$ מטריצה אלכסונית. אז $X_1, \ldots, X_n$ הם נורמליים ובלתי תלויים.

הוכחה. **מכיוון שאנו מניחים ש־Δ חיובית לחלוטין, בהכרח $\delta_{ii} > 0$. אם $X \sim N(0, \Delta)$ אז**

$$f(x_1, \ldots, x_n) = \frac{1}{\sqrt{(2\pi)^n \prod \delta_{ii}}} e^{-\frac{1}{2}\vec{x}^t \Delta^{-1} \vec{x}} = \prod_{i=1}^{n} \frac{1}{\sqrt{2\pi\delta_{ii}}} e^{-\frac{1}{2\delta_{ii}} x_i^2},$$

כלומר במקרה זה $X_i \sim N(0, \delta_{ii})$, והמשתנים בלתי תלויים. □

התכונה הבסיסית של משפחת ההתפלגויות הרב־נורמלית היא הסגירות להעתקות לינאריות הפיכות.

טענה 1.2.23 אם $X \sim N(0, \Sigma)$, אז לכל מטריצה הפיכה P, $PX \sim N(0, P\Sigma P^t)$.

הוכחה. **כפי שראינו בתרגיל 1.2.10, היעקוביאן של הטרנספורמציה הלינארית $\vec{y} = P\vec{x}$ הוא P. נסמן $\vec{y} = P\vec{x}$. לכן, אם $\vec{X} \sim N(0, \Sigma)$, הצפיפות של $P\vec{X}$ היא**

$$f_{P\vec{X}}(\vec{y}) = \frac{1}{\sqrt{(2\pi)^n \det(P\Sigma P^t)}} e^{-\frac{1}{2}\vec{x}^t (P\Sigma P^t)^{-1} \vec{x}}.$$

□

מסקנה 1.2.24 אם Σ מטריצה חיובית לחלוטין, אז הפונקציה המופיעה בהגדרה 1.2.21 היא אכן פונקציית צפיפות.

הוכחה. **אכן, לפי ההנחה אפשר לכתוב $\Sigma = PP^t$. ניקח $X \sim N(0, I)$; זוהי אכן התפלגות מוגדרת היטב לפי טענה 1.2.22. לפי טענה 1.2.23, הצפיפות של PX מתוארת על־ידי ההתפלגות $N(0, PP^t) = N(0, \Sigma)$, ולכן גם זו התפלגות מוגדרת היטב.** □

כעת נוכל להכליל את טענה 1.2.22.

טענה 1.2.25 נניח ש־$X = (X_1, \ldots, X_n)$ מתפלג $X \sim N(0, \Sigma)$, אז $\Sigma = \mathrm{Cov}(X)$.

הוכחה. **נכתוב $\Sigma = PP^t$ עבור מטריצה הפיכה P. לפי טענה 1.2.23, $Z = P^{-1}X \sim N(0, P^{-1}\Sigma P^{-t}) = N(0, I)$. כלומר, $Z_1, \ldots, Z_n$ הם נורמליים סטנדרטיים ובלתי תלויים, ולכן $\mathrm{Cov}(Z) = I$. נתבונן מחדש ב־$X = PZ$: לפי תרגיל 1.1.2,**

$$\mathrm{Cov}(X) = \mathrm{Cov}(P\vec{Z}) = P\mathrm{Cov}(\vec{Z})P^t = PP^t = \Sigma.$$

□

כלומר, אפשר לשחזר את המטריצה Σ המגדירה התפלגות רב־נורמלית, מתוך השונויות המשותפות של הרכיבים בווקטור. עובדה זו מאפשרת לחלץ מסקנה חזקה על אי־תלות מהנחה חלשה על שונויות משותפות:

מסקנה 1.2.26 נניח ש־$\vec{X} \sim N(0,\Sigma)$. אם $X_1,\dots,X_n$ בלתי מתואמים בזוגות, אז הם בלתי תלויים במשותף.

הוכחה. לפי ההנחה $\Sigma = \mathrm{Cov}(X)$ מטריצה אלכסונית, ולפי טענה 1.2.22 המשתנים X_i נורמליים ובלתי תלויים. □

וקל־וחומר:

מסקנה 1.2.27 נניח ש־$\vec{X} \sim N(0,\Sigma)$. אם $X_1,\dots,X_n$ בלתי תלויים בזוגות, אז הם בלתי תלויים במשותף (ולכן Σ אלכסונית).

מעלה ומטה

טענה 1.2.28 יהיו X,Y וקטורים מקריים בגודל n,m, ותהיינה Σ',Σ'' מטריצות חיוביות לחלוטין בגודל n,m, בהתאמה. אז התכונות הבאות שקולות:

1. $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma' & 0 \\ 0 & \Sigma'' \end{pmatrix}\right)$;
2. $X \sim N(0,\Sigma')$, $Y \sim N(0,\Sigma'')$, ו־X,Y בלתי תלויים.

הוכחה. $(\vec{x},\vec{y})^t \begin{pmatrix} \Sigma' & 0 \\ 0 & \Sigma'' \end{pmatrix}^{-1} \begin{pmatrix} \vec{x} \\ \vec{y} \end{pmatrix} = \vec{x}^t\Sigma'^{-1}\vec{x} + \vec{y}^t\Sigma''^{-1}\vec{y}$, ולכן פונקציית הצפיפות המשותפת של ההתפלגות הנורמלית $N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} \Sigma' & 0 \\ 0 & \Sigma'' \end{pmatrix}\right)$ מתפרקת למכפלת פונקציות הצפיפות של ההתפלגויות $N(0,\Sigma')$ ו־$N(0,\Sigma'')$. □

(בעזרת הכללה מיידית של הטענה הזו אפשר לתאר לא רק זוגות, אלא קבוצות של משתנים וקטוריים בלתי תלויים במשותף.)

טענה 1.2.29 יהיו X,Y וקטורים מקריים בגודל n,m בהתאמה, כך ש־$\begin{pmatrix} X \\ Y \end{pmatrix} \sim N(0,\Sigma)$, כאשר $\Sigma = \begin{pmatrix} \Sigma' & R \\ R^t & \Sigma'' \end{pmatrix}$ היא מטריצת בלוקים בגודל $(n+m)\times(n+m)$. אז $X \sim N(0,\Sigma')$.

הוכחה. כתת־מטריצה של Σ, Σ' חיובית לחלוטין ולכן הפיכה. נבחר $P = \begin{pmatrix} I & 0 \\ -R^t\Sigma'^{-1} & I \end{pmatrix}$, ונחשב:

$$\begin{aligned} P\Sigma P^t &= \begin{pmatrix} I & 0 \\ -R^t\Sigma'^{-1} & I \end{pmatrix}\begin{pmatrix} \Sigma' & R \\ R^t & \Sigma'' \end{pmatrix}\begin{pmatrix} I & -\Sigma'^{-1}R \\ 0 & I \end{pmatrix} \\ &= \begin{pmatrix} \Sigma' & R \\ 0 & \Sigma'' - R^t\Sigma'^{-1}R \end{pmatrix}\begin{pmatrix} I & -\Sigma'^{-1}R \\ 0 & I \end{pmatrix} \\ &= \begin{pmatrix} \Sigma' & 0 \\ 0 & \Sigma'' - R^t\Sigma'^{-1}R \end{pmatrix}. \end{aligned}$$

לפי טענה 1.2.28, רכיבי הווקטור $P\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} I & 0 \\ -R^t\Sigma'^{-1} & I \end{pmatrix}\begin{pmatrix} X \\ Y \end{pmatrix} = \begin{pmatrix} X \\ Y - R^t\Sigma'^{-1}X \end{pmatrix}$ בלתי תלויים, ומתפלגים רב־נורמלית עם מטריצות השונויות Σ' ו־$(\Sigma'' - R^t\Sigma'^{-1}R)$, בהתאמה. □

מסקנה 1.2.30 אם $X \sim N(0,\Sigma)$, אז $X_i \sim N(0,\Sigma_{ii})$.

מסקנה 1.2.31 בתנאי טענה 1.2.29, ההתפלגות של Y בהנתן $X = 0$ היא נורמלית, $(Y|X=0) \sim N(0, \Sigma'' - R^t\Sigma'^{-1}R)$.

תרגיל 1.2.32 נניח ש־$X_i \sim N(0,\sigma_i^2)$ משתנים נורמליים בלתי תלויים $(i = 1,\dots,n)$. אז $\sum a_iX_i \sim N(0, \sum a_i^2\sigma_i^2)$.

הדרכה. נסמן $\Delta = \mathrm{diag}(\sigma_1^2,\dots,\sigma_n^2)$. לפי ההנחה, $X \sim N(0,\Delta)$. נבחר מטריצה הפיכה כלשהי P שהשורה הראשונה שלה היא $(a_1,\dots,a_n)$ (למשל על־ידי השלמה לבסיס). לפי טענה 1.2.23, $PX \sim N(0, P\Delta P^t)$, ולפי תרגיל 1.2.30

$$\sum a_iX_i = (PX)_1 \sim N(0, (P\Delta P^t)_{11}) = N(0, \sum a_i^2\sigma_i^2).$$

הזזה הצידה

עד כה הנחנו שמרכז ההתפלגות הוא בראשית הצירים. המעבר למקרה הכללי חלק:

הגדרה 1.2.33 *תהי Σ מטריצה בגודל n ויהי μ וקטור באורך n. **ההתפלגות הרב־נורמלית** $N(\mu,\Sigma)$ היא ההתפלגות בעלת פונקציית הצפיפות ה־n ממדית*

$$f(\vec{x}) = \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^t\Sigma^{-1}(\vec{x}-\mu)}.$$

התפלגות זו מתקבלת על־ידי הוספת הווקטור הקבוע μ למשתנה בעל ההתפלגות $N(0,\Sigma)$. אם $X \sim N(\mu,\Sigma)$, אז $\mu = \mathbf{E}(X)$ ו־$\Sigma = \mathrm{Cov}(X)$ כמקודם.

תרגיל 1.2.34 אם $X \sim N(\mu,\Sigma)$, אז לכל וקטור a מתקיים $X + a \sim N(\mu + a,\Sigma)$.

טענה 1.2.35 הפונקציה יוצרת המומנטים של ההתפלגות $N(\mu,\Sigma)$ היא

$$M_{\vec{X}}(\vec{s}) = e^{\mu\cdot\vec{s}+\frac{1}{2}\vec{s}^t\Sigma\vec{s}}.$$

הוכחה. **נניח ש־**$X \sim N(\mu,\Sigma)$. **אז**

$$\begin{aligned} M_{\vec{X}}(\vec{s}) &= \mathbf{E}(e^{s\cdot X}) \\ &= \int_{\mathbb{R}^n} \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu})^t\Sigma^{-1}(\vec{x}-\mu)} e^{s\cdot\vec{x}} d\vec{x} \\ &= \int_{\mathbb{R}^n} \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} e^{-\frac{1}{2}[(\vec{x}-\vec{\mu})^t\Sigma^{-1}(\vec{x}-\mu)-2s\cdot\vec{x}]} d\vec{x} \\ &= e^{\mu\cdot\vec{s}+\frac{1}{2}\vec{s}^t\Sigma\vec{s}} \int_{\mathbb{R}^n} \frac{1}{\sqrt{(2\pi)^n \det(\Sigma)}} e^{-\frac{1}{2}(\vec{x}-\vec{\mu}-\Sigma\vec{s})^t\Sigma^{-1}(\vec{x}-\mu-\Sigma\vec{s})} d\vec{x} \\ &= e^{\mu\cdot\vec{s}+\frac{1}{2}\vec{s}^t\Sigma\vec{s}}. \end{aligned}$$

□

תרגיל 1.2.36 יהיו $X_1,\dots,X_m$ משתנים מקריים (למשל הציונים בשאלות של מבחן), ו־$Y = \sum X_i$ (הציון הסופי). **מקדם** α **של** Cronbach, המוגדר לפי

$$\alpha = \frac{m}{m-1}\left(1 - \frac{\sum \mathrm{V}(X_i)}{\mathrm{V}(Y)}\right)$$

הוא מדד מקובל לאמינות של המבחן. ערכים מעל 0.8-0.7 נחשבים למעידים על מבחן בעל אמינות טובה.

1. בדוק שתמיד $\alpha \leq 1$, ומתקבל שוויון רק כאשר כל ה־X_i שווים זה לזה (בהסתברות 1).
2. נניח ש־$\vec{X} \sim N(\mu,\Sigma)$, וקטור ממימד m. הראה ש־$\alpha = \frac{m}{m-1}\left(1 - \frac{\sum \Sigma_{ii}}{\sum_{i,j}\Sigma_{ij}}\right)$.
3. בפרט, אם $\mathrm{V}(X_i) = 1$ ו־$\mathrm{Cov}(X_i,X_j) = \rho$ לכל $i \neq j$, אז $\alpha = \frac{1}{1+\frac{1-\rho}{m\rho}}$.
4. השווה את המבחן על $X_1,\dots,X_m$ למבחן דומה על $X = X_1 + \cdots + X_k$ ו־$X' = X_{k+1} + \cdots + X_m$. השווה לממוצע המבחנים על פני כל החלוקות.

1.2.4 התפלגויות נלוות

יש כמה התפלגויות הנולדות מתוך ההתפלגות הנורמלית, ומשחקות תפקיד בהערכה של פרמטרים מההתפלגות הזו. ההתפלגויות האלה מאזכרות הכללה ידועה לפונקציית העצרת, הקרויה **פונקציית גמא**. פונקציה זו מוגדרת על־ידי הנוסחה

$$\Gamma(t) = \int_0^\infty x^{t-1}e^{-x}dx,$$

ומקיימת את המשוואה הפונקציונלית

$$\Gamma(z+1) = z \cdot \Gamma(z).$$

בפרט, $\Gamma(1) = \int_0^\infty e^{-x}dx = 1$, ולכן $\Gamma(n+1) = n!$ לכל מספר טבעי n. ערכים אחרים קשה יותר לחשב. למשל, $\Gamma(1/2) = \sqrt{\pi/2}$.

הגדרה 1.2.37 *יהיו $X_1, \dots, X_n$ משתנים מקריים. המשתנים $X_{(1)}, \dots, X_{(n)}$ מוגדרים כסידור מחדש של הערכים $X_1, \dots, X_n$, כך ש־$X_{(1)} \leq \cdots \leq X_{(n)}$. משתנים אלה נקראים **סטטיסטיי הסדר**. בפרט,*

$$X_{(1)} = \min\{X_1, \dots, X_n\},$$

$$X_{(n)} = \max\{X_1, \dots, X_n\}.$$

תרגיל 1.2.38 נניח ש־$X_1, \dots, X_n \sim U[0,1]$. הראה שהצפיפות של סטטיסטי הסדר $X_{(k)}$ היא $f_{X_{(k)}}(x) = \frac{n!}{(k-1)!(n-k)!}x^{k-1}(1-x)^{n-k}$. הראה ש־

$$\mathbf{E}(X_{(k)}^r) = \frac{n!\Gamma(r+k)}{(k-1)!\Gamma(n+r+1)};$$

בפרט, אם r שלם, $\mathbf{E}(X_{(k)}^r) = \frac{\binom{r+k-1}{r}}{\binom{n+r}{r}}$.

תרגיל 1.2.39 הצפיפות המשותפת של $X_{(j)}, X_{(k)}$ (כאשר $j < k$) היא

$$f_{X_{(j)},X_{(k)}}(x,y) = \frac{n!}{(j-1)!(k-j-1)!(n-k)!}x^{j-1}(y-x)^{k-j-1}(1-y)^{n-k}.$$

בפרט, הצפיפות המשותפת של המינימום והמקסימום היא

$$f_{X_{(1)},X_{(n)}}(x,y) = n(n-1)(y-x)^{n-2}.$$

תרגיל 1.2.40 הנקודות $P_1, \ldots, P_n$ מפוזרות בהתפלגות אחידה על מעגל שהיקפו 1. מה הסיכוי שיש קשת באורך α המכסה את כל הנקודות, בהנחה ש־$\alpha \leq 1/2$? **הדרכה**. קטע כזה קיים אם ורק אם אחת הנקודות P_i היא "שמאלית ביותר", במובן שכל שאר הנקודות מימינה ובמרחק לכל היותר α (ורק אחת, משום ש־$\alpha \leq 1/2$). הסיכוי לכך הוא α^{n-1}, ומכיוון שיש n נקודות העשויות להיות השמאלית ביותר, ההסתברות היא $n\alpha^{n-1}$.

התפלגות גמא

אחרי שהגדרנו את פונקציית גמא, אי אפשר שלא להציג את התפלגות גמא. התפלגות זו קשורה קשר הדוק להתפלגות המעריכית, אבל פרט למקרה פרטי חשוב שנפגוש מיד, אין לה קשר מיוחד להתפלגות הנורמלית.

הגדרה 1.2.41 *משתנה X הוא בעל התפלגות גמא $X \sim \Gamma(k, \lambda)$ אם יש לו פונקציית הצפיפות* $f_X(x) = \frac{1}{\Gamma(k)\lambda^k} x^{k-1} e^{-x/\lambda}$, $x > 0$.

למשתנה $X \sim \Gamma(k, \lambda)$ יש תוחלת $\mathbf{E}(X) = k\lambda$ ושונות $\mathbf{V}(X) = k\lambda^2$.

תרגיל 1.2.42 ההתפלגות $\Gamma(1, \lambda)$ אינה אלא ההתפלגות המעריכית $\mathrm{Exp}(\lambda)$, עם צפיפות $f_X(x) = \frac{1}{\lambda} e^{-x/\lambda}$.

תרגיל 1.2.43 הפונקציה יוצרת המומנטים של ההתפלגות $\Gamma(k, \lambda)$ היא $(1 - \lambda t)^{-k}$.

טענה 1.2.44 **לכל k, k', אם $Y \sim \Gamma(k, \lambda)$ ו־$Y' \sim \Gamma(k', \lambda)$ הם משתנים מקריים בלתי תלויים, אז $Y + Y' \sim \Gamma(k + k', \lambda)$.**

הוכחה. קונוולוציה, או השוואה של הפונקציות יוצרות המומנטים. □

מסקנה 1.2.45 **נניח ש־$X_1, X_2, \cdots \sim \mathrm{Exp}(\lambda)$ הם משתנים בלתי תלויים. לכל n, הסכום $S_n = X_1 + \cdots + X_n$ מתפלג $\Gamma(n, \lambda)$.**

תרגיל 1.2.46 $X_1, \ldots, X_n \sim U[0, 1]$ והם בלתי תלויים. חשב את ההתפלגות של המכפלה $X_1 \cdots X_n$. **הדרכה**. $-\log X_i \sim \mathrm{Exp}(1)$.

התפלגות χ^2

נניח ש־$Z_1, \ldots, Z_n \sim N(0, 1)$ והם בלתי תלויים. לסכום הריבועים

$$W = Z_1^2 + \cdots + Z_n^2$$

יש התפלגות, הקרויה **התפלגות חי־בריבוע עם n דרגות חופש**, ומסומנת ב־$W \sim \chi_n^2$.

הערה 1.2.47 התפלגות χ^2 היא מקרה פרטי של התפלגות גמא: $\chi^2_n = \Gamma(\frac{n}{2}, 2)$.

פונקציית הצפיפות היא

$$\frac{x^{n/2-1}e^{-x/2}}{\Gamma(n/2)2^{n/2}}. \tag{1.1}$$

התוחלת של $W \sim \chi^2_n$ היא $\mathbf{E}(W) = n$, והשונות $\mathbf{V}(W) = 2n$. הפונקציה יוצרת המומנטים היא $M_W(t) = (1-2t)^{-n/2}$.

דוגמא 1.2.48 1. פונקציית הצפיפות של התפלגות χ^2_1 היא $\frac{x^{-1/2}e^{-x/2}}{\sqrt{2\pi}}$, שאינה חסומה.

2. $\chi^2_2 = \mathrm{Exp}(2)$. כלומר, ההתפלגות χ^2_2 היא ההתפלגות המעריכית עם תוחלת 2; אכן, לפי (1.1) עם $n = 2$, הצפיפות של $W = Z_1^2 + Z_2^2$ היא $\frac{1}{2}e^{-x/2}$.

3. את ההתפלגות של $\sqrt{\chi^2_2} = \sqrt{Z_1^2 + Z_2^2}$ חישבנו בטענה 1.2.12.

הערה 1.2.49 נניח ש־$Z_i \sim N(0,1)$. הממוצע של $Z_1^2, \dots, Z_n^2$ הוא $\frac{1}{n}W_n$, ולפי משפט הגבול המרכזי, $\frac{W_n - n}{\sqrt{2n}} \xrightarrow{D} N(0,1)$. לכן כאשר n גדול, בקירוב $\chi^2_n \sim N(n, 2n)$. הסטיה המקסימלית בין שתי ההתפלגויות היא $\left|P(\chi^2_n < n) - P(\frac{Z-n}{\sqrt{2n}} < n)\right| \sim \frac{0.188}{\sqrt{n}}$ (כך נראה מבדיקה נומרית).

התפלגות t

אם $Z \sim N(0,1)$ ו־$W \sim \chi^2_n$ בלתי תלויים, אז $T = \frac{Z}{\sqrt{W/n}}$ הוא בעל התפלגות הנקראת **התפלגות t של סטודנט**, עם n דרגות חופש ('סטודנט' הוא הכינוי שאימץ לעצמו הסטטיסטיקאי ויליאם סילי גוסט, שפרסם את ההתפלגות ב־1908).

פונקציית הצפיפות של ההתפלגות הזו היא

$$f(t) = \frac{\Gamma(\frac{n+1}{2})}{\sqrt{n\pi}\,\Gamma(\frac{n}{2})}\left(1 + \frac{t^2}{n}\right)^{-\frac{n+1}{2}}.$$

התוחלת היא $\mathbf{E}(T) = 0$ (כאשר $n > 1$; ב־ $n = 1$ אין תוחלת: האינטגרל אינו מתכנס) והשונות $\mathbf{V}(T) = \frac{n}{n-2}$.

כאשר n גדול (למשל $n = 30$), התפלגות זו קרובה להתפלגות הנורמלית, משום שלפי משפט הגבול המרכזי $W/n = 1 + O(1/\sqrt{n})$ ולכן גם $\frac{1}{\sqrt{W/n}} = 1 + O(\frac{1}{\sqrt{n}})$.

התפלגות F

להתפלגות היחס $X = \frac{U/n}{V/m}$ כאשר $U \sim \chi^2_n$ ו־$V \sim \chi^2_m$ קוראים **התפלגות** F, ומסמנים $X \sim F_{n,m}$. פונקציית הצפיפות היא

$$\frac{\Gamma(\frac{n+m}{2})}{\Gamma(\frac{n}{2})\Gamma(\frac{m}{2})} \left(\frac{n}{m}\right)^{\frac{n}{2}} x^{\frac{n}{2}-1} \left(1+\frac{n}{m}x\right)^{-\frac{n+m}{2}}.$$

התוחלת היא $\mathbf{E}(X) = \mathbf{E}(\frac{1}{n}U)\mathbf{E}(m\frac{1}{V}) = \frac{m}{m-2}$. השונות היא $\mathbf{V}(X) = \frac{2m^2(m+n-2)}{n(m-2)^2(m-4)}$.

תרגיל 1.2.50 $T \sim t_n$ אם ורק אם $T^2 \sim F_{1,n}$.

תרגיל 1.2.51 אם $X \sim F_{n,m}$ אז $\frac{1}{X} \sim F_{m,n}$.

תרגיל 1.2.52 אם $X \sim F_{n,m}$ ו־$X' \sim F_{m,n}$, אז $P(X' < a) = 1 - P(X < 1/a)$.

אמידת התוחלת והשונות

אם $X_1, \dots, X_n$ משתנים מקריים, מגדירים $\bar{X} = \frac{1}{n}(X_1 + \cdots + X_n)$ ו־

$$S^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2. \qquad (1.2)$$

טענה 1.2.53 נניח ש־$X_1, \dots, X_n$ משתנים בלתי מתואמים, שיש להם תוחלת μ ושונות σ^2. אז $\mathbf{E}(\bar{X}) = \mu$ ו־$\mathbf{E}(S^2) = \sigma^2$.

הוכחה. לפי ההנחה $\mathbf{E}(X_i) = \mu$ לכל i, ולכן גם $\mathbf{E}(\bar{X}) = \mu$. כעת נחשב

$$\begin{aligned}(n-1)S^2 &= \sum(X_i - \bar{X})^2 \\ &= \sum X_i^2 - 2n\bar{X}^2 + n\bar{X}^2 \\ &= \sum X_i^2 - n\bar{X}^2\end{aligned}$$

ולכן

$$\begin{aligned}\mathbf{E}((n-1)S^2) &= \sum \mathbf{E}(X_i^2) - n\mathbf{E}(\bar{X}^2) \\ &= \sum(\sigma^2+\mu^2) - n(\mu^2 + \frac{1}{n}\sigma^2) \\ &= (n-1)\sigma^2\end{aligned}$$

□

האומדים לפרמטרים של ההתפלגות הנורמלית

נתבונן במודל השכיח $X \sim N(\mu,\sigma^2)$. נתון מדגם מן המודל הזה, כלומר $X_1,\dots,X_n \sim N(\mu,\sigma^2)$, שהם בלתי תלויים. האומדים הסטנדרטיים ל־μ ולשונות σ^2 מוכרים; אבל מה ההתפלגות שלהם? כדי לטפל בסוגיה זו, נפעיל את המנגנון שפיתחנו עבור ההתפלגות הרב־נורמלית. לפי ההנחה, $\frac{1}{\sigma}(\vec{X}-\mu) \sim N(0,I)$.

תרגיל 1.2.54 נניח ש־$X_1,\dots,X_n \sim N(\mu,\sigma^2)$. אז $\bar{X} \sim N(\mu,\frac{\sigma^2}{n})$. **הדרכה.** לפי תרגיל 1.2.32.

משפט 1.2.55 *נניח ש־$X_1,\dots,X_n \sim N(\mu,\sigma^2)$ משתנים נורמליים בלתי תלויים. אז:*

1. $\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}} \sim N(0,1)$.
2. *$\bar{X}_n$ ו־S בלתי תלויים.*
3. $\frac{n-1}{\sigma^2}S^2 = \sum(\frac{X_i-\bar{X}_n}{\sigma})^2 \sim \chi^2_{n-1}$.
4. *$Q = \frac{\bar{X}_n-\mu}{S/\sqrt{n}}$ הוא משתנה מקרי בעל התפלגות t_{n-1}.*

הוכחה. **1.** לפי תרגיל **1.2.32**, $\bar{X} \sim N(\mu,\sigma^2/n)$.

2. הווקטור $(X_1-\bar{X},\dots,X_{n-1}-\bar{X},\bar{X})$ הוא טרנספורמציה לינארית הפיכה של $\vec{X}$, ולכן, לפי טענה **1.2.23**, התפלגותו רב־נורמלית. קל לחשב ש־$\mathrm{Cov}(\bar{X},X_i-\bar{X})=0$, ולפי טענה **1.2.26** נובע מכאן שהרכיב $\bar{X}$ בלתי תלוי בשאר הרכיבים. מכיוון שסכום המשתנים $X_i-\bar{X}$ הוא אפס, אפשר להציג את הסטטיסטי S כפונקציה של $X_1-\bar{X},\dots,X_{n-1}-\bar{X}$, ומכאן ש־$\bar{X}$ בלתי תלוי ב־S^2.

3. מכיוון ש־$\frac{X_i-\mu}{\sigma} \sim N(0,1)$ והם בלתי תלויים, $\sum\left(\frac{X_i-\mu}{\sigma}\right)^2 \sim \chi^2_n$. אבל לפי חישוב

$$\begin{aligned}\sum\left(\frac{X_i-\mu}{\sigma}\right)^2 &= \frac{1}{\sigma^2}\sum(X_i-\bar{X}+\bar{X}-\mu)^2\\ &= \frac{1}{\sigma^2}\left[\sum(X_i-\bar{X})^2+2(\bar{X}-\mu)\sum(X_i-\bar{X})+n(\bar{X}-\mu)^2\right]\\ &= \frac{n-1}{\sigma^2}S^2+\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2.\end{aligned}$$

לפי הסעיף הראשון, $\left(\frac{\bar{X}-\mu}{\sigma/\sqrt{n}}\right)^2 \sim \chi^2_1$. לפי טענה **1.1.6** (המשתמשת בפונקציה יוצרת המומנטים) אפשר לצמצם את הרכיב הזה מההתפלגות $\sum((X_i-\mu)/\sigma)^2 \sim \chi^2_n$, ולהסיק ש־$\frac{n-1}{\sigma^2}S^2 \sim \chi^2_{n-1}$.

4. **לפי שלושת הסעיפים הקודמים והגדרת ההתפלגות,**

$$Q = \frac{\bar{X}_n - \mu}{S/\sqrt{n}} = \frac{\frac{\bar{X}_n-\mu}{\sigma/\sqrt{n}}}{\sqrt{\frac{(n-1)S^2}{\sigma^2}/(n-1)}} \sim t_{n-1}.$$

□

מסקנה 1.2.56 נניח ש־$X_1,\ldots,X_n \sim N(\mu_1,\sigma^2)$ ו־$Y_1,\ldots,Y_m \sim N(\mu_2,\sigma^2)$ בלתי תלויים (עם אותה סטיית תקן). אז

$$\frac{\sum(X_i-\bar{X}_n)^2}{\sum(Y_i-\bar{Y}_m)^2} \sim F_{n-1,m-1}.$$

הוכחה. לפי המשפט, $\sum(X_i-\bar{X}_n)^2/\sigma^2 \sim \chi^2_{n-1}$ ו־$\sum(Y_i-\bar{Y}_m)^2/\sigma^2 \sim \chi^2_{m-1}$, ואלו כמובן משתנים בלתי תלויים. □

1.2.5 האומד ה־r־סימטרי

תרגיל 1.2.57 יהיו X_1, X_2 משתנים מקריים בלתי־תלויים, עם שונות σ^2. נגדיר S^2 כמקודם (עבור $n=2$). הראה ש־$S^2 = \frac{1}{2}(X_1-X_2)^2$.

תרגיל 1.2.58 תהי $h(x_1,\ldots,x_r)$ פונקציה סימטרית של r משתנים. יהיו $X_1,\ldots$ משתנים מקריים בלתי תלויים ושווי התפלגות. לכל n, נסמן

$$H_n = \frac{1}{\binom{n}{r}} \sum_{1\le i_1\le\cdots\le i_r\le n} h(X_{i_1},\ldots,X_{i_r}),$$

ממוצע הערכים של h תחת ההצבות השונות. אומד זה נקרא **אומד r־סימטרי**. נסמן ב־

$$\zeta_c = \operatorname{Cov}(h(X_{i_1},\ldots,X_{i_r}),\ h(X_{j_1},\ldots,X_{j_r}))$$

את השונות המשותפת של שני ערכים של הפונקציה, אם יש $c = |\{i_1,\ldots,i_r\}\cap\{j_1,\ldots,j_r\}|$ משתנים משותפים. (נניח שכל התוחלות מתכנסות). הוכח ש־

$$V(H_n) \sim \frac{r^2\zeta_1}{n}.$$

הדרכה. הצג את $\operatorname{Cov}(H_n,H_n)$ כצירוף לינארי של $\zeta_1,\ldots,\zeta_r$, והשאף $n\to\infty$.

סדרת האומדים H_n מתכנסת אל התוחלת $\gamma = \mathbf{E}(h(X_1,\dots,X_r))$, עם שגיאה נורמלית דועכת, במובן הבא.

הערה 1.2.59 (משפט Hoeffding, 1948) בתנאי תרגיל 1.2.58, הסדרה $\frac{H_n-\gamma}{\sqrt{r^2\zeta_1/n}}$ מתכנסת בהתפלגות אל ההתפלגות הנורמלית $N(0,1)$. *[Introduction to the theory of Nonparametric Statistics, Randles and Wolfe, 1979, Thm 3.3.13.]*

תרגיל 1.2.60 נגדיר H_n כבתרגיל 1.2.58 עבור הפונקציה $h(x_1,x_2)=\frac{1}{2}(x_1-x_2)^2$. הראה ש־$H_n = S_n^2$. **הדרכה**. הממוצע של $h(X_i,X_j)$ הוא תבנית ריבועית סימטרית. הסק ש־$\mathbf{V}(S^2)\sim\frac{1}{n}V(X_0^2)$, כאשר $X_0 = X_1 - \mathbf{E}(X_1)$.

תרגיל 1.2.61 (הכללה של תרגיל 1.2.60) יהיו $X_1,\dots,X_n$ משתנים מקריים בלתי־תלויים שווי התפלגות, עם $\mu=\mathbf{E}(X_i)$, $\sigma^2=\mathbf{V}(X_i)$, $\sigma_4=\mathbf{E}((X_i-\mu)^4)$.
אם $A\subseteq\{1,\dots,n\}$ קבוצה עם $|A|>1$, נסמן

$$S_A^2 = \frac{1}{|A|-1}\sum_{i\in A}(X_i - (\frac{1}{|A|}\sum_{j\in A}X_j))^2.$$

בפרט, $S^2_{\{1,\dots,n\}} = S^2$ כפי שהוגדר ב־(1.2).

1. לכל קבוצה A (כלעיל), $\mathbf{E}(S_A^2)=\sigma^2$.
2. נקבע $2\leq k\leq n$. הראה שהממוצע של כל האומדים S_A^2 על־פני כל הקבוצות עם $|A|=k$, שווה ל־S^2. **הדרכה**. כמו בתרגיל 1.2.60.
3. תהיינה $A,B\subseteq\{1,\dots,n\}$ קבוצות. אז

$$\mathrm{Cov}(S_A^2,S_B^2)=\frac{|A\cap B|}{|A|\cdot|B|}\left(\sigma_4-\left(1-\frac{2(|A\cap B|-1)}{(|A|-1)(|B|-1)}\right)\sigma^4\right).$$

בפרט, אם $|A|=|B|=r$ ו־$|A\cap B|=c$ אז $\zeta_c=\mathrm{Cov}(S_A^2,S_B^2)=\frac{c}{r^2}\left(\sigma_4-\left(1-\frac{2(c-1)}{(r-1)^2}\right)\sigma^4\right)$, ובפרט $\zeta_1=\mathrm{Cov}(S_A^2,S_B^2)=\frac{1}{r^2}\left(\sigma_4-\sigma^4\right)$.

4. אם $A\subseteq B$ אז $\mathrm{Cov}(S_A^2,S_B^2)=\frac{1}{|B|}\left(\sigma_4-\frac{|B|-3}{|B|-1}\sigma^4\right)$.
5. $\mathbf{V}(S^2)=\frac{1}{n}\left(\sigma_4-\frac{n-3}{n-1}\sigma^4\right)$.

פרק 2

אמידה

2.1 מודל, אוכלוסיה ומדגם

כפי שנסביר מיד, יש שני סוגי מקורות לנתונים סטטיסטיים: מודל ואוכלוסיה. גם המודל וגם האוכלוסיה אינם נגישים לנו ישירות, אלא רק דרך מדגם, ולכן הטיפול בשני המצבים דומה.

תוצאות של מדידה או ניסוי מדעי מגיעות לעתים קרובות מהתפלגות מוכרת, למשל התפלגות מעריכית או נורמלית. הכרת ההתפלגות הזו, כלומר ההנחה שהמשתנה מתפלג באופן מסויים ולא אחר נקראת **מודל**. בניית המודל (סטטיסטי או אחר) היא הצעד הראשון בשיטה המדעית. עם זאת, גם כאשר סוג ההתפלגות ידוע, בדרך כלל איננו יודעים את הפרמטרים לאשורם. למשל, מספר המטאורים הנצפים מדי דקה בלילה מסויים מתפלג פואסונית. המודל קובע שהמספר $X \sim \mathrm{Poi}(\lambda)$, כאשר λ אינו ידוע. תוצאות הניסוי, היינו ערכים $X_1, \dots, X_n$ של המשתנה המקרי שאת ההתפלגות שלו אנו דוגמים, נקראות **מדגם**.

הטבע אינו חושף את סודותיו בנקל, ולכן איננו יכולים לדעת כיצד בדיוק התקבלו המספרים האלה; המודל מציע משפחה של אפשרויות לתהליך כזה, ומשאיר בידי הסטטיסטיקאי את הצורך להכריע מי מהן היא הסבירה ביותר. **תורת האמידה** משתמשת בתוצאות המדגם כדי להעריך פרמטרים של ההתפלגות. הדוגמא השכיחה היא אמידה של התוחלת או השונות של ההתפלגות הנורמלית, אבל לגיטימי לנסות לאמוד גם דברים כמו p/q בהתפלגות בינומית, $e^{-\lambda}$ בהתפלגות פואסון, וכן הלאה.

מקור אפשרי אחר לנתונים הוא **אוכלוסיה** קבועה. למשל, אם רוצים ללמוד את הגובה של אזרח בוגר בישראל, עלינו לבחור אנשים מן האוכלוסיה ולמדוד את הגובה שלהם. גם במקרה כזה, ה**מדגם** הוא אוסף (בדרך כלל קטן) של ערכים שנאספו מן האוכלוסיה. גם במקרה כזה, הניתוח הסטטיסטי מבוסס על ההנחה שמקורם של נתוני האוכלוסיה הוא מודל, והמדגם מהווה בחירה של משתנים מקריים מתוך רשימה

קיימת. כמעט מכל בחינה, הניתוח זהה למקרה הנקי שבו המדגם נוצר ישירות מן המודל הסטטיסטי עצמו.

פרט לעניין אחד קטן. כשדוגמים מודל מדעי, סביר להניח שהתוצאות השונות אינן תלויות זו בזו. בדגימה מתוך אוכלוסיה סופית יש בעיה מהותית: אם הדגימה אקראית, יש סיכוי חיובי לחזרה על אותו ערך; אבל אם תתכן חזרה, נוצרת תלות מובנית בין הערכים שנדגמו (הם מן הסתם בלתי תלויים כשהם שונים זה מזה, אבל בוודאי תלויים אם דוגמים את אותו ערך פעמיים). מאידך, אם קובעים מראש שלא תתכן חזרה, ערכי המדגם נעשים תלויים (משום שאם דגמנו ערך גדול מהממוצע, והערך הבא חייב להיות שונה ממנו, הוא מוטה להיות נמוך מהממוצע). הפתרון שלנו לבעיה הזו יהיה להניח שהאוכלוסיה גדולה מספיק עד שהאפקט נעלם. עם זאת יש לדעת שדגימה מאוכלוסיה קטנה דורשת תשומת לב גם להיבט הזה.

בדגימה מאוכלוסיה יש בעיות נוספות. הניתוח המתמטי של המדגם, כמייצג של האוכלוסיה, מבוסס על ההנחה שהדגימה אינה מוטה (כלומר, לכל פרט יש סיכוי שווה להופיע במדגם) - אחרת יש לבצע התאמות שונות ומשונות. זהו אינו קורס בסטטיסטיקה מעשית, ולכן לא נעסוק בהרחבה בהטיות דגימה אפשריות. מכיוון שפטור בלא כלום אי אפשר, נסתפק ברמזים: לא כל הפרטים באוכלוסיה זמינים לצרכי דגימה (חסרי בית; תושבי חו"ל שיגיעו ליום הבחירות); תהליך הדגימה יוצר הטיות (בדגימה לפי מספרי טלפון יש לבעלי שני קווים סיכוי מוגבר להופיע; בדגימת אנשים בתחנת רכבת או בסניף דואר מגיעים לנוסעי רכבות ושולחי דואר); אנשים נוטים לשקר בסקרים (למשל בנושא שכר, הרגלים אישיים, אמונות, 'האם עברת הטרדה מינית') או להבין את השאלות אחרת מן הפרשן, כשעוסקים בהעדפות ודעות התשובה אינה יציבה (האם אתה אוהב כרוב? לפעמים.); מאגרי מידע מכילים ערכי דמה ושגיאות; ועוד ועוד.

2.2 אמידה נקודתית

המודל קובע ש־$X_1,\dots,X_n \sim F_\theta$, כאשר הפרמטר θ קובע את ההתפלגות המסויימת, אבל אינו ידוע. זה עשוי להיות פרמטר רב־ממדי, כמו בהתפלגות הנורמלית, הנקבעת על־ידי התוחלת והשונות שלה. לשם הכלליות, תהי $\tau = \tau(\theta)$ פונקציה של הפרמטר הזה. עלינו לאמוד את הערך של $\tau(\theta)$.

הניתוח המתמטי מאפשר לנו לדון בפונקציות של נתוני המדגם והפרמטר, שאותן נסמן ב־$s(X_1,\dots,X_n;\theta)$ או לשם הקיצור $s(X;\theta)$. מכיוון שהפרמטר θ אינו ידוע, האמידה מוכרחה להתבצע בלעדיו. פונקציה $T = t(X_1,\dots,X_n) = t(X)$ של נתוני המדגם, שאינה תלויה בפרמטר θ, נקראת **סטטיסטי**.

סטטיסטי שאנו משתמשים בו כדי לאמוד את $\tau(\theta)$ הוא **אומד נקודתי**, או סתם **אומד** (ההבדל בין סטטיסטי לאומד הוא דידקטי ולא מתמטי). כדי לציין זאת, נסמן אומד ל־θ ב־$\hat{\theta}$, ובאופן כללי אומד ל־$\tau = \tau(\theta)$ יהיה $\hat{\tau}$. ההבדל הטיפוגרפי הדק הזה מסתיר הבדל מהותי בין שני הביטויים: θ הוא פרמטר, ואילו $\hat{\theta}(X_1,\dots,X_n)$ הוא

אומד, התלוי בנתוני המדגם ואינו תלוי ב־θ.

לאחר הדגימה, כאשר מתקבלים הערכים $X_i = x_i$, האומד $\hat{\tau}(X_1, \dots, X_n)$ נעשה **אומדן** מספרי, $t(x_1, \dots, x_n)$.

דוגמא 2.2.1 נניח ש־$X_1, \dots, X_n \sim N(\mu, 1)$, והפרמטר שרוצים לאמוד הוא μ. הממוצע $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$ הוא סטטיסטי, שאפשר לראות בו אומד $\hat{\mu} = \bar{X}_n$ של μ. מדוע זה כך? משום שבהנתן נתוני המדגם, הממוצע מספק הערכה טובה לתוחלת. בדגימה מההתפלגות $N(3, 1)$ אנו מצפים לקבל ממוצע קרוב ל־3, ובדגימה מההתפלגות $N(12, 1)$ נקבע ערכים קרובים יותר ל־12. בכך שהפרמטר משפיע על נתוני המדגם, הוא מספק לנו אפשרות להפוך את המגמה, ולהסיק מנתוני המדגם על ערכו של הפרמטר.

זו הזדמנות להתבונן בסכום או הממוצע של משתנים בלתי תלויים, גם במקרה הכללי וגם עבור התפלגויות מוכרות.

תרגיל 2.2.2 אם $X_1, \dots, X_n$ משתנים בלתי תלויים בעלי תוחלת μ ושונות σ^2, אז $\mathbf{E}(\bar{X}_n) = \mu$ ו־$\mathbf{V}(\bar{X}_n) = \frac{\sigma^2}{n}$.

תרגיל 2.2.3 נניח ש־$X_1, \dots, X_n$ בלתי תלויים. נסמן $\bar{X}_n = \frac{1}{n}(X_1 + \cdots + X_n)$.

1. אם $X_i \sim b(p)$ אז $\bar{X}_n \sim \mathrm{Bin}(n, p)$.
2. אם $X_i \sim \mathrm{Poi}(\lambda)$ אז $\bar{X}_n \sim \mathrm{Poi}(n\lambda)$.
3. אם $X_i \sim \mathrm{Exp}(\lambda)$ אז $\bar{X}_n \sim \Gamma(n, \lambda)$.
4. אם $X_i \sim N(\mu, \sigma^2)$ אז $\bar{X}_n \sim N(\mu, \sigma^2/n)$.

2.2.1 שיטת המומנטים

איך בונים אומד מוצלח לפרמטר θ? שיטת המומנטים מציעה פתרון נאיבי למדי. מצד אחד, ההתפלגות של נתוני המדגם $X_1, \dots, X_n \sim F_\theta$ תלויה ב־θ, ולכן התוחלת שלה, $\mathbf{E}_\theta(X) = f(\theta)$, היא פונקציה של θ.

מצד שני, לפי החוק החלש של המספרים הגדולים, ממוצע המדגם שואף לתוחלת. לכן סביר להשוות $\bar{X}_n = f(\theta)$, לפתור את המשוואה, ולקבל את האומד $\hat{\theta} = f^{-1}(\bar{X}_n)$.

שימושים בשיטת המומנטים מיוחסים לגאוס, בסל וצ'ביצ'ב, אך קרל פירסון (Carl Pearson) היה בלי ספק הסניגור והפופולריזטור הגדול שלה. על השיטה נמתחה גם לא מעט ביקורת: ראו "Professor Karl Pearson and the Method of Moments", R.A.Fischer, Annals of Eugenics (!), June 1937.[1]

[1] קישור: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1469-1809.1937.tb02149.x.

דוגמא 2.2.4 1. נתבונן בהתפלגות הנורמלית $N(\mu,\sigma^2)$, כאשר σ^2 ידוע ומבקשים לאמוד את μ. הפרמטר שווה במקרה זה לתוחלת, ולכן שיטת המומנטים מציעה את האומד הטבעי $\hat{\mu}=\bar{X}_n$.

2. גם בהתפלגות פואסון $P(\lambda)$ התוחלת שווה לפרמטר λ, ולכן שיטת המומנטים מציעה גם כאן את האומד $\hat{\lambda}=\bar{X}_n$.

3. לעומת זאת, בהתפלגות גאומטרית $G(p)$ התוחלת היא $1/p$, ולכן האומד המתקבל משיטת המומנטים הוא $\hat{p}=1/\bar{X}_n$.

מדוע, אם כך, נקראת השיטה "שיטת המומנטים" ולא "שיטת הממוצע"? לפעמים רוצים לאמוד כמה פרמטרים בו־זמנית (היינו, לאמוד פרמטר רב־ממדי). במקרה כזה משוואה אחת, על התוחלת, אינה מספיקה, ויש להשוות את המומנטים הבאים. אכן, לפי החוק החלש של המספרים הגדולים, מומנטים של המדגם מתקרבים למומנטים של האוכלוסיה, ולכן ההשוואה מוצדקת גם במקרה הזה.

דוגמא 2.2.5 נציע אומד דו־ממדי לפרמטרים μ,σ בהתפלגות $N(\mu,\sigma^2)$. המומנטים הראשונים של ההתפלגות הזו הם μ ו־$\mathbf{E}(X^2)=\mu^2+\sigma^2$. השוואתם לממוצעים $\bar{X}_n$ ו־$\overline{X^2}_n=\frac{1}{n}(X_1^2+\cdots+X_n^2)$ מציעה את האומדים

$$\hat{\mu}=\bar{X}_n,\qquad \hat{\sigma}^2=\overline{X^2}_n-\bar{X}_n^2.$$

דוגמא 2.2.6 פירסון השתמש בשיטת המומנטים כדי לאמוד את הפרמטרים בהתפלגות המשלבת (בפרופורציות לא ידועות) שתי התפלגויות נורמליות (עם פרמטרים לא ידועים). כאן יש לאמוד חמישה פרמטרים, והשוואת חמשת המומנטים הראשונים מובילה למשוואה פולינומית ממעלה 9 בנעלם $\mu_1\mu_2$ (שממנה אפשר למצוא גם את שאר הפרמטרים). [“Tapas of Algebraic Statistics”, Notices of the AMS, **65**(8), 2018.][2].

תרגיל 2.2.7 מצא אומד לפי שיטת המומנטים לפרמטרים a,b בהתפלגות $U[a,b]$.
הדרכה. פתור את המשוואות $\bar{X}=\frac{1}{2}(\hat{a}+\hat{b})$ ו־$\frac{n-1}{n}S^2=\frac{1}{12}(\hat{b}-\hat{a})^2$.

דוגמא 2.2.8 אפשר לאמוד פרמטרים סבוכים יותר, לאו דווקא מומנטים. למשל, ההסתברויות $\mathcal{P}(X>a)$, או תוחלת כמו $\mathbf{E}(e^{X_1X_2})$.

[2] קישור: https://www.ams.org/journals/notices/201808/rnoti-p936.pdf.

2.2.2 אומדים חסרי הטיה

נניח ש־$\hat{\tau} = t(X_1,\dots,X_n) = T$ הוא אומד של $\tau(\theta)$. כפי שהתפלגות נתוני המדגם X_i תלויה ב־θ, גם ההתפלגות של האומד תלויה ב־θ. בפרט, התוחלת של T תלויה ב־θ (אבל לא בנתוני המדגם).

הגדרה 2.2.9 נניח ש־$X_1,\dots,X_n \sim F_\theta$, כאשר θ הוא פרמטר הקובע את ההתפלגות. סטטיסטי $T = t(X_1,\dots,X_n)$ הוא **אומד חסר הטיה** של $\tau(\theta)$ אם לכל ערך של θ מתקיים $\mathbf{E}(T) = \tau$.

אומד שאינו חסר הטיה הוא **אומד מוטה**.

תרגיל 2.2.10 בכל התפלגות, הממוצע $\frac{1}{n}(X_1+\cdots+X_n)$ הוא אומד חסר הטיה של התוחלת.

הערה 2.2.11 למרות ש־S^2 אומד חסר הטיה ל־σ^2, אין זה נכון ש־S הוא אומד חסר הטיה ל־σ.

תרגיל 2.2.12 בהתפלגות ברנולי $b(p)$, העזר בנתוני מדגם $X_1,\dots,X_n$ כדי למצוא אומדים חסרי הטיה ל־p ול־p^2. האם תוכל למצוא אומד חסר הטיה ל־$1/p$?

תרגיל 2.2.13 נניח ש־$X \sim \mathrm{Poi}(\lambda)$.

1. הראה ש־X הוא אומד חסר הטיה ל־λ.
2. מצא אומד חסר הטיה ל־$\lambda(\lambda-1)\cdots(\lambda-k+1)$.
3. בדוק שהפונקציה יוצרת המומנטים היא $M(t) = \mathbf{E}e^{tX} = e^{\lambda(e^t-1)}$. הסק שלכל $\alpha > 0$, α^X הוא אומד חסר הטיה ל־$e^{(\alpha-1)\lambda}$.
4. מצא אומד חסר הטיה ל־$e^{-\lambda}$. **הדרכה.** $T = \delta_{X,0}$ עונה על תנאי השאלה. כדי "לגלות" את הפתרון הזה, השאף $t \to -\infty$ בסעיף הקודם.
5. מצא אומד חסר הטיה ל־$e^{(\alpha-1)\lambda}$ לכל $\alpha \neq 0$. **הדרכה.** כתוב $T = f(X)$. מ־$e^{(\alpha-1)\lambda} = \mathbf{E}(T) = \sum \frac{e^{-\lambda}\lambda^n}{n!} f(n)$ הסק ש־$T = \alpha^X$ עובד גם במקרה זה.

דוגמא 2.2.14 הזמן הנדרש לתוכנית מחשב מסויימת להתחיל לרוץ מתפלג מעריכית עם תוחלת $30 + 12\theta$ שניות, כאשר θ הוא מספר הווירוסים המתרוצצים בזכרון. יהיה נאיבי מצידנו לצפות שזמן הריצה X יהיה בדיוק $30+12\theta$ שניות, ובכל זאת, $T = \frac{1}{12}(X-30)$ הוא אומד חסר הטיה למספר הווירוסים, משום שהתוחלת שלו היא θ.

תרגיל 2.2.15 במה עדיף הממוצע $\frac{1}{n}(X_1+\cdots+X_n)$ על־פני האומד X_2, אם ממילא שניהם חסרי הטיה?

תרגיל 2.2.16 כשרוצים לאמוד את מספר התומכים במפלגת 'כרוב וחסה', מניחים שיש באוכלוסיה שעור מסויים, p, של תומכים. כשבוחרים את המדגם האקראי, הסיכוי של כל משתתף במדגם להשתייך לקבוצת התומכים הוא p, ולכן תוצאות המדגם מתפלגות $X_1,\dots,X_n \sim b(p)$. בדוק שהממוצע הוא אומד חסר הטיה ל־p. מה שונותו? איך היא תלויה ב־n (האם כדאי לקחת מדגם גדול פי 100?) איך תלויה השונות ב־p, ומה משמעות התוצאה הזו בבואנו לאמוד את מספר התומכים במפלגה שעל גבול אחוז החסימה?

תרגיל 2.2.17 מה דעתך על פרסום (אפרת וייס, 3/11/2002, **Ynet**) שכותרתו "מחקר: 141,710 נשים מוכות בישראל"?

תרגיל 2.2.18 חשב את התוחלת של שונות המדגם $s^2 = \frac{1}{n}\sum(X_i - \bar{X})^2$, והסק שזהו אינו אומד חסר הטיה לשונות האוכלוסיה σ^2.

תרגיל 2.2.19 רוצים לאמוד את האחוז p של הילדים הפוחדים מחושך. חוששים שהילדים לא יענו על שאלה כזו בכנות, ולכן מבקשים מהם לפעול כדלקמן: כל ילד יטיל מטבע; אם יצא 'עץ' הוא יאמר את התשובה הנכונה, ואם יצא 'פלי' הוא יטיל מטבע שני, ויתן תשובה אקראית (כן או לא) על־פי התוצאה שקיבל שם. הילדים לא יחששו לשתף פעולה, משום שתשובה חיובית יכולה להתקבל גם כתוצאה מהטלת מטבע.

1. אם ילד עונה שהוא פוחד מחושך, מה הסיכוי שזה אכן כך?
2. כתוב את הסטטיסטי המתקבל באופן כזה, ובנה ממנו אומד חסר הטיה של הפרמטר המבוקש. (מהי שונות האומד? השווה אותה לשונות של האומד שהיה מתקבל אם אפשר היה לסמוך על התשובות כלשונן.)

רעיון 2.2.20 פיקוד העורף מבצע תרגיל לבדיקת צופרי האזעקה. אם יודיעו מראש שכל אזרח שאינו שומע את האזעקה מתבקש להתקשר ולהתריע על כך, קווי הטלפון בפיקוד יקרסו. נתח את האפשרות שרק שהאזרחים שמספר הזהות שלהם מסתיים ב־345 יתבקשו להודיע במקרה הצורך. מה דעתך על האלטרנטיבה, שלפיה רק אלו שמספר הזהות שלהם מתחיל ב־345 יתקשרו?

תרגיל 2.2.21 בבחירות לנשיאות בארצות הברית זוכה המועמד שצבר את מספר האלקטורים הגדול ביותר. ממדינה i ($i=1,\dots,k$, כאשר למשל $k=50$) מגיעים n_i אלקטורים, מתוך $N=\sum n_i$ אלקטורים בסך־הכל. הסיכוי של מועמד A לנצח במדינה i (על־פי

הסקרים המקומיים) הוא p_i בדיוק. נסמן $p^0 = p$ ו־$p^1 = (1-p)$. הראה שהסיכוי של A לנצח בסופו של דבר הוא $P = \sum_{(\epsilon_1,\dots,\epsilon_k)\in\{0,1\}^k : \sum \epsilon_i n_i > \frac{1}{2}N} (p_1^{\epsilon_1}\cdots p_k^{\epsilon_k})$.

קבוצת הווקטורים $(\epsilon_1,\dots,\epsilon_k)\in\{0,1\}^k$ המקיימים את התנאי $\sum \epsilon_i n_i > \frac{1}{2}N$ היא מסובכת למדי, וחישוב ישיר של P עשוי להיות בלתי אפשרי. הצע דרך מהירה לאמוד את הפרמטר P. **הדרכה.** בכל אחת מ־1000 החזרות על הניסוי, מטילים k מטבעות.

חוסר הטיה אסימפטוטי

בדרך כלל, אנו רואים את גודל המדגם n כקבוע. עם זאת, כאשר בוחרים אומד $T_n = T_n(X_1,\dots,X_n)$ לכל n, אפשר להתבונן גם בהתנהגות האסימפטוטית של הסדרה.

הגדרה 2.2.22 *סדרת האומדים T_n **היא עקבית** אם לכל θ,* $\lim \mathbf{E}((T_n-\tau(\theta))^2) = 0$.

הגדרה 2.2.23 *סדרת האומדים T_n **היא חסרת הטיה אסימפטוטית** אם לכל θ,* $\lim \mathbf{E}(T_n) = \tau(\theta)$.

(למשל, אם כל T_n חסר הטיה, קל־וחומר שהסדרה חסרת הטיה אסימפטוטית.)

טענה 2.2.24 סדרת האומדים T_n היא עקבית אם ורק אם היא חסרת הטיה אסימפטוטית וכן $\mathbf{V}(T_n)\to 0$.

דוגמא 2.2.25 בעקבות תרגיל 2.2.2, סדרת הממוצעים $\bar{X}_n$ היא סדרת אומדים עקבית לתוחלת.

טענה 2.2.26 S^2 הוא אומד עקבי ל־σ^2.

הוכחה. לפי משפט **1.2.55**, $\mathbf{V}(S^2) = \frac{2\sigma^4}{n-1}$. □

הגדרה 2.2.27 *סדרת האומדים T_n **היא עקבית־פשוטה** אם לכל $\epsilon > 0$,* $P(|T_n-\tau(\theta)| < \epsilon)\to 1$.

טענה 2.2.28 סדרת אומדים עקבית היא בפרט עקבית־פשוטה.

הוכחה. אי־שוויון מרקוב. □

2.2.3 השוואת אומדים

השגיאה באמידת τ על־ידי T היא המרחק $|T-\tau|$, ובאופן נגיש יותר לאנליזה מתמטית, הריבוע של המרחק הזה. מתברר שתוחלת השגיאה מתפרקת לשני רכיבים:

טענה 2.2.29 $\mathbf{E}((T-\tau)^2) = V(T) + (\mathbf{E}(T)-\tau)^2$.

הוכחה.

$$\begin{aligned}\mathbf{E}((T-\tau)^2) &= \mathbf{E}(T^2 - 2\tau T + \tau^2) \\ &= \mathbf{E}(T^2) - \mathbf{E}(T)^2 + \mathbf{E}(T)^2 - 2\tau\mathbf{E}(T) + \tau^2 \\ &= \mathbf{V}(T) + (\mathbf{E}(T)-\tau)^2.\end{aligned}$$

□

הגודל $(\mathbf{E}(T)-\tau)^2$ הוא **ההטיה** של T. היינו, תוחלת ריבוע השגיאה מורכבת משני רכיבים: השונות וההטיה. כדי למזער את השגיאה, עלינו לתקוף את שני הרכיבים האלה: לבחור T שההטיה שלו קטנה עד כמה שאפשר (חסר הטיה, אם יש כזה), ולהעדיף אומד בעל שונות קטנה עד כמה שאפשר.

המודל קובע ש־$X_1,\dots,X_n \sim F_\theta$. נניח ש־$T = t(X_1,\dots,X_n)$ הוא אומד חסר הטיה של θ. כדי להעריך את איכות האומד, ולבחור בין כמה אומדים אפשריים, מתבוננים בשונות $V_\theta(T)$. גם כאן חשוב להדגיש שהשונות תלויה ב־θ, ולכן יתכן שאומד מסויים יהיה בעל שונות נמוכה משל אומד אחר עבור ערכים מסויימים של הפרמטר, אבל בעל שונות גבוהה יותר במקומות אחרים.

הגדרה 2.2.30 *אם T, T' אומדים חסרי הטיה ל־$\tau(\theta)$, ולכל θ מתקיים $\mathbf{V}(T) \leq \mathbf{V}(T')$, אז T עדיף על T'.*

דוגמא 2.2.31 נניח שרוצים לאמוד את $1/p$ בהתפלגות מעריכית $\mathrm{G}(p)$. נתוני המדגם (הבלתי תלויים) הם $X_1,\dots,X_n$. כל אומד מהצורה $T = \sum a_i X_i$ עם $\sum a_i = 1$ הוא אומד חסר הטיה, משום ש־$\mathbf{E}(T) = \sum a_i \mathbf{E}(X_i) = (\sum a_i)/p = 1/p$. השונות של האומד הזה היא $\mathbf{V}(T) = \sum a_i^2 \mathbf{V}(X_i) = \frac{q}{p}\sum a_i^2$. אם כן, האומד עדיף ככל ש־$\sum a_i^2$ קטן יותר.

תרגיל 2.2.32 הוכח, בהמשך לדוגמא 2.2.31, שהאומד העדיף ביותר מהצורה $\sum a_i X_i$ הוא הממוצע.

בהנתן שני אומדים, יתכן שלכל אחד מהם שונות נמוכה עבור ערכים אחרים של הפרמטר, ואז אף אחד מהם אינו עדיף על משנהו.

דוגמא 2.2.33 הראה שהאומד $\bar{X}$ עדיף על $\frac{X_1+X_n}{2}$ בתור אומדים לתוחלת בהתפלגות $N(\mu,\sigma^2)$.

תרגיל 2.2.34 כל אחד מן המשתנים המקריים הבלתי תלויים $X_1,\dots,X_n$ הוא אומד חסר הטיה לגודל מסויים, עם שונות $\mathbf{E}(X_i)=\sigma_i^2$. צירוף ליניארי $Y=\alpha_1X_1+\cdots+\alpha_nX_n$ הוא אומד חסר הטיה אם $\alpha_1+\cdots+\alpha_n=1$. מצא את האומד בעל השונות הקטנה ביותר מצורה זו. הראה שעבורו $\mathbf{V}(Y)=(\sum\sigma_i^{-2})^{-1}$, וערך זה קטן מכל אחת מן השונויות σ_i^2.

תרגיל 2.2.35 נניח ש־$X_1,\dots,X_n\sim U[0,\theta]$.
הראה שסטטיסטיי הסדר המתוקנים $\frac{n+1}{k}X_{(k)}$ מהווים אומדים חסרי הטיה לפרמטר θ; השונות של $\frac{n+1}{k}X_{(k)}$ היא $\frac{n+1-k}{(n+2)k}\theta^2$. השווה את האומד $\frac{n+1}{n}X_{(n)}=\frac{n+1}{n}\max\{X_i\}$ לכפליים הממוצע $\frac{2}{n}(X_1+\cdots+X_n)$. איזה אומד עדיף?

תרגיל 2.2.36 בהמשך לתרגיל 2.2.13, נניח ש־$X_1,\dots,X_n\sim\mathrm{Poi}(\lambda)$ בלתי תלויים.

1. הראה ש־$X_1\cdots X_k$ הוא אומד חסר הטיה ל־λ^k.
2. הראה שכל האומדים הבאים של λ^2 הם חסרי הטיה: X_1X_2, $X_1^2-X_1$, $\bar{X}^2-\frac{1}{n}\bar{X}$, $\frac{1}{n}(\sum X_i)^2-\bar{X}$. איזה מהם עדיף?

2.2.4 אומד נראות מקסימלית

נניח שלפי המודל, $X_1,\dots,X_n\sim\mathrm{Poi}(\lambda)$, כאשר λ פרמטר לא ידוע. ההסתברות לקבל וקטור ערכים מסויים במדגם היא מכפלת ההסתברויות, $\prod\frac{e^{-\lambda}\lambda^{X_i}}{X_i!}=\frac{e^{-n\lambda}\lambda^{n\bar{X}}}{\prod X_i!}$. ערך זה נקרא **הנראות** של המדגם, והוא תלוי כמובן בפרמטר. אם הפרמטר קבוע, התוצאה סבירה יותר ככל שמתקבלת הסתברות גבוהה יותר. אבל אותו שיקול פועל גם בכיוון ההפוך: אם המדגם נתון, ככל שההסתברות גבוהה יותר כך סביר יותר ערכו של הפרמטר. הלך מחשבה זו מוביל אותנו לבחור את נקודת המקסימום של ההסתברות (שאותה אפשר למצוא על־ידי גזירה לפי הפרמטר והשוואה לאפס) כאומד $\hat{\lambda}=\bar{X}$.
ערכה של ההסתברות המתקבלת חשוב, אבל צריך לטפל בו בזהירות:

דוגמא 2.2.37 נניח שבמדגם מהתפלגות פואסון התקבלו הערכים $(2,0,7)$. ההסתברות לווקטור הזה היא $\frac{e^{-3\lambda}\lambda^9}{10080}$, והמקסימום המתקבל בערך $\hat{\lambda}=3$ הוא בערך $1/4000$. ההסתברות של המדגם הסביר ביותר עבור הפרמטר הזה, כלומר $(3,3,3)$, היא בערך אחוז אחד. מה משמעותה של ההסתברות הנמוכה של נתוני המדגם? היא עשויה לגרום לנו לפקפק במודל, או בנתוני המדגם עצמם; אבל עבור המודל הזה, ובנתוני המדגם האלה, ערכו של הפרמטר $\hat{\lambda}=3$ הוא הסביר ביותר.

כאשר המודל מצביע על התפלגות רציפה ההסתברות לכל ווקטור מדגם היא אפס. בכל זאת, אם נחליף את ההסתברות בצפיפות, נגיע למסקנות דומות. נסמן ב־$f(x;\theta)$ את פונקציית הצפיפות. **הנראות** של הערך $X=x$ היא הצפיפות בנקודה, כלומר $f(x;\theta)$. באופן כללי יותר, **הנראות** של המדגם $X_1,\dots,X_n$ היא הצפיפות המשותפת בנקודה, כלומר המכפלה

$$L(X;\theta)=\prod_{i=1}^{n} f(X_i;\theta).$$

אם עלינו להכריע איזה ערך של θ סביר יותר, נעדיף את זה שמציע נראות גבוהה יותר. הבחנה זו מוליכה להגדרה הבאה:

הגדרה 2.2.38 ***אומד נראות מקסימלית*** *ל־θ הוא אומד $\hat{\theta}=\hat{\theta}(X)$, שעבורו $L(X;\hat{\theta})$ מקסימלי.*

בדרך כלל אפשר למצוא אומד נראות מקסימלית על־ידי גזירה והשוואת הנגזרת לאפס. מכיוון שהנראות היא מכפלה, נוח יותר למקסם אותה אחרי לקיחת הלוגריתם; כלומר, לגזור

$$\frac{\partial}{\partial\theta}\log L(X;\theta)=\sum\frac{\partial}{\partial\theta}\log f(X_i;\theta)=\sum\frac{\frac{\partial f}{\partial\theta}(X_i;\theta)}{f(X_i;\theta)},$$

להשוות את הנגזרת לאפס, ולפתור את המשוואה

$$\sum\frac{\frac{\partial f}{\partial\theta}(X_i;\hat{\theta})}{f(X_i;\hat{\theta})}=0. \qquad (2.1)$$

דוגמא 2.2.39 נמצא אומד נראות מקסימלית לפרמטר בהתפלגות מעריכית. לפי המודל, $X_1,\dots,X_n\sim \mathrm{Exp}(\mu)$, עם צפיפות $f(x;\mu)=\frac{1}{\mu}e^{-x/\mu}$. הנראות של המדגם היא $L(X;\mu)=\prod\frac{1}{\mu}e^{-X_i/\mu}=\frac{1}{\mu^n}e^{-n\bar{X}/\mu}$. לכן $\log L(X;\mu)=-n\log\mu-n\bar{X}/\mu$; הנגזרת היא $\frac{\partial}{\partial\mu}L(X;\mu)=-\frac{n}{\mu}+\frac{n\bar{X}}{\mu^2}$. השוואה לאפס נותנת $\hat{\mu}=\bar{X}$.

הטכניקה מוצאת בדיוק באותו אופן אומד לפרמטר רב־ממדי:

דוגמא 2.2.40 נמצא אומד נראות מקסימלית לשני הפרמטרים של התפלגות נורמלית $X_1,\dots,X_n\sim N(\mu,\sigma^2)$. פונקציית הנראות היא $L(X;\mu,\sigma^2)=\prod\frac{1}{\sqrt{2\pi\sigma^2}}e^{-\frac{(\mu-X_i)^2}{2\sigma^2}}$. הלוגריתם הוא $\log L(X;\mu,\sigma^2)=-\frac{n}{2}\log(2\pi\sigma^2)-\sum\frac{(\mu-X_i)^2}{2\sigma^2}$. כדי למצוא את המקסימום, נגזור לפי μ ולפי σ^2 (!):

$$\frac{\partial}{\partial\mu}\log L(X;\mu,\sigma^2)=-\frac{1}{2\sigma^2}\sum 2(\mu-X_i)=\frac{n}{\sigma^2}(\bar{X}-\mu);$$

$$\frac{\partial}{\partial\sigma^2}\log L(X;\mu,\sigma^2) = -\frac{n}{2\sigma^2} + \sum\frac{(\mu-X_i)^2}{2\sigma^4}.$$

השוואת שתי הנגזרות לאפס מביאה לנקודת הקיצון

$$\hat{\mu} = \bar{X}, \qquad \hat{\sigma^2} = \frac{1}{n}\sum(\bar{X}-X_i)^2 = \frac{n-1}{n}S^2.$$

(אומד הנראות המקסימלית לשונות אינו חסר הטיה.)

תרגיל 2.2.41 מצא אומד נראות מקסימלית לתוחלת μ בהתפלגות $N(\mu,\alpha\mu^2)$, כאשר α פרמטר ידוע.

תרגיל 2.2.42 מצא אומד נראות מקסימלית להתפלגות ברנולי ולהתפלגות בינומית עם מספר ניסויים קבוע.

תרגיל 2.2.43 מצא אומד נראות מקסימלית לפרמטר p בהתפלגות שהצפיפות שלה $f(x;p) = 2pxe^{-px^2}$.

אנו רגילים למצוא נקודת מקסימום על־ידי גזירה והשוואה לאפס; זו הדרך הנכונה אם המקסימום נמצא בפנים תחום ההגדרה, אבל הוא עלול להמצא גם על השפה.

דוגמא 2.2.44 נמצא אומד נראות מקסימלית להתפלגות האחידה $U[0,\theta]$. פונקציית הצפיפות היא $\frac{1}{\theta}$ בתוך הקטע, ואפס מחוץ לו. כדי לא לאבד את המידע הזה, נכתוב את הצפיפות כך: $f(x;\theta) = \frac{1}{\theta}I_{[0,\theta]}(x)$, כאשר I_A היא הפונקציה המציינת של קבוצה A. כרגיל, הצפיפות היא המכפלה $\prod\frac{1}{\theta}I_{[0,\theta]}(X_i) = \frac{1}{\theta^n}I_{[0,\theta]}(\max X_i)$. הפונקציה הזו יורדת עם θ, ולכן נראה ש־θ יהיה קטן ככל האפשר, בכפוף לתנאי $\max X_i \leq \theta$. כלומר, $\hat{\theta} = \max X_i$. זהו אינו אומד חסר הטיה.

טענה 2.2.45 נניח ש־$\hat{\theta}$ הוא אומד נראות מקסימלית לפרמטר θ. אז לכל פונקציה τ, אומד הנראות המקסימלית ל־$\tau(\theta)$ הוא $\hat{\tau} = \tau(\hat{\theta})$.

הוכחה. האומד מתקבל בנקודת המקסימום של הנראות, ואין חשיבות לשאלה איזו פונקציה של הפרמטר מבקשים למקסם. □

בפרט, אומד הנראות המקסימלית של σ מתקבל מהוצאת שורש מן האומד של σ^2: $\hat{\sigma} = \sqrt{\frac{n-1}{n}}S$.

2.2.5 סטטיסטיים מספיקים ומספיקים במשותף

כפי שראינו בסעיף הקודם, ההסתברות של מדגם $(X_1,\dots,X_n)$ מהתפלגות פואסון היא $e^{-n\lambda}\lambda^{\sum X_i}(\prod X_i!)^{-1}$. כדי לחשב את המספר הזה, כפונקציה של λ, אין צורך בכל נתוני המדגם: מספיקים הערכים של $\sum X_i$ ו־$\prod X_i!$. יתרה מזו, ההשפעה של $\prod X_i!$ קבועה, ואינה תלויה בפרמטר. הסתכלות זו מוליכה להגדרה הבאה.

הגדרה 2.2.46 *סטטיסטי S הוא מספיק אם בהנתן S, צפיפות המדגם אינה תלויה בפרמטר. לחילופין, אם בהנתן S ההתפלגות של כל סטטיסטי אחר אינה תלויה בפרמטר.*

דוגמא 2.2.47 בהתפלגות פואסון, הסכום הוא סטטיסטי מספיק. אכן, התפלגות הסכום היא $\sum X_i \sim \mathrm{Poi}(n\lambda)$, ובהנתן הסכום $\sum X_i = s$, ההסתברות לווקטור המדגם $(X_1,\dots,X_n)$ היא

$$P((X_1,\dots,X_n)|\sum X_i = s) = \frac{e^{-n\lambda}\lambda^s/\prod X_i!}{e^{-n\lambda}(n\lambda)^s/s!} = \frac{s!}{n^s\prod X_i!},$$

וזה אינו תלוי בפרמטר λ.

תרגיל 2.2.48 מספר הכרטיסים הזוכים בחבילה בת מאה כרטיסי חיש־גד מתפלג פואסונית, עם פרמטר λ. דוגמים חמש חבילות, ומספר הכרטיסים הזוכים בסך־הכל הוא 20. מהו אומד הנראות המקסימלית ל־λ? הסבר את התפלגות וקטורי המדגם בהנתן מספר הזוכים הכולל (שאותה חישבנו בסוף דוגמא 2.2.47).

תרגיל 2.2.49 הראה שעבור ההתפלגות $N(\mu,\sigma^2)$ התלויה בפרמטר μ, כאשר σ קבוע וידוע, סכום המדגם הוא סטטיסטי מספיק.

תרגיל 2.2.50 נתבונן במדגם בלתי־תלוי מהתפלגות כלשהי, $(X_1,\dots,X_n)$. בהנתן הסכום $S = X_1+\dots+X_n$, התוחלת של כל X_i היא $\mathbf{E}(X_i|S) = S/n$, ואינה תלויה בפרמטרים של ההתפלגות. האם זה אומר שהסכום תמיד מספיק? **הדרכה**. לא: התוחלת של כל רכיב אינה תלויה בפרמטר, אבל התפלגות הווקטור – אולי כן.

הגדרנו מתי סטטיסטי בודד הוא מספיק. כמו במקרים אחרים, אפשר להכליל את ההגדרה: הסטטיסטיים $S_1,\dots,S_k$ **מספיקים במשותף** אם בהנתן $S_1,\dots,S_k$, צפיפות המדגם אינה תלויה בפרמטר. אפשר לקצר ולומר שהסטטיסטי (הרב ממדי) $(S_1,\dots,S_k)$ מספיק.

דוגמא 2.2.51 ערכי המדגם עצמן, $X_1,\dots,X_n$, מספיקים במשותף. גם סטטיסטיי הסדר $X_{(1)},\dots,X_{(n)}$ (הגדרה 1.2.37) הם מספיקים במשותף.

משפט 2.2.52 (משפט הפירוק) *הסטטיסטי S מספיק עבור ההתפלגות F_θ אם ורק אם אפשר לפרק את פונקציית הנראות בצורה*

$$L(X;\theta) = g(S;\theta)h(X),$$

כלומר למכפלה שבה אחד הגורמים אינו תלוי ב־θ, ואילו השני תלוי ב־X רק דרך S.

הוכחה. אם S מספיק, אז לפי ההנחה הצפיפות המותנית $f_{X|S}(X)$ אינה תלויה ב־θ. מכיוון ש־S תלוי רק בערכי X, הצפיפות המותנית היא

$$f_{X|S}(X) = \frac{L(X;\theta)}{f_S(S;\theta)}, \tag{2.2}$$

כאשר f_S היא הצפיפות (המשותפת) של S, ולכן $L(X;\theta) = f_{X|S}(X;\theta) \cdot f_S(S;\theta) = f_{X|S}(X)f_S(S;\theta)$, כפי שרצינו לפרק.

מאידך, אם נתון פירוק $L(X;\theta) = g(S;\theta)h(X)$, אפשר לחשב את הצפיפות של S לפי $f_S(s;\theta) = \int_{X:\ S=s} L(X;\theta)dx = g(s;\theta)\int_{X:\ S=s} h(X)dx$, ואז הצפיפות המותנית היא $f_{X|S=s}(X;\theta) = \frac{L(X;\theta)}{\int_{\{X:S=s\}} L(X;\theta)} = \frac{h(X)}{\int_{\{X:S=s\}} h(X)}$, ואינה תלויה ב־$\theta$. □

המשפט מספק שיטה לגילוי סטטיסטיים מספיקים: כל שעלינו לעשות הוא למצוא בפונקציית הנראות את הסטטיסטיים שלא ניתן לבודד מן הפרמטר.

תרגיל 2.2.53 הראה ש־$\max X_i$ ו־$\min X_i$ מספיקים במשותף להתפלגות אחידה $U[\theta_1,\theta_2]$, וגם להתפלגות אחידה $U[\theta,\theta+1]$.

דוגמא 2.2.54 הסטטיסטי $T = X_1X_2 + X_3$ אינו מספיק עבור מדגם (X_1,X_2,X_3) בהתפלגות ברנולי $b(p)$, בגלל ההתפלגות של X_1X_2 בהינתן T.

תרגיל 2.2.55 מצא סטטיסטי מספיק עבור ההתפלגות הנורמלית $N(\theta,1)$. מצא סטטיסטיים מספיקים במשותף עבור ההתפלגות הנורמלית $N(\mu,\sigma^2)$.

תרגיל 2.2.56 התבונן במדגם מתוך ההתפלגות $\begin{pmatrix} X \\ Y \end{pmatrix} \sim N\left(\begin{pmatrix} 0 \\ 0 \end{pmatrix}, \begin{pmatrix} 1 & \rho \\ \rho & 1 \end{pmatrix}\right)$. הראה שהסטטיסטיים $\sum X_i^2 + Y_i^2$ ו־$\sum X_iY_i$ מספיקים במשותף עבור הפרמטר ρ.

משפט 2.2.57 *יהי S סטטיסטי מספיק. אז אומד הנראות המקסימלית תלוי במדגם רק דרך התלות ב־S.*

הוכחה. משפט הפירוק. □

דוגמא 2.2.58 בהתפלגות $U[\alpha,\beta]$, אפשר לפרק את הנראות בצורה

$$L(X;\alpha,\beta)=\prod_i \frac{1}{\beta-\alpha}I_{[\alpha,\beta]}(X_i)=\frac{1}{(\beta-\alpha)^n}I_{[\alpha,\beta]}(\min X_i)I_{[\alpha,\beta]}(\max X_i),$$

ולכן הסטטיסטיים $\min X_i$ ו־$\max X_i$ מספיקים במשותף. אומד הנראות המקסימלי $\hat{\alpha}=\min X_i$ ו־$\hat{\beta}=\max X_i$ אכן תלוי רק בסטטיסטיים האלה, כפי שחוזה משפט 2.2.57. לעומת זאת, האומד של שיטת המומנטים (תרגיל 2.2.7) אינו מסתפק במינימום והמקסימום, ודורש את $\bar{X}$ ו־S^2.

נזכיר ש־$Y_1=\min X_i$ ו־$Y_n=\max X_i$ הם סטטיסטיי הסדר של $X_1,\dots,X_n$. הצפיפות המשותפת שלהם היא $f_{Y_1,Y_n}(y_1,y_n)=\frac{n(n-1)(y_n-y_1)^{n-2}}{(\beta-\alpha)^n}I_{[\alpha,\beta]}(y_1)I_{[\alpha,\beta]}(y_n)$, ולפי (2.2) הצפיפות המותנית היא $f_{\vec{X}|Y_1,Y_n}=\frac{1}{n(n-1)(Y_n-Y_1)^{n-2}}$, שאכן אינה תלויה ב־$\alpha,\beta$.

הערה 2.2.59 סטטיסטי מספיק שכל סטטיסטי מספיק אחר הוא פונקציה שלו, נקרא **סטטיסטי מינימלי**.

2.2.6 אינפורמציית פישר

יהי X משתנה מקרי בעל פונקציית צפיפות רציפה $f_X(x;\theta)$. נגדיר

$$W=W_\theta=\frac{\partial}{\partial\theta}\log f_X(X;\theta)=\frac{\frac{\partial}{\partial\theta}f_X(X;\theta)}{f_X(X;\theta)}.$$

בתור פונקציה של X זהו משתנה מקרי, התלוי ב־θ.

(באותו אופן אפשר להגדיר את W אם X הוא משתנה בדיד; לשם הזהירות נכתוב $p(x;\theta)=P(X=x)$, ואז נגדיר $W=W_\theta=\frac{\partial}{\partial\theta}\log p(X;\theta)=\frac{\frac{\partial}{\partial\theta}p(X;\theta)}{p(X;\theta)}$. בהמשך הפיתוח נתייחס למשתנים רציפים, אך הטיפול במשתנה מקרי בדיד הוא זהה.)

טענה 2.2.60 $\mathbf{E}(W)=0$ (לכל θ).

הוכחה. נחשב:

$$\begin{aligned}\mathbf{E}(W) &= \int f_X(x;\theta)\frac{\frac{\partial}{\partial\theta}f_X(x;\theta)}{f_X(x;\theta)}dx\\ &= \int\frac{\partial}{\partial\theta}f_X(x;\theta)dx=\frac{\partial}{\partial\theta}\int f_X(x;\theta)dx=\frac{\partial}{\partial\theta}1=0.\end{aligned}$$

□

טענה 2.2.61 $\mathbf{V}(W)=\mathbf{E}(W^2)=-\mathbf{E}(\frac{\partial}{\partial\theta}W)$.

הוכחה. השוויון $\mathbf{V}(W)=\mathbf{E}(W^2)-\mathbf{E}(W)^2=\mathbf{E}(W^2)$ ברור מטענה 2.2.60. עבור הטענה השניה, נחשב

$$\begin{aligned}
\mathbf{E}(\frac{\partial}{\partial\theta}W) &= \mathbf{E}(\frac{\partial}{\partial\theta}\frac{\frac{\partial}{\partial\theta}f_X(X;\theta)}{f_X(X;\theta)}) \\
&= \mathbf{E}(\frac{f_X(X;\theta)\frac{\partial^2}{\partial\theta^2}f_X(X;\theta)-(\frac{\partial}{\partial\theta}f_X(X;\theta))^2}{f_X(X;\theta)^2}) \\
&= \mathbf{E}(\frac{\frac{\partial^2}{\partial\theta^2}f_X(X;\theta)}{f_X(X;\theta)})-\mathbf{E}(W^2),
\end{aligned}$$

אבל $\mathbf{E}(\frac{\frac{\partial^2}{\partial\theta^2}f_X(X;\theta)}{f_X(X;\theta)})=\int f_X(x;\theta)\frac{\frac{\partial^2}{\partial\theta^2}f_X(x;\theta)}{f_X(x;\theta)}dx=\int\frac{\partial^2}{\partial\theta^2}f_X(x;\theta)dx=\frac{\partial^2}{\partial\theta^2}1=0$ בדיוק כמו בטענה 2.2.60. $\square$

הערה 2.2.62 1. בטענה 2.2.60 נעזרנו בכך ש־$\frac{\partial}{\partial\theta}\int\square dx=\int\frac{\partial}{\partial\theta}\square dx$. לעומת זאת, בדרך כלל $\mathbf{E}(\frac{\partial}{\partial\theta}\square)\neq\frac{\partial}{\partial\theta}\mathbf{E}(\square)$, שהרי התוחלת היא אינטגרל של הפונקציה המוכפלת בצפיפות של X.

2. החלפת סדר הגזירה והאינטגרציה נכונה בדרך כלל, אבל לא תמיד. לדוגמא, עבור ההתפלגות האחידה $U[0,\theta]$, פונקציית הצפיפות אינה גזירה בנקודת הקצה $x=\theta$, ואכן במקרה זה W אינו מוגדר. ההסבר האינטואיטיבי לכך הוא ש־W מודד את האינפורמציה שערכי X נותנים על הפרמטר θ. אם באותה נקודה $X=x$ יש ערכים של θ שעבורם הפונקציה מתאפסת, ואחרים שעבורם היא אינה מתאפסת, אז האינפורמציה שם היא כביכול אינסופית, והחישוב משתבש.

הגדרה 2.2.63 *הגודל* $I(\theta)=\mathbf{V}(W)$ *נקרא **אינפורמציית פישר** של הפרמטר* θ *(בדרך כלל קל יותר לחשב לפי* $I(\theta)=-\mathbf{E}(\frac{\partial}{\partial\theta}W)$*).*

דוגמא 2.2.64 בהתפלגות גמא $\Gamma(k,\lambda)$ עם k ידוע, פונקצית הצפיפות היא

$$f_X(x;\lambda)=\frac{1}{\Gamma(k)\lambda^k}x^{k-1}e^{-x/\lambda}.$$

לכן

$$\log f_X(x;\lambda)=-\log\Gamma(k)-k\log\lambda+(k-1)\log x-x/\lambda,$$

ו־

$$W=\frac{\partial}{\partial\lambda}\log f_X(x;\lambda)=\frac{X-k\lambda}{\lambda^2}.$$

אם כך $\frac{\partial}{\partial\lambda}W=\frac{\partial}{\partial\lambda}(\frac{X}{\lambda^2}-\frac{k}{\lambda})=-\frac{2X}{\lambda^3}+\frac{k}{\lambda^2}$, ומכיוון ש־$\mathbf{E}(X)=k\lambda$, $I(\lambda)=\frac{k}{\lambda^2}$.

דוגמא 2.2.65 בהתפלגות פואסון,

$$W = \frac{\partial}{\partial\lambda}\log\frac{e^{-\lambda}\lambda^X}{X!} = \frac{\partial}{\partial\lambda}(-\lambda + X\log\lambda - \log X!) = \frac{X-\lambda}{\lambda}.$$

אם כך $\frac{\partial}{\partial\lambda}W = -\frac{X}{\lambda^2}$, ולכן $I(\lambda) = -\mathbf{E}(\frac{\partial}{\partial\lambda}W) = \frac{1}{\lambda}$.

אינפורמציית פישר מוגדרת עבור הפרמטר θ בהתייחס למשתנה בודד X. אפשר להכליל את ההגדרה, ולקחת $I_n(\theta) = \mathbf{V}(W_n)$, כאשר $W_n = \frac{\partial}{\partial\theta}L(X_1,\dots,X_n;\theta)$; כלומר, הנראות של הווקטור $(X_1,\dots,X_n)$ מחליפה את הנראות (שהיא הצפיפות) של דגימה בודדת.

טענה 2.2.66 אם $X_1,\dots,X_n$ בלתי תלויים, אז $I_n(\theta) = nI(\theta)$.

הוכחה. הרי $\log L(X_1,\dots,X_n;\theta) = \sum\log f_X(X_i;\theta)$. □

(בהמשך נניח תמיד שערכי המדגם בלתי תלויים זה בזה, אבל יש לדעת שאם אין הדבר כך, ההכללה הנכונה לאינפורמציה היא $I_n(\theta)$ ולא $nI(\theta)$.)

מן האינפורמציה של θ אפשר לעבור בקלות לאינפורמציה של פרמטרים התלויים ב־θ:

טענה 2.2.67 לכל פונקציה הפיכה τ, $I(\tau(\theta)) = \frac{1}{\tau'(\theta)^2}I(\theta)$.

הוכחה. סימנו $W_\theta = \frac{\partial}{\partial\theta}\log f_X(X;\theta)$, ואז $I(\theta) = \mathbf{E}(W_\theta^2)$. עבור $\tau(\theta)$ נקבל לפי כלל השרשרת

$$W_\tau = \frac{\partial}{\partial\tau}\log f_X(X;\theta) = \frac{\partial\theta}{\partial\tau}\frac{\partial}{\partial\theta}\log f_X(X;\theta) = \frac{1}{\tau'(\theta)}W_\theta,$$

ולכן

$$I(\tau(\theta)) = \mathbf{E}(W_\tau^2) = \frac{1}{\tau'(\theta)^2}\mathbf{E}(W_\theta^2) = \frac{1}{\tau'(\theta)^2}I(\theta).$$

□

2.2.7 אי־שוויון קרמר־ראו

אנחנו מעדיפים אומד של $\tau(\theta)$ ככל שהשונות שלו קטנה יותר. עד כמה אפשר לשפר את השונות הזו? מתברר שיש חסם תחתון מוחלט על השונות של אומדים, וקריטריון הקובע מתי מתקבל החסם הזה.

הערה 2.2.68 אם יש פירוק

$$W_n = \frac{\partial}{\partial\theta}\log L(X;\theta) = k(\theta,n)[T(X)-\tau(\theta)],$$

אז T אומד חסר הטיה ל־$\tau(\theta)$. אכן, לפי טענה 2.2.60

$$0 = \mathbf{E}(W_n) = k(\theta,n)\mathbf{E}(T-\tau(\theta))$$

ולכן $\mathbf{E}(T)=\tau(\theta)$.

משפט 2.2.69 (אי־שוויון קרמר־ראו) *1. לכל אומד חסר הטיה T של $\tau(\theta)$ מתקיים*

$$\mathbf{V}(T) \geq \frac{1}{I_n(\tau)} = \frac{\tau'(\theta)^2}{nI(\theta)}.$$

2. קיים שוויון בסעיף 1 אם ורק אם יש פירוק

$$W_n = \frac{\partial}{\partial\theta}\log L(X;\theta) = k(\theta,n)[T(X)-\tau(\theta)].$$

הוכחה. מכיוון ש־T הוא אומד חסר הטיה של $\tau(\theta)$,

$$\begin{aligned}
\mathbf{E}(TW_\tau) &= \int L(x_1,\ldots,x_n;\theta)T(x_1,\ldots,x_n)W_\tau(x_1,\ldots,x_n;\theta)dx \\
&= \int T(x_1,\ldots,x_n)\frac{\partial}{\partial\tau}L(x_1,\ldots,x_n;\theta)dx \\
&= \frac{\partial}{\partial\tau}\int T(x_1,\ldots,x_n)L(x_1,\ldots,x_n;\theta)dx \\
&= \frac{\partial}{\partial\tau}\mathbf{E}(T) = \frac{\partial}{\partial\tau}\tau = 1.
\end{aligned}$$

מכיוון ש־$\mathbf{E}(W_\tau)=0$ (טענה 2.2.60), מקבלים

$$\mathrm{Cov}(T,W_\tau) = \mathbf{E}(TW_\tau) - \mathbf{E}(T)\mathbf{E}(W_\tau) = 1.$$

טענת הסעיף הראשון נובעת מיד מאי־שוויון קושי־שוורץ,

$$1 = \mathrm{Cov}(T,W_\tau)^2 \leq \mathbf{V}(T)\mathbf{V}(W_\tau) = \mathbf{V}(T)\cdot I_n(\tau).$$

יש כאן שוויון אם ורק אם W תלוי לינארית ב־T, כלומר $W=k(T-\beta)$ עבור קבועים k,β. אבל אז $0=\mathbf{E}(W)=k(\mathbf{E}(T)-\beta)$, כלומר $\beta=\mathbf{E}(T)=\tau$ ו־$W=k(T-\tau)$; להשלמת ההוכחה יש להבחין שהקבוע k הוא קבוע ביחס למשתנים המקריים, אבל מותר לו להיות תלוי בפרמטר θ, וכמובן בגודל המדגם n. □

דוגמא 2.2.70 נתבונן בהתפלגות המעריכית, עם הצפיפות $f_X(x;\mu) = \frac{1}{\mu}e^{-x/\mu}$. זוהי התפלגות $\Gamma(1,\lambda)$, ובדוגמא 2.2.64 ראינו שעבור ההתפלגות הזו $I(\lambda) = \frac{1}{\lambda^2}$. לפי אי־שוויון קרמר־ראו, לכל אומד חסר הטיה T של λ יש שונות $\mathbf{V}(T) \geq \frac{\lambda^2}{n}$.

תרגיל 2.2.71
1. מצא את חסם קרמר־ראו על אומדים חסרי הטיה לתוחלת בהתפלגות פואסון.
2. מצא את חסם קרמר־ראו על $e^{-\lambda}$ בהתפלגות פואסון $\mathrm{Poi}(\lambda)$.

משפט 2.2.72 ***נניח ש־$\hat{\theta}$ הוא אומד נראות מקסימלית של θ, המאפס את הנגזרת. אז האומד חסר ההטיה היחיד של $\tau(\theta)$ ששונותו שווה לחסם קרמר־ראו (אם יש כזה), הוא $\tau(\hat{\theta})$.***

הוכחה. **לפי ההנחה, $\frac{\partial}{\partial\theta}\log L(X;\hat{\theta}) = 0$. אם T הוא אומד חסר הטיה לפונקציה $\tau(\theta)$ שהשונות שלו מקיימת את חסם קרמר־ראו, אז לפי המשפט $\frac{\partial}{\partial\theta}\log L(X;\theta) = k(\theta,n)[T(X)-\tau(\theta)]$, ולכן $k(\hat{\theta},n)[T(X)-\tau(\hat{\theta})] = \frac{\partial}{\partial\theta}\log L(X;\hat{\theta}) = 0$. מכאן ש־$T = \tau(\hat{\theta})$.** □

2.2.8 אומדים חסרי הטיה בעלי שונות מינימלית במידה שווה

הגדרה 2.2.73 אומד חסר הטיה שהוא עדיף על כל אומד חסר הטיה אחר (התלוי במדגם $(X_1,\dots,X_n)$), נקרא **אומד חסר הטיה בעל שונות מינימלית במידה שווה** (ובקיצור, על־פי ראשי התיבות של השם האנגלי, $UMVUE$; זה עדיף על אחהבשמבמ"ש...).

אומד כזה הוא "הגביע הקדוש" של תורת האמידה, וממבט ראשון לא ברור שהוא קיים בכלל. מתברר למרבה ההפתעה שעבור מודלים רבים, קיים UMVUE; בהמשך נראה איך אפשר לבנות אותו. התרגיל הבא מראה שאם קיים UMVUE, אז הוא יחיד.

תרגיל 2.2.74 נניח ש־ T_0,T_1 הם אומדים חסרי הטיה לפרמטר θ, עם שונויות $V(T_i) = \sigma_i^2 > 0$ ושונות משותפת $\mathrm{Cov}(T_0,T_1) = \rho\sigma_0\sigma_1$. כמובן, $-1 \leq \rho \leq 1$.

1. כל $T_\alpha = \alpha T_1 + (1-\alpha)T_0$ הוא אומד חסר הטיה של θ ($\alpha \in \mathbb{R}$).
2. חשב במפורש את הפונקציה $f(\alpha) = V(T_\alpha)$ [בדוק ש־ $f(\alpha) = \sigma_\alpha^2$ עבור $\alpha = 0,1$].
3. אם $\rho = 1$ אז $T_0 = T_1$ (בהסתברות 1). נניח מעתה ש־$\rho < 1$.
4. הפונקציה f היא פולינום ריבועי עם מקדם מוביל חיובי. נסמן ב־ α^* את נקודת הקיצון של f. הסק שזו נקודת מינימום ולא מקסימום.
5. הראה ש־$\alpha^* = 0$ אם ורק אם $\rho = \frac{\sigma_0}{\sigma_1}$ (ובפרט $\sigma_0^2 < \sigma_1^2$). בדומה לזה, $\alpha^* = 1$ אם ורק אם $\rho = \frac{\sigma_1}{\sigma_0}$ (ובפרט $\sigma_1^2 < \sigma_0^2$).

6. אם $\sigma_0^2=\sigma_1^2$, אז T_{α^*} הוא אומד עדיף על שני האומדים T_0,T_1 (אגב, במקרה זה $\alpha^*=1/2$).

7. הסק: אם T_0,T_1 הם אומדים חסרי הטיה בעלי שונות מינימלית במדה שווה, אז $T_0=T_1$ (בהסתברות 1).

הערה 2.2.75 אם לאומד חסר הטיה T יש שונות השווה לחסם קרמר־ראו CR_θ, אז T הוא האומד האולטימטיבי - $UMVUE$. זאת משום שלכל אומד חסר הטיה T', $\mathbf{V}(T')\geq CR_\theta=\mathbf{V}(T)$.

החלק השני במשפט קרמר־ראו נותן שיטה אפקטיבית למצוא UMVUE לאיזושהי פונקציה $\tau(\theta)$ של θ, אם כי אין לנו דרך לשלוט ב־τ עצמה.

דוגמא 2.2.76 נמצא UMVUE לפרמטר של התפלגות המעריכית, אם קיים כזה. לפי דוגמא 2.2.64, $W_n=\sum\frac{X_i-\lambda}{\lambda}=\frac{1}{\lambda}\sum X_i-n=\frac{n}{\lambda}(\bar{X}_n-\lambda)$. מכאן ש־$T=\bar{X}$ הוא אומד חסר הטיה ל־λ המקיים את חסם קרמר־ראו, ואין אף פונקציה אחרת של λ (עד כדי כפל בקבוע העשוי להיות תלוי ב־n), שיש לה אומד כזה.

סיכום ביניים: משפט הפירוק מוצא עברונו את הסטטיסטיים המספיקים. משפט קרמר־ראו מוצא UMVUE עבור איזושהי פונקציה. אומדים חסרי הטיה אפשר למצוא בשיטות של תרגיל 2.2.13.

המשפט החשוב הבא מראה שאם ידוע לנו סטטיסטי מספיק, אז אפשר לקבל מאומד חסר הטיה כלשהו אומד חסר הטיה טוב יותר. במשפט 2.2.86 נראה שבתנאים מסויימים אומד זה הוא לא פחות מאשר UMVUE.

משפט 2.2.77 (משפט ראו־בלקוול) *יהי T אומד חסר הטיה ל־$\tau(\theta)$, ויהי S סטטיסטי מספיק. ניקח $T'=\mathbf{E}(T|S)$. אז:*

1. *T' סטטיסטי;*
2. *T' הוא אומד חסר הטיה;*
3. *$\mathbf{V}(T')\leq\mathbf{V}(T)$, ושוויון מתקיים רק אם $T=T'$ בהסתברות אחת.*

הוכחה. 1. ההתפלגות של T תלויה ב־θ, ולכן כך גם התוחלת שלו; אבל בהנתן S, שהוא מספיק, ההתפלגות אינה תלויה עוד ב־θ, ולכן T' הוא פונקציה של S בלבד, כלומר סטטיסטי.

2. $\mathbf{E}(T')=\mathbf{E}(\mathbf{E}(T|S))=\mathbf{E}(T)=\tau(\theta)$.

3. $\mathbf{V}(T)=\mathbf{V}(\mathbf{E}(T|S))+\mathbf{E}(\mathbf{V}(T|S))=\mathbf{V}(T')+\mathbf{E}(\mathbf{V}(T|S))\geq\mathbf{V}(T')$, וכדי שיתקיים שוויון דרוש $\mathbf{V}(T|S)=0$ לכל ערך של θ, כלומר T הוא פונקציה של S, וממילא $T'=T$.

□

הערה 2.2.78 אם נחזור שנית על הבניה של משפט ראו־בלקוול, נקבל $T'' = T'$. אכן, T' הוא פונקציה של S, ולכן $\mathbf{E}(T'|S) = T'$.

דוגמא 2.2.79 בדוגמא 2.2.47 ראינו שהסכום $S = X_1 + \cdots + X_n$ הוא סטטיסטי מספיק עבור התפלגות פואסון $\mathrm{Poi}(\lambda)$. (למעשה $X_1|S \sim \mathrm{Bin}(S, \frac{1}{n})$).

1. המשתנה $T = X_1$ הוא אומד חסר הטיה לפרמטר λ. התוחלת המותנית $T' = \mathbf{E}(X_1|S) = \frac{1}{n}S$ גם היא אומד חסר הטיה, עם שונות $\mathbf{V}(T') = \frac{\lambda}{n}$ לעומת השונות של X_1 שהיא $\mathbf{V}(X_1) = \lambda$.

2. נקבע פרמטר $\alpha \neq -1$. הסטטיסטי $T = (\alpha+1)^{X_1}$ הוא אומד חסר הטיה ל־$e^{\alpha\lambda}$ (תרגיל 2.2.13). במקרה זה $T' = \mathbf{E}((\alpha+1)^{X_1}|S) = (1+\frac{\alpha}{n})^S$ גם הוא אומד חסר הטיה, התלוי רק ב־S.

   השונות של T היא $\mathbf{V}(T) = e^{2\alpha\lambda}(e^{\alpha^2\lambda}-1)$, בעוד שהשונות של T' היא $\mathbf{V}(T') = e^{2\alpha\lambda}(e^{\alpha^2\lambda/n}-1)$.

תרגיל 2.2.80 המקסימום $S = \max X_i$ הוא סטטיסטי מספיק עבור ההתפלגות האחידה $U[0,\theta]$. האומד $T = 2X_1$ הוא חסר הטיה עבור θ. חשב את האומד $T' = \mathbf{E}(T|S)$ של משפט ראו־בלקוול, והשווה את השונויות.

תרגיל 2.2.81 $T = X_1X_2$ הוא אומד חסר הטיה ל־p^2 בהתפלגות ברנולי. העזר באומד המספיק $S = X_1 + \cdots + X_n$ כדי למצוא אומד חסר הטיה עדיף.

2.2.9 סטטיסטיים שלמים

הגדרה 2.2.82 *סטטיסטי S הוא **שלם** אם אין אף פונקציה שלו (פרט לפונקציית האפס) שהתוחלת שלה היא זהותית אפס.*

דוגמא 2.2.83 בהתפלגות ברנולי, $S = X_1 + \cdots + X_n$ שלם, ו־$X_1 - X_2$ אינו שלם.

דוגמא 2.2.84 המקסימום $\max X_i$ הוא סטטיסטי שלם בהתפלגות האחידה $U[0,\theta]$.

טענה 2.2.85 אם S סטטיסטי שלם, אז יש לכל היותר אומד חסר הטיה אחד ל־$\tau(\theta)$ שהוא פונקציה של S.

הוכחה. אחרת $\mathbf{E}(T_1 - T_2) = 0$. □

משפט 2.2.86 (משפט להמן־שפה) *אם S מספיק ושלם, ו־ $T^* = t(S)$ הוא אומד חסר הטיה ל־ $\tau(\theta)$, אז הוא* UMVUE.

הוכחה. יהי T אומד חסר הטיה כלשהו. לפי משפט ראו־בלקוול, $T' = \mathbf{E}(T|S)$ הוא אומד חסר הטיה שהוא פונקציה של S. לפי היחידות (טענה 2.2.85), $T' = T^*$, ולכן $\mathbf{V}(T^*) = \mathbf{V}(T') \leq \mathbf{V}(T)$. □

לסיכום, מסטטיסטי מספיק ושלם, ואומד חסר הטיה, נוצר UMVUE:

טענה 2.2.87 נניח שלמשפחת התפלגויות F_θ יש סטטיסטי מספיק ושלם S. אם יש אומד חסר הטיה T של $\tau(\theta)$, אז $T' = \mathbf{E}(T|S)$ הוא UMVUE עבור $\tau(\theta)$.

תרגיל 2.2.88 מצא UMVUE בהתפלגות מעריכית ל־λ ול־$1/\lambda$.

תרגיל 2.2.89 מצא UMVUE ל־$e^{-\lambda}$ בהתפלגות פואסון.

תרגיל 2.2.90 נניח שמספר התקלות X במוצר הוא $0,1,2$ ביחסים $1 : \theta : \theta^2$, כאשר θ הוא פרמטר לא ידוע. מצא UMVUE להסתברות $P(X=0)$. **הדרכה.** $S = \sum X_i$ מספיק, $T = \delta_{X_1,0}$ הוא חסר־הטיה, ו־$E(T|S) = \frac{F_{n-1}(S)}{F_n(S)}$ כאשר $F_n(s)$ מוגדר כמספר הדרכים לפתור $x_1 + \cdots + x_n = s$ עם $x_i \in \{0,1,2\}$.

2.2.10 משפחות מעריכיות

משפחת התפלגויות F_θ נקראת **משפחה מעריכית** אם הצפיפות שלה היא מהצורה

$$f(x;\theta) = a(\theta)b(x)e^{c(\theta)d(x)}.$$

דוגמא 2.2.91
1. משפחת ההתפלגויות הנורמליות עם שונות ידועה היא מעריכית.
2. משפחת התפלגויות גמא היא מעריכית.
3. משפחת התפלגויות פואסון היא מעריכית, כשההסתברות מחליפה את הצפיפות.

משפט 2.2.92 *במשפחה מעריכית של התפלגויות,* $S = \frac{1}{n}\sum d(X_i)$ *הוא סטטיסטי מספיק, שלם ומינימלי. בנוסף, הוא מהווה* UMVUE *עבור הפרמטר* $\tau(\theta) = -\frac{a'(\theta)}{a(\theta)c'(\theta)}$.

הוכחה. נוכיח שהסטטיסטי מספיק (אבל לא שהוא שלם ומינימלי). הנראות של משפחה מעריכית היא

$$L(X;\theta) = a(\theta)^n \cdot \prod_{i=1}^{n} b(X_i) \cdot e^{c(\theta)\sum d(X_i)}. \tag{2.3}$$

לפי משפט הפירוק 2.2.52, S מספיק מכיוון שיש פירוק של הנראות

$$a(\theta)^n e^{c(\theta)\sum d(X_i)} \cdot \prod_{i=1}^{n} b(X_i)$$

שבו הרכיב התלוי בפרמטר זקוק לנתוני המדגם רק דרך $\sum d(X_i)$.
הטענה על כך ש־S הוא UMVUE נובעת ממשפט קרמר־ראו, בעזרת הפירוק

$$\frac{\partial}{\partial\theta}\log L(X;\theta) = nc'(\theta)\left(\frac{1}{n}\sum d(X_i) + \frac{a'(\theta)}{a(\theta)c'(\theta)}\right).$$

□

תרגיל 2.2.93 הכלל את המשפט למשפחה שהצפיפות שלה היא מהצורה

$$f(x;\theta) = a(\theta)b(x)e^{c_1(\theta)d_1(x)+\cdots+c_t(\theta)d_t(x)}.$$

2.3 רווחי סמך

באמידה נקודתית אנחנו מנסים לקלוע אל הערך המדוייק של הפרמטר, ומודדים את ההצלחה בתוחלת של ריבוע השגיאה. אמידת רווח היא גישה אחרת, שבה מנסים להציע רווח שבו שוכן הפרמטר, ומודדים את ההצלחה בסיכויים לכך שהרווח עומד במשימה.

הגדרה 2.3.1 *נניח ש־$X_1,\dots,X_n \sim F_\theta$, כאשר F_θ הוא מודל התלוי בפרמטר θ. זוג סדור של סטטיסטים T_1, T_2 נקרא **רווח סמך בעל רמת מובהקות** α עבור הפרמטר, אם הסיכוי למאורע $T_1 < \theta < T_2$ הוא $1-\alpha$.*

במלים אחרות, α הוא הסיכוי לכך שהפרמטר לא יפול ברווח. גודל טיפוסי ל־α הוא 0.05 או 0.01. ככל שנבחר ערך קטן יותר, נצטרך לשלם בהגדלה של הרווח, השקולה לטענה חלשה יותר על הפרמטר.

לאחר ביצוע הליך הדגימה בפועל, המשתנים המקריים $X_1,\dots,X_n$ מקבלים ערכים מספריים, וכך הופכים גם קצות הקטע T_1, T_2 למספרים, נאמר t_1, t_2. זו מכשלה נפוצה לומר שבמקרה זה, "הסיכוי לכך שהפרמטר θ נמצא בין t_1 ל־t_2 הוא $1-\alpha$". ניסוח זה שגוי בתכלית, משום שלפרמטר אין התפלגות - הוא מספר (וגם אם הוא נקבע על־פי התפלגות כלשהי, התפלגות זו אינה נלקחת בחשבון בחישוב הרווח). אם כך, הסיכוי לכך שהפרמטר יהיה בין שני מספרים הוא או אפס או אחד (גם אם איננו יודעים איזו אפשרות היא הנכונה). א־פריורי, הסיכוי לכך שהפרמטר יהיה שייך לקטע הוא בדיוק $1-\alpha$; אבל לאחר מעשה, גורל הניסוי כבר נגזר, לשבט (בסיכוי α) או לחסד, והוא איננו מאורע הסתברותי.

2.3.1 שיטת הכמות הצירית

השיטה הפשוטה ביותר לבניית רווח סמך היא **שיטת הכמות הצירית**, שלפיה אנו מוצאים פונקציה של המדגם ושל הפרמטר, שהתפלגותה אינה תלויה בפרמטר.

נניח ש־$X_1,\dots,X_n \sim F_\theta$. הפונקציה $Q = q(\vec{X},\theta)$ היא **כמות צירית** אם ההתפלגות של Q אינה תלויה בפרמטר θ. מכיוון שההתפלגות קבועה, יש q_1,q_2 כך ש־$P(q_1 < Q < q_2) = 1-\alpha$. את רווח הסמך אפשר למצוא על־ידי חילוץ θ מאי־השוויון $q_1 < Q < q_2$.

הערה 2.3.2 1. שיטת הכמות הצירית מספקת אינסוף רווחי סמך בעלי רמת מובהקות נתונה: לכל q_1 (קטן מספיק) יש q_2 שעבורו הסיכוי ל־$q_1 < Q < q_2$ שווה ל־$1-\alpha$.

2. כמקרי קיצון, יש רווחי סמך $q < Q$ ו־$Q < q'$.

3. מבין כל אלה, מקובל להעדיף את הרווח שבו הסיכויים ליפול משני קצות הקטע שווים (ל־$\alpha/2$).

4. גישה אחרת מבקשת למזער את אורך הקטע, $q_2 - q_1$; גזירה מראה שזה שקול לפתרון המשוואה $f_Q(q_1) = f_Q(q_2)$.

דוגמא 2.3.3 נניח ש־$X \sim U(0,\theta)$. הראה שלכל t, $\left(\frac{X_1}{1-\alpha+t}, \frac{X_1}{t}\right)$ הוא רווח סמך בעל רמת מובהקות α.

להלן כמה כמויות ציריות סטנדרטיות.

דוגמא 2.3.4 1. אם $F_X(t) = P(X < t)$ פונקצית ההצטברות רציפה של X, אז $Q = F_X(X)$ היא בעלת התפלגות אחידה.

2. בהנתן מדגם $X_1,\dots,X_n$, המכפלה $\prod F_X(X_i;\theta)$ היא כמות צירית.

3. כמות צירית שימושית יותר: $\sum -logF(X_i;\theta)$, עם התפלגות $\Gamma(n,1)$.

תרגיל 2.3.5 מצא רווח סמך לפרמטר בהתפלגות $\lambda x^{\lambda-1}$ (בקטע $[0,1]$)

2.3.2 רווחי סמך עבור ההתפלגות הנורמלית

לו היה זה קורס בסטטיסטיקה, היינו מתארים כאן בהרחבה רווחי סמך:

1. לתוחלת של התפלגות נורמלית,

(א) כאשר סטיית התקן ידועה (הכמות הצירית מתפלגת נורמלית);

(ב) כאשר סטיית התקן אינה ידועה (הכמות הצירית מתפלגת t, לפי משפט (4)1.2.55).

2. לשונות של התפלגות נורמלית (χ^2).

3. להפרש התוחלות של שתי התפלגויות נורמליות

   (א) שהשונויות שלהן ידועות (נורמלית);
   (ב) שהשונויות שלהן אינן ידועות, אך אנו מניחים שהן שוות (t);
   (ג) שהשונויות שלהן אינן ידועות (t בקירוב).

4. לייחס בין השונויות של שתי התפלגויות נורמליות (התפלגות F);

5. לפרופורציה p בהתפלגות בינומית (בעזרת הקירוב הנורמלי)

   (א) אם אפשר להניח שהפרופורציה רחוקה מחצי (נורמלית בקירוב);
   (ב) ללא הנחות (נורמלית בקירוב).

פרק 3

בדיקת השערות

בדיקת השערות היא פרוצדורה המשתמשת בנתוני מדגם שנאסף מהתפלגות, כדי להכריע בהשערה לגבי הפרמטרים של ההתפלגות. למשל, אפשר לבנות פרוצדורה שתכריע האם משקלו הממוצע של קלח כרוב שנמכר ברשת מסויימת גדול מ־2.2 קילוגרם, על־פי מדגם של קלחי כרוב שנמכרו באותה רשת. ההשערה אינה מתייחסת לנתוני המדגם (שאותם אנו יודעים), אלא לפרמטרים של האוכלוסיה. היא אינה מספקת תשובות חד־משמעיות, אלא כאלו שיש בהן סיכוי (ידוע מראש!) לשגיאה.

3.1 השערות, הכרעות, והליך הבדיקה

3.1.1 השערת האפס וההשערה האלטרנטיבית

הבדיקה מכריעה בין שתי השערות לגבי ההתפלגות: **השערת האפס** H_0 שהיא לרוב ההשערה השמרנית אותה מנסים לדחות, **וההשערה האלטרנטיבית** H_1, שהיא הטענה שאותה רוצים להוכיח. דחיית H_0 נחשבת להצלחה, משום שהיא מוכיחה את H_1, ברמת מובהקות שנקבעה מראש.

השערת האפס היא **נקודתית** אם היא קובעת את ההתפלגות באופן חד־משמעי (היינו, זו השערה מהצורה $\theta = 5$, ולא $\theta \geq 5$ או $\theta \in (5,7)$. בבדיקת השערות מורכבות נעסוק בהמשך.)

דוגמא 3.1.1 שעור האנשים בעלי הפרעת אישיות חרדתית באוכלוסיה הוא 3%. חוקר משער ששיעורם בקרב בני מזל טלה שונה מן השיעור בקרב בני המזלות האחרים. הפרמטר הלא־ידוע הוא ההסתברות p לכך שבן מזל טלה יפתח הפרעת אישיות חרדתית. השערת האפס תהיה $H_0 : p = 0.03$. ההשערה האלטרנטיבית היא $H_1 : p \neq 0.03$.

דוחים את השערת האפס אם ההנחה שהיא נכונה מובילה למסקנה שההסתברות

לקבל את נתוני המדגם קטנה מערך קבוע מראש, הקרוי **רמת מובהקות**. הערך המקובל לרמת המובהקות הוא $\alpha = 0.05$.

דוגמא 3.1.2 סיכויי ההצלחה בסדרה של ניסויי ברנולי הם p, שאינו ידוע. עורכים n ניסויים, וכולם נכשלו. מהו הערך הגדול ביותר של p שאותו לא ניתן לדחות? תחת השערת האפס $H_0 : p = p_0$, הסיכויים לתוצאה שהתקבלה הם $(1-p_0)^n$. ההשוואה ל־$\alpha = 0.05$ מראה ש־$p_0 \approx \frac{-\log 0.05}{n} \approx \frac{3}{n}$.

3.1.2 הכרעות ושגיאות

לבדיקת השערות יש שתי תוצאות אפשריות: **דחיה** של השערת האפס (המהווה הוכחה סטטיסטית לנכונותה של H_1), ואי־דחיה שלה, שמשמעותה אחת משתיים: או ש־H_0 נכונה, או שאין מספיק נתונים להוכיח שלא. לאי־דחיה קוראים לפעמים "קבלה" של H_0, אבל חשוב להבין שאם H_1 אינה נקודתית, אז אין בכוחה של בדיקת השערות להוכיח את H_0, משום שאין בכוחו של מדגם סופי להפריד את ההשערה $\mu = 14$ מכל האלטרנטיבות $\mu = 14.1$, $\mu = 14.01$, וכו'.
גם במציאות יש שתי אפשרויות: או שהשערת האפס נכונה, או שההשערה האלטרנטיבית היא הנכונה. בהתאם לכך, יש ארבע אפשרויות:

1. H_0 נכונה, והליך בדיקת ההשערות דוחה אותה. אזעקת שווא. זוהי טעות, הנקראת **טעות מסוג ראשון** (או **false-positive**).

2. H_0 היא הנכונה, והליך בדיקת ההשערות לא דחה אותה. בהתחשב בנסיבות, זו כמובן התוצאה הרצויה של ההליך.

3. H_1 נכונה, והליך בדיקת ההשערות דוחה את H_0, ובכך מוכיח את H_1. הניסוי הצליח.

4. H_1 נכונה, והליך בדיקת ההשערות לא דוחה את H_0, ובכך מאשש בטעות את H_0. גם זו טעות, הנקראת **טעות מסוג שני** (**false-negative**).

דוגמא 3.1.3 חוקר רוצה להוכיח שתוספת ויטמין K לתזונת פגים מעלה את המשקל שלהם מעבר לתזונה הרגילה. ידוע שהוויטמין הזה אינו מזיק. נסמן את העליה במשקל מעבר לצפוי ב־X. השערת האפס (שאותה רוצים לדחות) קובעת שהתוחלת של X היא אפס, כלומר $H_0 : \mu = 0$. כנגדה, ההשערה האלטרנטיבית היא $H_1 : \mu > 0$. נפרש את ארבע האפשרויות, לפי הסדר שבו הן מופיעות לעיל.

1. הוויטמין אינו מועיל, והחוקר 'הוכיח' שהוא כן מועיל. נעשה כאן חוכא ואטלולא מן השיטה המדעית. טעות מסוג ראשון היא טעות חמורה.

2. הוויטמין אינו מועיל, והחוקר נכשל בנסיונו להוכיח את ההיפך. החוקר בזבז את זמנו ואת כספן של קרנות המחקר, אבל הגיע בסופו של דבר לתוצאה הנכונה.

3. הוויטמין מועיל, והחוקר מצליח לדחות את השערת האפס ולהוכיח שזה אכן כך. כולם מרוצים: המחקר מתפרסם בספרות המקצועית, ותזונת הפגים משתפרת.

4. הוויטמין מועיל, אלא שהניסוי לא הצליח להוכיח זאת. זו טעות מסוג שני - לא נעים (בזבוז זמן וכסף), אבל לא נורא (בשנה הבאה יבצע מישהו אחר מחקר על קבוצת תינוקות גדולה פי ארבעה, ואולי יצליח להוכיח את האפקט).

דוגמא 3.1.4 הסיווג של טעות מסוג ראשון כחמורה יותר מן הטעות מסוג שני אינו תורה מסיני. לדוגמא, אם מנסים לחזות פריצת מלחמה באמצעים סטטיסטיים, עדיף לדחות בטעות את השערת האפס (ולגייס את המילואים לחינם), מאשר לקבל בטעות את השערת האפס (ולספוג מתקפת פתע).

דוגמא 3.1.5 לקראת תחילת שנת הלימודים האקדמית, סטודנט אינו בטוח האם הקורס שנרשם אליו מתקיים בסמסטר הראשון או בסמסטר השני. הוא מתכוון להכריע על־ידי סקר בין סטודנטים אחרים (שרבים מהם מבולבלים לפחות כמוהו). דון במשמעות של טעות משני הסוגים במקרה זה.

3.1.3 הליך הבדיקה

בדיקת ההשערות נעשית על־ידי חישוב סטטיסטי T מנתוני המדגם, ובדיקה: האם ערך הסטטיסטי נופל **באזור דחיה**, שהוא תחום הערכים שאם הסטטיסטי יפול לתוכו נכריז על דחיית השערת האפס. משום כך, אזור הדחיה מגדיר את המבחן הסטטיסטי.

כדי להבהיר את הנושא, נציג תאור מפורט (אם כי פשטני לפרקים) של הצעדים הדרושים לבדיקת השערות.

- רקע תאורטי:
  - זיהוי והגדרת התופעה הנמדדת; זהו המשתנה המקרי X.
  - קביעת המודל, על־פי שיקולים תאורטיים, היורסיטיים וניסויים; המודל הוא משפחת ההתפלגויות $\{F_\theta\}$, שהמשתנה X מתפלג לפי אחת מהן.
  - מנסחים את השערת האפס, $H_0 : \theta = \theta_0$, כאשר θ_0 הוא הערך של הפרמטר שאותו מנסים **לשלול**.
  - מנסחים את ההשערה האלטרנטיבית לפי הידע התאורטי; בדרך כלל $H_1 : \theta \neq \theta_0$, אבל יתכן גם $H_1 : \theta > \theta_0$, $H_1 : \theta = \theta_1$, וכדומה.
  - קובעים את רמת המובהקות α. פירוש הדבר הוא ש־α תהיה ההסתברות לטעות מסוג ראשון.

- רקע סטטיסטי:
  - על פי משפחת ההתפלגויות והפרמטר שבו מדובר, קובעים סטטיסטי (למשל, אומד של הפרמטר).
  - על־פי ההשערות ורמת המובהקות, קובעים את אזור הדחיה.
  - קובעים את גודל המדגם n, בהתחשב באילוצים תקציביים או בחישוב המבוסס על הסיכוי הרצוי לטעות מסוג שני (שאותו מסמנים ב־β).
- ביצוע הניסוי:
  - אוספים נתוני מדגם $X_1, \dots, X_n$.
  - מחשבים את הסטטיסטי S.
  - בודקים האם הסטטיסטי נופל לאזור הדחיה.
- פרשנות התוצאות:
  - אם הסטטיסטי נפל לאזור הדחיה, דוחים את השערת האפס. זוהי 'הוכחה' שהשערת האפס אינה נכונה. אם ההשערה נכונה, הסיכוי לדחיה הוא α.
  - אם הסטטיסטי נפל מחוץ לאזור הדחיה, ההשערה אינה נדחית.

3.1.4 פונקציית העוצמה

כפי שהסברנו לעיל, אפשר לזהות את המבחן עם אזור הדחיה שלו, C: דוחים את H_0 אם ורק אם $\vec{X} \in C$. אם כך אפשר להגדיר

הגדרה 3.1.6 ***פונקציית העוצמה*** *של המבחן C היא הפונקציה $\Pi_C(\theta) = P_\theta(X \in C)$, כלומר הסיכוי לדחות את H_0 בהנחה שהפרמטר שווה ל־θ.*

ההסתברות המקסימלית לטעות מסוג ראשון, כלומר $\sup_{\theta \in H_0} \Pi_C(\theta)$, נקראת **רמת המובהקות** של המבחן (וגם **גודל המבחן**). את ההסתברות הזו אנחנו מבקשים לשמור קטנה ככל האפשר. מאידך, לכל $\theta \in H_1$ היינו רוצים ש־$\Pi_C(\theta)$ יהיה גדול ככל האפשר.

פונקציית העוצמה מונוטונית ב־C: אם $C \subseteq C'$, אז ברור ש־$\Pi_C(\theta) \leq \Pi_{C'}(\theta)$. ככל שבוחרים אזור דחיה גדול יותר, עולה הסיכוי לדחות את ההשערה. אם כך, יש כאן איזון תמורות (trade-off): ככל שאזור הדחיה גדל, עולה רמת המובהקות (לרעתנו), ומאידך עולה הסיכוי להצליח בדחיה כאשר $\theta \in H_1$ (לטובתנו).

רמת המובהקות המקובלת היא 5%. כלומר, בבדיקת השערות סטנדרטית, אם השערת האפס נכונה, הסיכוי לדחות אותה הוא 5%. מכאן שכ־5% מן התוצאות החיוביות המוכרזות בספרות המדעית כמובהקות, הוכרזו ככאלה בטעות; ראו את http://xkcd.com/882/ לאיור הכשל הזה ברפואה (אכן רופאים אינם מאמצים שיטות טיפול חדשות על־פי תוצאות של מחקר יחיד, אלא מחכים שהתוצאה תחזור בניסויים אחרים).

ציטוט 3.1.7 "נעשו מחקרים ובהם בדקו יותר ממאה פרמטרים שמאבחנים באמצעות הרורשך, 95% מהם התגלו כלא תקפים מדעית, ורק 5% תקפים" (אביבה לורי, "מרד הפסיכולוגים", "הארץ", *13.12.2007*).

תרגיל 3.1.8 נניח שבודקים סדרה של השערות על מכונה לייצור מספרים אקראיים; בכל המקרים השערת האפס נכונה, ומשערים, כרגיל, שהיא שגויה. איזה אחוז מההשערות "ימצאו "תקפות מבחינה מדעית"?

תרגיל 3.1.9 האם ההשערה $\theta = \theta_0$ מתקבלת על הדעת? מדד חשוב לכך הוא **ערך־**p של התוצאה $\vec{X} = \vec{x}$, שהוא סכום כל ההסתברויות בעלות סיכוי קטן או שווה ל־$\vec{x}$. דון במשמעות המדד הזה, וביתרונות והחסרונות שלו.

ציטוט 3.1.10 אלוף פיקוד העורף הנכנס, האלוף תמיר ידעי (*2/2/2017*): "במזרח התיכון, אירועים בעלי סבירות נמוכה מתרחשים בסבירות גבוהה". למה התכוון האלוף?

רעיון 3.1.11 מטילים קוביה שלוש פעמים, ומקבלים את התוצאה $4, 2, 5$. הסיכוי לתוצאה זו הוא $6^{-3} \approx 0.45\%$. האם אפשר להסיק שהקוביות מזוייפות ("ברמת מובהקות 0.5%"), שהרי התקבלה בהן תוצאה כל־כך בלתי סבירה?
במקום לחשב את ההסתברות למאורע עצמו, אפשר להגדיר $\tilde{P}(\omega_0) = \sum_{P(\omega) \leq \tilde{P}(\omega_0)} P(\omega)$; דוחים את ההשערה שזו ההתפלגות רק אם $\tilde{P}(\omega_0) < \alpha$. נתח באופן זה את הציטוט 3.1.10 לעיל.

3.2 השערות פשוטות

לאחר הסקירה של הסעיף הקודם, נתמקד כעת במקרה שבו גם ההשערה H_0 וגם ההשערה האלטרנטיבית H_1 הן השערות נקודתיות. כלומר,

$$H_0 : \theta = \theta_0; \qquad H_1 : \theta = \theta_1.$$

במקרה כזה המבחן הטבעי ביותר (שמיד נוכיח שהוא המבחן האולטימטיבי) משווה את הנראות של שתי ההשערות, והוא מבוסס על **יחס הנראות**

$$\lambda = \frac{L(\vec{X}; \theta_0)}{L(\vec{X}; \theta_1)}.$$

אזור הדחיה של מבחן יחס הנראות הוא $U_\kappa = \left\{\vec{X} : \lambda(\vec{X}) < \kappa\right\}$. המבחן הזה מונוטוני ב־$\kappa$: ככל ש־$\kappa$ גדל, גם אזור הדחיה גדל, ואיתו עולה **רמת המובהקות**, שהיא

כזכור ההסתברות לטעות מסוג ראשון, כלומר $\alpha = P_{\theta_0}(\lambda(\vec{X}) < \kappa)$. הקשר הזה בין κ לבין α מאפשר לקבוע את κ לכל α רצוי. לצד זה, אזור הדחיה קובע גם את הסיכוי לטעות מסוג שני, $\beta = 1 - \Pi_{U_\kappa}(\theta_1)$.

התרגיל הבא מסכם את הדוגמא הקלאסית של הפרדת שתי התפלגויות נורמליות בעלות אותה סטיית תקן. "כושר ההפרדה" הוא הפרש הפרש התוחלות ביחידות של סטיית התקן. כדי להשיג הפרדה ברמת מובהקות טובה, נחוץ מדגם בגודל $\approx \frac{1}{\delta^2}$.

תרגיל 3.2.1 נניח ש־$X \sim N(\pm\delta\sigma, \sigma^2)$, כאשר δ קבוע ידוע, אבל הסימן אינו ידוע. כתוב את אזור הדחיה עבור המבחן המשווה את ההשערה $H_0 : \mu = \delta\sigma$ להשערה האלטרנטיבית $H_1 : \mu = -\delta\sigma$. מהי רמת המובהקות אם נתון מדגם בגודל n?

הגדרה 3.2.2 *נאמר שמבחן C^* עבור השוואה בין שתי השערות פשוטות הוא **בעל עוצמה מרבית** עבור α **אם** $\Pi_{C^*}(\theta_0) \leq \alpha$, ולכל מבחן C המקיים את התנאי הזה, $\Pi_{C^*}(\theta_1) \geq \Pi_C(\theta_1)$. כלומר, $\Pi_{C^*}(\theta_1)$ הוא המקסימלי האפשרי מבין כל המבחנים C המקיימים את האילוץ $\Pi_C(\theta_0) \leq \alpha$.*

למה 3.2.3 (הלמה של ניימן־פירסון) לכל κ, מבחן יחס הנראות U_κ הוא בעל עוצמה מרבית (עבור α התלוי ב־κ).

הוכחה. נשווה את מבחן יחס הנראות U_κ למבחן אחר C. נסמן $\alpha = \Pi_{U_\kappa}(\theta_0)$, כך שלפי ההנחה $\Pi_C(\theta_0) \leq \alpha$. עלינו להראות ש־$\int_{U_\kappa} L(x;\theta_1) = \Pi_{U_\kappa}(\theta_1) \geq \Pi_C(\theta_1) = \int_C L(x;\theta_1)$. נתבונן בהפרש:

$$\begin{aligned}
\Pi_{U_\kappa}(\theta_1) - \Pi_C(\theta_1) &= \int_{U_\kappa \cap C^c} L(x;\theta_1) - \int_{U_\kappa^c \cap C} L(x;\theta_1) \\
&\geq \frac{1}{\kappa}\int_{U_\kappa \cap C^c} L(x;\theta_0) - \frac{1}{\kappa}\int_{U_\kappa^c \cap C} L(x;\theta_0) \\
&= \frac{1}{\kappa}\left(\int_{U_\kappa} L(x;\theta_0) - \int_C L(x;\theta_0)\right) \\
&= \frac{1}{\kappa}\left(\Pi_{U_\kappa}(\theta_0) - \Pi_C(\theta_0)\right) \geq 0.
\end{aligned}$$

□

תרגיל 3.2.4 כתוב את מבחן יחס הנראות עבור ההתפלגות המעריכית.

הערה 3.2.5 בסעיף זה התייחסנו לבדיקת השערות כאל תופעה בינארית: הצלחה או כשלון. גישה קצת אחרת מצמידה לכל החלטה **מחיר**; המחירים קובעים את תוחלת המחיר, התלויה כמובן ב־θ. לפי גישה זו, עלינו לקבוע מבחן שימזער את תוחלת המחיר (מבחן כזה נקרא **מבחן מינמקס**). מתברר שמבחן יחס הנראות, עם κ מתאים, הוא מבחן מינמקס.

3.3 בדיקת השערות כללית

3.3.1 מבחן יחס הנראות המוכלל

בסעיף זה נעסוק בבדיקת השערות כללית, המשווה את השערת האפס H_0 למצב הכללי $H = H_0 \cup H_1$. אנו מגדירים את **מבחן יחס הנראות המוכלל**

$$\lambda = \frac{\sup_{\theta \in H_0} L(\vec{X};\theta)}{\sup_{\theta \in H} L(\vec{X};\theta)}.$$

דוגמא 3.3.1 נניח ש־$X \sim N(\mu, \Sigma)$, התפלגות רב־נורמלית. בדוק את ההשערה $H_0 : \mu = \mu_0$ לעומת המקרה הכללי $H_1 : \mu \neq \mu_0$. מצא את יחס הנראות המוכלל כפונקציה של X. מה ההתפלגות של λ?

[החישוב מראה ש־$-2\log(\lambda) = (\bar{X}-\mu)'(\frac{1}{n}\Sigma)^{-1}(\bar{X}-\mu)$.]

בדוק את ההשערה $H_0 : \mu \in U_0$ במרחב האפשרויות $\mu \in U_1$, כאשר $U_0 \subseteq U_1 \subseteq \mathbb{R}^n$ הם תת־מרחבים של המרחב האוקלידי.

משפט 3.3.2 (משפט Wilks) *נניח ש־$H_0 \subseteq H$ הם מרחבים וקטוריים. אם H_0 נכונה, אז בגבול $n \to \infty$, $-2\log(\lambda) \sim \chi^2_d$ (התכנסות בהתפלגות), כאשר $d = \dim H - \dim H_0$ הוא מספר דרגות החופש של ההשערות.*

דוגמא 3.3.3 נניח ש־$X \sim \mathrm{Exp}(1/\theta)$, עם פונקציית הצפיפות $f(x) = \theta e^{-\theta x}$. נבדוק באמצעות מבחן יחס הנראות המוכלל את ההשערה $H_0 : \theta = \theta_0$. פונקציית הנראות היא $L(\vec{X},\theta) = \prod \theta e^{-\theta X_i} = \theta^n e^{-\theta n \bar{X}}$. הסופרימום בנקודה $\theta = \theta_0$ הוא כמובן $\theta_0^n e^{-\theta_0 n \bar{X}}$, ובמכנה יש לחשב את הסופרימום על הישר כולו, המתקבל בנקודה $\theta = 1/\bar{X}$ ושווה ל־$\bar{X}^{-n}e^{-n}$. לכן יחס הנראות הוא

$$\lambda = \frac{\theta_0^n e^{-\theta_0 n \bar{X}}}{\bar{X}^{-n}e^{-n}} = (e\theta_0 \bar{X} e^{-\theta_0 \bar{X}})^n.$$

כלומר, $-2\log\lambda = 2n[\theta_0\bar{X} - \log(\theta_0\bar{X}) - 1]$.

כדי להבין את היחס הזה, נזכר שלפי משפט הגבול המרכזי, בגבול מתקיים $\frac{\bar{X}-1/\theta}{1/(\theta\sqrt{n})} \sim N(0,1)$, כלומר $Z = \sqrt{n}(\theta\bar{X}-1) \sim N(0,1)$. תחת השערת האפס $\theta = \theta_0$, אפשר לחשב קירוב טיילור בסביבה $\theta_0\bar{X} \approx 1$, ולקבל

$$-2\log\lambda = 2n[\theta_0\bar{X} - \log(\theta_0\bar{X}) - 1] \sim n(\theta_0\bar{X}-1)^2 = Z^2.$$

כלומר, $-2\log\lambda \sim \chi^2_1$, כפי שמשפט וילקס חוזה.

3.3.2 בדיקת השערות למבחנים חד־צדדיים

נפעיל את מבחן יחס הנראות המוכלל על המקרה שבו המודל קובע התפלגות ממשפחה מעריכית (כלומר, $f(x;\theta) = a(\theta)b(x)e^{c(\theta)d(x)}$). כפי שראינו ב־(2.3),

$$L(X;\theta) = a(\theta)^n \cdot \prod_{i=1}^{n} b(X_i) \cdot e^{c(\theta)\sum d(X_i)}$$

ולכן יחס הנראות עבור $H_0 : \theta = \theta_0$ הוא

$$\lambda = \frac{a(\theta)^n e^{c(\theta)\sum d(X_i)}}{\sup_{\theta\in H_1} a(\theta)^n e^{c(\theta)\sum d(X_i)}},$$

כלומר תלוי ב־$S = \sum d(X_i)$ בלבד. לכן אזור הדחיה של מבחן יחס הנראות, עבור השערה אלטרנטיבית חד־צדדית, הוא או מהצורה $\{\sum d(X_i) > \kappa\}$ או מהצורה $\{\sum d(X_i) < \kappa\}$, תלוי האם $c(\theta)$ פונקציה עולה או יורדת.

3.3.3 בדיקת השערות באמצעות רווחי סמך

כפי שראינו בסעיף 2.3, רווח סמך הוא אזור שהפרמטר נמצא בתוכו בסיכוי $1-\alpha$. אזור דחיה הוא אזור ש(אם השערת האפס נכונה) הסטטיסטי נופל מחוץ לו בסיכוי $1-\alpha$. לפי משפט הסיכום הבא, בדיקת השערות ובניית רווחי סמך הם צדדים שונים של אותו מטבע:

משפט 3.3.4 נניח ש־$X_1,\dots,X_n \sim F_\theta$ הם נתוני מדגם בלתי תלויים, כאשר הפרמטר θ אינו ידוע. נניח ש־(T_1,T_2) הוא רווח סמך ברמת מובהקות α. אז $\{\theta_0 \notin (T_1,T_2)\}$ הוא אזור דחיה אפשרי להשערה $H_0 : \theta = \theta_0$.

במלים פשוטות, אפשר לדחות את ההשערה $H_0 : \theta = \theta_0$ אם θ_0 אינו נופל בתוך רווח סמך של θ, והסיכוי לטעות מסוג ראשון היא המשלים α של רמת המובהקות של רווח הסמך.

המשפט מאפשר לתרגם כל רווח סמך שבנינו בסעיף הקודם להליך לבדיקת השערה לגבי פרמטר מתאים: התוחלת, הפרש תוחלות, השונות, וכן הלאה.

דוגמא 3.3.5 (בדיקת השערות על התוחלת בהתפלגות נורמלית, כאשר השונות σ^2 ידועה) כשבוחנים את $H_0 : \mu = \mu_0$ כנגד ההשערה האלטרנטיבית $H_1 : \mu \neq \mu_0$, אזור הדחיה הוא $\left|\bar{X} - \mu_0\right| > z_{1-\alpha/2}\frac{\sigma}{\sqrt{n}}$. לעומת זאת כשההשערה חד־צדדית, למשל $H_1 : \mu > \mu_0$, אזור הדחיה הוא $\bar{X} - \mu_0 > z_{1-\alpha}\frac{\sigma}{\sqrt{n}}$.

נשווה בין אזורי הדחיה של ההשערה הדו־צדדית להשערות חד־צדדיות כאשר $\alpha = 0.05$. הערכים בטבלה מתייחסים להכרעה (H_1 לדחייה, H_0 לאי־דחייה), על־פי הערך של הסטטיסטי $Z = \frac{\bar{X}-\mu_0}{\sigma/\sqrt{n}}$, בהתאם להשערה האלטרנטיבית.

	-1.96	-1.645	0	1.645	1.96
$H_1: \mu \neq \mu_0$	H_1	H_0	H_0	H_0	H_1
$H_1: \mu < \mu_0$	H_1	H_1	H_0	H_0	H_0
$H_1: \mu > \mu_0$	H_0	H_0	H_0	H_1	H_1

דוגמא 3.3.6 (בדיקת השערות על התוחלת בהתפלגות נורמלית, כאשר השונות לא ידועה) פועלים בדיוק כמו בדוגמא 3.3.5 פרט לשני הבדלים: את הערך $\frac{\sigma^2}{n}$, שאינו ידוע, מחליפים באומד חסר ההטיה $\frac{s^2}{n-1}$; ואת ערכי הטבלה $z_{1-\alpha}$ וכדומה מחליפים ב־$t_{n-1,1-\alpha}$.

נסיים בהסבר הרמז שניתן למעלה על הקשר בין גודל המדגם לסיכוי לטעות מסוג שני. נסמן ב־ $\Phi(z) = P(Z < z)$ את פונקציית ההצטברות של המשתנה הנורמלי הסטנדרטי $Z \sim N(0,1)$. כלומר, $\Phi(z_\gamma) = \gamma$.

טענה 3.3.7 עבור התפלגות נורמלית $N(\mu, \sigma^2)$, עם שונות ידועה, משערים את השערת האפס $H_0 : \mu = \mu_0$, וכנגדה את ההשערה האלטרנטיבית $H_1 : \mu = \mu_1$. נניח ש־$\mu_1 > \mu_0$. עבור אזור הדחיה $\{\bar{X}_n > t\}$[1], הסיכוי לטעות מסוג ראשון הוא $\alpha = \Phi(-\frac{t-\mu_0}{\sigma/\sqrt{n}})$, והסיכוי לטעות מסוג שני הוא $\beta = \Phi(-\frac{\mu_1 - t}{\sigma/\sqrt{n}})$.

מכאן מתקבלת נוסחת ה־*trade-off* הקושרת בין α ל־β (אם n קבוע):

$$z_\beta + z_\alpha = -\frac{\mu_1 - \mu_0}{\sigma/\sqrt{n}};$$

ככל שהסיכוי לטעות מסוג ראשון יורד, הסיכוי לטעות מסוג שני עולה, ולהיפך.

תרגיל 3.3.8 כדי להגיע לסיכויי שגיאה α, β, כאשר μ_0, μ_1, σ קבועים, יש לבחור $n \geq \left(\frac{\sigma}{\mu_1 - \mu_0}\right)^2 (z_\alpha + z_\beta)^2$.

דוגמא 3.3.9 נניח, בדוגמא הקודמת, ש־$\sigma/\sqrt{n} = 0.2$, $\mu_0 = 0$ ו־$\mu_1 = 1$. מצא את β אם $\alpha = 0.05$.

[1] ככל שדוגמים ערך גדול יותר של $\bar{X}$, יותר מתקבל על הדעת שהתוחלת היא μ_1; לכן לוקחים את אזור הדחיה להיות קרן ימנית.

פרק 4

רגרסיה לינארית

4.1 רגרסיה דו־ממדית

מודל הרגרסיה הליניארית נועד לבדוק קשר ליניארי בין שני משתנים מקריים. לפי המודל, $X_1,\dots,X_n$ הם ערכים ידועים, וקיימים קבועים α,β, שאינם ידועים, כך ש־

$$Y_i = \alpha + \beta X_i + e_i, \quad e_i \sim N(0,\sigma^2),$$

באופן בלתי תלוי. (אם כך, במודל משתתפים שלושה פרמטרים: α,β,σ.) המשתנה X נקרא **משתנה מסביר**, בעוד ש־Y הוא **המשתנה המוסבר**. כרגיל, מסמנים $\bar{X} = \frac{1}{n}\sum X_i$, $\bar{Y} = \frac{1}{n}\sum Y_i$.

הערה 4.1.1 לפי המודל $\mathbf{E}(Y_i) = \alpha + \beta X_i$. סיכום השוויונים האלה מספק את הקשר בין התוחלות של הממוצעים,

$$\mathbf{E}(\bar{Y}) = \alpha + \beta\bar{X}, \tag{4.1}$$

ובפרט, $\mathbf{E}(Y_i - \bar{Y}) = \beta(X_i - \bar{X})$. למעשה, קל להווכח ש־$\bar{Y} \sim N(\beta\bar{X}+\alpha, \frac{\sigma^2}{n})$ ו־$Y_i - \bar{Y} \sim N(\beta(X_i-\bar{X}), \frac{n-1}{n}\sigma^2)$.

הערה 4.1.2 למודל הרגרסיה הליניארית יש הכללות חשובות רבות. החשובה ביותר היא המעבר לרגרסיה רב־ממדית, שבה יש כמה משתנים מסבירים (ראו סעיף 4.2). יש גם מודלים שנראים לא ליניאריים, ואפשר להפוך אותם בקלות לליניאריים. למשל, אם $Y_i = \alpha e^{\beta X_i + e_i}$ עם $e_i \sim N(0,\sigma^2)$, אפשר לעבור למודל ליניארי $\log Y_i = \alpha' + \beta' X_i + e_i$.

ציטוט 4.1.3 "הכפלת מהירות הגלישה באינטרנט מייצרת צמיחה שנתית של 0.3% בשנה - כך טוען מחקר חדש של חברת הייעוץ ארתור די ליטל בשיתוף עם חברת תשתיות התקשורת אריקסון ואוניברסיטת צ'אלמרס לטכנולוגיה. ... עוד עולה מהמחקר כי הגדלה של פי

ארבעה במהירות הגלישה תורמת כ־0.6% לקצב הצמיחה השנתי" (אמיר טייג, "האינטרנט בישראל מהאיטיים בעולם", *TheMarker*, *3.10.2011*, http://www.themarker.com/hitech/1.1488642). התוכלו להציע עוד מסקנות מרגשות העולות מן המחקר, בלי לקרוא אותו? האם המסקנות עולות מתוצאות המחקר, או מן ההנחות שלו?

תרגיל 4.1.4 קרא על חוק טיטיוס־בודה (הקשור למרחקי כוכבי הלכת מהשמש).

4.1.1 אומדים לקו הרגרסיה

ערכי המשתנה המסביר $(X_1,\dots,X_n)$ הם קבועים. לפי המודל, הנראות של הווקטור $(y_1,\dots,y_n)$ עבור המשתנה המוסבר Y היא

$$\prod \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(y_i-\beta x_i-\alpha)^2}{2\sigma^2}};$$

כדי לבנות אומד נראות מקסימלית ל־α,β, מחשבים את פונקציית הנראות:

$$\log L(x_i,y_i;\alpha,\beta) = -n\log\sqrt{2\pi\sigma^2} - \frac{1}{2\sigma^2}\sum(y_i-\beta x_i-\alpha)^2.$$

הבאת הנראות למקסימום מוליכה אותנו אל **שיטת הריבועים הפחותים** של גאוס: יש למצוא $\hat{\alpha},\hat{\beta}$ כך ש־ $s(\hat{\alpha},\hat{\beta})$ מינימלי, כאשר $s(\alpha,\beta) = \sum(Y_i-\beta X_i-\alpha)^2$. הסטטיסטים $(\hat{\alpha},\hat{\beta})$ הם אומדי נראות מקסימלית ל־(α,β).

המשוואות $\frac{\partial s}{\partial \alpha} = \frac{\partial s}{\partial \beta} = 0$ מתורגמות ל־

$$\begin{pmatrix} \sum X_i^2 & \sum X_i \\ \sum X_i & n \end{pmatrix} \begin{pmatrix} \hat{\beta} \\ \hat{\alpha} \end{pmatrix} = \begin{pmatrix} \sum X_iY_i \\ \sum Y_i \end{pmatrix},$$

ופתרונן

$$\begin{aligned}
\begin{pmatrix} \hat{\beta} \\ \hat{\alpha} \end{pmatrix} &= \frac{1}{n\sum X_i^2 - (\sum X_i)^2} \begin{pmatrix} n & -\sum X_i \\ -\sum X_i & \sum X_i^2 \end{pmatrix} \begin{pmatrix} \sum X_iY_i \\ \sum Y_i \end{pmatrix} \\
&= \frac{1}{n\sum X_i^2 - (\sum X_i)^2} \begin{pmatrix} n\sum X_iY_i - \sum X_i \sum Y_i \\ \sum X_i^2 \sum Y_i - \sum X_i \sum X_iY_i \end{pmatrix} \\
&= \frac{1}{\overline{X^2} - \overline{X}^2} \begin{pmatrix} \overline{XY} - \overline{X}\,\overline{Y} \\ \overline{X^2}\,\overline{Y} - \overline{X}\,\overline{XY} \end{pmatrix}.
\end{aligned}$$

בפרט,

$$\hat{\beta} = \frac{R}{S_X^2},$$

כאשר

$$R = \frac{1}{n-1}\sum(Y_i - \bar{Y})(X_i - \bar{X})$$

(סטטיסטי) ו־$S_X^2 = \frac{1}{n-1}\sum(X_i - \bar{X})^2$ (קבוע). להשלמת התמונה:

תרגיל 4.1.5 אומד הנראות המקסימלית ל־$\hat{\sigma}$ הוא $\hat{\sigma}^2 = \frac{1}{n}\sum(Y_i - \hat{Y}_i)^2$, כאשר $\hat{Y}_i = \hat{\alpha} + \hat{\beta}X_i$.

את ההתפלגות של האומד $\hat{\beta}$ אפשר לתאר בקלות:

טענה 4.1.6 $\hat{\beta} \sim N(\beta, \frac{\sigma^2}{(n-1)S_X^2})$. בפרט, $\hat{\beta}$ הוא אומד חסר הטיה של β.

הוכחה. מכיוון שה־X_i הם קבועים, S_X^2 הוא קבוע, ו־R הוא צירוף ליניארי של משתנים נורמליים בלתי תלויים. לכן הוא מתפלג נורמלית, ויש לחשב את המומנטים שלו. $\mathbf{E}(R) = \frac{1}{n-1}\beta\sum(X_i - \bar{X})^2 = \beta S_X^2$, ומכן $\mathbf{E}(\hat{\beta}) = \beta$; ובאשר לשונות,

$$\begin{aligned}
\mathbf{V}((n-1)R) &= \sum_{i,j}\mathrm{Cov}((X_i - \bar{X})(Y_i - \bar{Y}), (X_j - \bar{X})(Y_j - \bar{Y})) \\
&= \sum_{i,j}(X_i - \bar{X})(X_j - \bar{X})\mathrm{Cov}((Y_i - \bar{Y}), (Y_j - \bar{Y})) \\
&= \frac{(n-1)\sigma^2}{n}\sum(X_i - \bar{X})^2 - \frac{\sigma^2}{n}\sum_{i\neq j}(X_i - \bar{X})(X_j - \bar{X}) \\
&= \sigma^2\sum(X_i - \bar{X})^2 - \frac{\sigma^2}{n}\sum_{i,j}(X_i - \bar{X})(X_j - \bar{X}) \\
&= \sigma^2\sum(X_i - \bar{X})^2 = (n-1)S_X^2\sigma^2,
\end{aligned}$$

ולכן $\mathbf{V}(\hat{\beta}) = \mathbf{V}(\frac{(n-1)R}{(n-1)S_X^2}) = \frac{\sigma^2}{(n-1)S_X^2}$. □

הנוסחה ל־α מסובכת יותר, אבל אין בה צורך: כפל (4.2) מימין ב־$(\bar{X}, 1)$ נותן מיד

טענה 4.1.7 $\bar{Y} = \hat{\beta}\bar{X} + \hat{\alpha}$.

כלומר, קו הרגרסיה $y = \hat{\alpha} + \hat{\beta}x$ עובר דרך מרכז הכובד של גרף הפיזור.

מסקנה 4.1.8 $\hat{\alpha}$ הוא אומד חסר הטיה ל־α.

תרגיל 4.1.9 באמידת קו רגרסיה כל דגימה מספקת אומדים לשני הפרמטרים α, β. בניסוי אחד דגמו 12 פעמים 400 נקודות על הקו, ומיצעו את 12 האומדים שהתקבלו; בניסוי אחר דגמו 400 פעמים 12 נקודות על הקו, ומיצעו את 400 האומדים שהתקבלו. איזו שיטה עדיפה לדעתך? ערוך סימולציה כדי להכריע בשאלה זו. **הערה.** מכיוון שהאומדים מבוססים על שיטת הנראות המקסימלית, הם ממצים את האינפורמציה שהדגימה מחלצת מן המערכת, ולכן - בכפוף לטענה 4.1.18 להלן - מספר הדגימות הוא הנתון המשמעותי היחיד; אין הבדל עקרוני בין השיטות.

4.1.2 פירוק השונות

נסמן $\hat{Y}_i = \hat{\beta} X_i + \hat{\alpha}$. זהו אומד לערך של Y לפי המודל בנקודה X_i, היוצא מאומדי הנראות המקסימלית למקדמים. חיסור טענה 4.1.7 נותן $\hat{Y}_i - \bar{Y} = \hat{\beta}(X_i - \bar{X})$, ולכן $\frac{1}{n}\sum \hat{Y}_i = \bar{Y}$.

ההפרש $Y_i - \hat{Y}_i$ הוא גודל השגיאה באמידת הערך של קו הרגרסיה בנקודה X_i (אם כי האומד עצמו לוקח בחשבון גם את הערך Y_i, כמובן).

הערה 4.1.10 $\hat{Y}_i$ הוא אומד חסר הטיה ל־Y_i. אכן, $\mathbf{E}(Y_i - \hat{Y}_i) = 0$ כי $\mathbf{E}(\hat{Y}_i) = (X_i - \bar{X})\mathbf{E}(\hat{\beta}) + \mathbf{E}(\bar{Y})$.

דוגמא 4.1.11 נניח ש־$X_1, \dots, X_n$ משתנים מקריים בלתי מתואמים בעלי אותה שונות σ^2. נסמן, כרגיל, $\bar{X} = \frac{1}{n}(X_1 + \dots + X_n)$.

1. אם למשתנים אותה תוחלת, אז $\mathbf{E}(X_i - \bar{X}) = 0$.
2. $\mathrm{Cov}(\bar{X}, X_i - \bar{X}) = 0$.
3. $\mathrm{Cov}(X_i - \bar{X}, X_j - \bar{X}) = -\frac{1}{n}\sigma^2$ לכל $i \neq j$.
4. $\mathbf{V}(X_i - \bar{X}) = \frac{n-1}{n}\sigma^2$.

הערה 4.1.12 לפי דוגמא 4.1.11, $\mathrm{Cov}(\bar{Y}, \hat{\beta}) = 0$.

טענה 4.1.13

1. $\mathrm{Cov}(Y_i - \bar{Y}, \hat{\beta}) = \frac{(X_i - \bar{X})\sigma^2}{(n-1)S_X^2}$
2. $\mathbf{V}(\hat{Y}_i - \bar{Y}) = \frac{(X_i - \bar{X})^2\sigma^2}{(n-1)S_X^2}$;
3. $\mathbf{V}(Y_i - \bar{Y}) = \frac{n-1}{n}\sigma^2$;

4. $\mathbf{V}(Y_i-\hat{Y}_i)=\left(\frac{n-1}{n}-\frac{(X_i-\bar{X})^2}{(n-1)S_X^2}\right)\sigma^2$.

5. $\mathrm{Cov}(Y_i-\hat{Y}_i,\hat{Y}_i-\bar{Y})=0$.

הוכחה. 1. לפי הערה 4.1.11, $\mathrm{Cov}(Y_i-\bar{Y},Y_i-\bar{Y})=(\delta_{ij}-\frac{1}{n})\sigma^2$, והטענה נובעת מהגדרת R.

2. ישירות מ־$\hat{Y}_i-\bar{Y}=\hat{\beta}(X_i-\bar{X})$.

3. לפי תחילת הסעיף.

4. $Y_i-\hat{Y}_i=(Y_i-\bar{Y})-\hat{\beta}(X_i-\bar{X})$, ולפי סעיף 1

$$\mathbf{V}(Y_i-\hat{Y}_i)=\mathbf{V}(Y_i-\bar{Y})-2\frac{(X_i-\bar{X})^2\sigma^2}{(n-1)S_X^2}+\frac{(X_i-\bar{X})^2\sigma^2}{(n-1)S_X^2}.$$

5. נובע מכך ש־$\mathbf{V}(Y_i-\bar{Y})=\mathbf{V}(Y_i-\hat{Y}_i)+\mathbf{V}(\hat{Y}_i-\bar{Y})$.

□

הפירוק בסעיף 5, המתייחס לשונות של המשתנים המקריים, נכון גם עבור סכום הריבועים:

טענה 4.1.14

$$\sum(Y_i-\bar{Y})^2=\sum(Y_i-\hat{Y}_i)^2+\sum(\hat{Y}_i-\bar{Y})^2.$$

הוכחה. הווקטורים

$$(Y_i-\hat{Y}_i)=(Y_i-\bar{Y})-\hat{\beta}(X_i-\bar{X}),\qquad(\hat{Y}_i-\bar{Y})=\hat{\beta}(X_i-\bar{X})$$

מאונכים זה לזה לפי הנוסחה ל־$\hat{\beta}$. □

נוסחה חשובה זו מפרקת את השונות הפנימית של הערכים Y_i לשני חלקים: השונות של האומדים $\hat{Y}_i$, שהיא **השונות המוסברת**, והשונות של השגיאות $Y_i-\hat{Y}_i$, שהיא **השונות הבלתי מוסברת**.

הערה 4.1.15 סיכום על־פני i של סכום השונות וריבוע התוחלת מאפשר לחשב את התוחלת של כל סכום ריבועים בטענה 4.1.14:

$$\begin{aligned}\mathbf{E}(\sum(Y_i-\bar{Y})^2) &= (n-1)\sigma^2+\beta^2\sum(X_i-\bar{X})^2 \\ \mathbf{E}(\sum(Y_i-\hat{Y}_i)^2) &= (n-2)\sigma^2 \\ \mathbf{E}(\sum(\hat{Y}_i-\bar{Y})^2) &= \sigma^2+\beta^2\sum(X_i-\bar{X})^2.\end{aligned}$$

מאלה, הנוסחה עבור $\mathbf{E}(\sum(Y_i-\hat{Y}_i)^2)$ היא החשובה ביותר, משום שהיא מראה ש־

$$\hat{\sigma}^2=\frac{1}{n-2}\sum(Y_i-\hat{Y}_i)^2 \tag{4.2}$$

הוא אומד חסר הטיה ל־σ^2 (השווה לתרגיל 4.1.5).

4.1.3 אמידת ערך חדש

נניח ש־X הוא ערך חדש (שאינו משתתף בחישוב הממוצע $\bar{X}$), ונניח על־פי המודל ש־$Y\sim N(\beta X+\alpha,\sigma^2)$ (כמובן, באופן שאינו תלוי ב־$Y_1,\dots,Y_n$; כך Y אינו משתתף באמידת α,β). נגדיר $\hat{Y}=\hat{\beta}X+\hat{\alpha}$. חיסור טענה 4.1.7 מעביר נוסחה זו לצורה הסימטרית $\hat{Y}-\bar{Y}=\hat{\beta}(X-\bar{X})$, שממנה נובע לפי (4.1) ש־$\hat{Y}$ הוא אומד חסר הטיה ל־$\beta X+\alpha$, ובפרט $\mathbf{E}(Y-\hat{Y})=0$.

טענה 4.1.16 אם (X,Y) היא נקודת דגימה **חדשה**, אז מתקיים:

1. $\mathbf{V}(\hat{Y})=\left(\frac{1}{n}+\frac{(X-\bar{X})^2}{(n-1)S_X^2}\right)\sigma^2$.
2. $\mathbf{V}(Y-\bar{Y})=\frac{n+1}{n}\sigma^2$.
3. $\mathbf{V}(\hat{Y}-\bar{Y})=\frac{(X-\bar{X})^2}{(n-1)S_X^2}\sigma^2$
4. $\mathbf{V}(Y-\hat{Y})=\left(\frac{n+1}{n}+\frac{(X-\bar{X})^2}{(n-1)S_X^2}\right)\sigma^2$
5. $\mathrm{Cov}(Y-\bar{Y},\bar{Y}-\hat{Y})=0$.

הוכחה. 1. לפי הערה 4.1.12,

$$\begin{aligned}\mathbf{V}(\hat{Y}) &= \mathbf{V}(\bar{Y}+(X-\bar{X})\hat{\beta}) \\ &= \mathbf{V}(\bar{Y})+(X-\bar{X})^2\mathbf{V}(\hat{\beta}) \\ &= \frac{\sigma^2}{n}+(X-\bar{X})^2\frac{\sigma^2}{(n-1)S_X^2}.\end{aligned}$$

2. $\mathrm{Cov}(Y,\bar{Y})=0$ לפי ההנחה על Y.

3. לפי טענה 4.1.6, שהרי $\hat{Y}-\bar{Y}=\hat{\beta}(X-\bar{X})$.

4. $Y-\hat{Y}=(Y-\bar{Y})-\hat{\beta}(X-\bar{X})$, אבל Y ו־$\hat{\beta}$ בלתי תלויים, ו־$\bar{Y}$ בלתי מתואם עם $\hat{\beta}$ לפי טענה 4.1.12.

5. נובע מכך ש־$\mathbf{V}(Y-\hat{Y})=\mathbf{V}(Y-\bar{Y})+\mathbf{V}(\bar{Y}-\hat{Y})$.

□

פירושו של סעיף 4: ככל שהנקודה X רחוקה מן המרכז, וככל שערכי X_i במדגם קרובים למרכז, שונות השגיאה באמידת Y גדולה יותר; זהו כימות של הקושי לערוך אקסטרפולציה מנתונים צפופים.

לסיכום, עבור נקודת מדגם חדשה (X,Y), הפירוק

$$(Y-\hat{Y})=(Y-\bar{Y})+(\bar{Y}-\hat{Y})$$

מפרק את השגיאה לשני חלקים בלתי מתואמים. (השווה זאת לפירוק השונות

$$\mathbf{V}(Y_i-\bar{Y})=\mathbf{V}(Y_i-\hat{Y}_i)+\mathbf{V}(\hat{Y}_i-\bar{Y})$$

עבור נקודת מדגם שהשתתפה באמידת הפרמטרים.)

הערה 4.1.17 בבעיות מודרניות רבות, מספר הפרמטרים בווקטור β גדול ממספר נקודות הדגימה, ולכן אומד הנראות המקסימלית אינו מוגדר היטב. עם זאת, אפשר לאלץ בנוסף את הדרישה כי מספר הפרמטרים β_i השונים מאפס יהיה קטן; שיטת *Lasso* (**L***east* **a***bsolute* **s***hrinkage and* **s***election* **o***perator*), שמפעילים אחרי נרמול כל העמודות של X, עושה זאת על־ידי מינימיזציה של הערך

$$\|Y-X\beta\|_2^2/n+\lambda\|\beta\|_1,$$

כאשר $\lambda>0$ הוא פרמטר כוונון. בנקודת המינימום המגדירה את האומד $\hat{\beta}(\lambda)$ אכן מתקבל ש־$\hat{\beta}_j=0$ לערכי j רבים, וזו ראיה לכאורה לכך שהמשתנים X_j המתאימים אינם רלוונטיים.

4.1.4 בדיקת השערות על קו הרגרסיה

בכל דוגמא מציאותית, אמידת הפרמטר β תתן ערך שאינו אפס. האם נובע מכך ש־Y תלוי ב־X באופן לא טריוויאלי? הוכחת קשר ליניארי (דרך דחייה של השערת האפס הקובעת שאין קשר כזה) היא אחד השימושים החשובים ביותר של הרגרסיה הליניארית:

טענה 4.1.18 תחת השערת האפס $H_0:\beta=0$, הסטטיסטי

$$\frac{\hat{\beta}^2}{\frac{1}{\sqrt{n-1}}\frac{\hat{\sigma}}{S_X}} = \sqrt{\frac{(n-1)R^2}{\hat{\sigma}^2 S_X^2}} = \frac{\sum(Y_i-\bar{Y})(X_i-\bar{X})}{\sqrt{\frac{1}{n-2}\sum(X_i-\bar{X})^2\cdot\sum(Y_i-\hat{Y}_i)^2}}$$

מתפלג לפי ההתפלגות t_{n-2}.

הערה 4.1.19 על־ידי שינוי ליניארי של קנה המידה, אפשר להביא את הנתונים (X_i,Y_i) למצב שבו $\bar{X}=\bar{Y}=0$ ו־$S_X^2=S_Y^2=1$. במקרה זה $\hat{\alpha}=0$ ו־$\hat{\beta}=R=\frac{1}{n-1}\sum X_iY_i$. הסטטיסטי $\sqrt{\frac{(n-2)R^2}{1-R^2}}$ הוא שמתפלג t_{n-2} תחת השערת האפס $H_0:\beta=0$ לפי טענה 4.1.18.

תרגיל 4.1.20 רוצים לבדוק את הקשר בין ממוצע ציוני הבגרות (X) לבין ממוצע הציונים בסיום התואר (Y). ראש המחלקה מציע לחסוך בזמן החישוב, על־ידי קיבוץ התלמידים לפי טווח ציוני הבגרות והחלפת כל קבוצה של תלמידים שיש להם ציון X משותף בנתון אחד, שערך ה־Y שלו הוא ממוצע ערכי ה־Y שלהם. מה דעתך על קו הרגרסיה שיתקבל בצורה זו? איך תשפיע הטרנספורמציה על המבחן הבודק את ההשערה ששיפוע הקו הוא אפס?

4.2 מבוא לרגרסיה רב־ממדית

בסעיף הקודם כל נקודת מדגם כללה שני ערכים: X_i,Y_i, והמודל קבע ש־$Y_i\sim\beta X_i+\alpha+N(0,\sigma^2)$. ההכללה למקרה של מספר משתנים מסבירים פשוטה: כל נקודת מדגם היא זוג סדור (X_i,Y_i), אלא שהפעם X_i הוא משתנה מקרי רב־ממדי, ו־α וקטור מקדמים מאותו ממד. המודל קובע ש־

$$Y=\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k+\epsilon,$$

כאשר $\epsilon\sim N(0,\sigma^2)$. איננו זקוקים לערך הקבוע, α מן המקרה הדו־ממדי, משום שאפשר להציב זהותית $X_1=1$. בכתיב מקוצר, המודל הוא

$$Y=X\beta+\epsilon,$$

כאשר $\beta = (\beta_1, \dots, \beta_k)^t \in \mathbb{R}^k$ הוא וקטור המקדמים.

את נתוני המדגם $(X_1^{(j)}, \dots, X_k^{(j)}, Y^{(j)})$, $j = 1, \dots, n$, אפשר לאסוף במטריצות, שגם אותן מסמנים ב־X, Y: $X \in \mathrm{M}_{n\times k}(\mathbb{R})$ היא מטריצת הערכים המסבירים, ו־$Y \in \mathrm{M}_{n\times 1}(\mathbb{R})$ הוא וקטור המשתנה המוסבר.

תרגיל 4.2.1 הראה ש־$\hat{\beta} = (X^tX)^{-1}(X^tY)$ הוא אומד הנראות המקסימלית לווקטור המקדמים $\beta = (\beta_1, \dots, \beta_k)^t$.

אמידת המקדמים מוליכה מיד לווקטור $\hat{Y} = X\hat{\beta}$, המחזיק את האומדים של המשתנה המוסבר עבור כל אחד מנתוני המדגם.

תרגיל 4.2.2 הווקטורים $Y - \hat{Y}$ ו־$\hat{Y} - \bar{Y}$ מאונכים זה לזה (כאשר $\bar{Y}$ מסמן את הווקטור הקבוע שכל רכיביו שווים ל־$\frac{1}{n}\sum Y_i$).

מכאן מתקבל **פירוק השונות**

$$\|Y - \bar{Y}\|^2 = \|Y - \hat{Y}\|^2 + \|\hat{Y} - \bar{Y}\|^2;$$

השונות הכללית $\|Y - \bar{Y}\|^2$ מתפרקת לשני מרכיבים: **השונות המוסברת** על־ידי המודל, $\|\hat{Y} - \bar{Y}\|^2$ **והשונות הלא־מוסברת** $\|Y - \hat{Y}\|^2$. בפרט, היחס

$$R^2 = \frac{\|\hat{Y} - \bar{Y}\|^2}{\|Y - \bar{Y}\|^2}$$

נמצא תמיד בין 0 ל־1. ככל שהערך הזה גדול יותר, פירושו של דבר שהאומדים $\hat{Y}$ קרובים יותר לערכי האמת Y, והמודל טוב יותר. לכן זהו המדד המקובל לטיב המודל הלינארי.

4.2.1 בדיקת השערות

ההשערה העיקרית הנוגעת למודל הלינארי היא מהצורה

$$H_0 : \beta_{i_1} = \cdots = \beta_{i_{k-\ell}} = 0.$$

השערה כזו אומרת למעשה שאין ערך סטטיסטי למשתנים $X_{i_1}, \dots, X_{i_{k-\ell}}$ בנוכחות ℓ המשתנים המסבירים האחרים. כדי לבדוק השערה כזו, מחשבים את $\hat{\hat{\beta}}$, שהוא אומד הנראות המקסימלית של β תחת ההשערה H_0, ואת האומדים $\hat{\hat{Y}}$ הנובעים ממנו. ברור מאליו ש־$\|Y - \hat{Y}|^2 \leq \|Y - \hat{\hat{Y}}\|^2$, משום שבראשון נעשה שימוש בכל כח ההסבר של

המשתנים (היינו, $\hat{\beta}$ הוא אומד נראות מקסימלית במרחב ה־k ממדי, בעוד ש־$\hat{\hat{\beta}}$ הוא אומד נראות מקסימלית בתת־מרחב ℓ־ממדי).

כאן נחשף תפקידו האמיתי של R^2. תחת השערת האפס,

$$\frac{\|\hat{Y}-\hat{\hat{Y}}\|^2/(k-\ell)}{\|Y-\hat{Y}\|^2/(n-k)} \sim F_{k-\ell,n-k}.$$

עובדה זו מספקת מבחן להשערת האפס. דחיית השערת האפס מוכיחה שהמשתנים $X_{i_1},\dots,X_{i_{k-\ell}}$ (כולם או מקצתם) הם בעלי כח הסבר מובהק, מעבר למשתנים האחרים, ומכאן שהם שייכים בדין למודל.

משתנים שאי־אפשר להוכיח את נחיצותם, משליכים מן המודל. מציאת המודל "הנכון", כאשר המשתנים המסבירים הזמינים הם רבים, דורשת שיטות נוספות.

4.3 מבוא לניתוח שונות

רגרסיה לינארית מתאימה למקרים שבהם המשתנה המוסבר רציף, וגם המשתנים המסבירים רציפים. מודל עם משתנים מסבירים בדידים בודקים בניתוח שונות.

4.3.1 ניתוח שונות חד־ממדי

דוגמא 4.3.1 לנבדקים בקבוצת המדגם יש מין ומשקל. אנו מניחים שבכל אחת משתי קבוצות המין, המשקל מתפלג נורמלית. נסמן ב־A את משתנה המין (0 או 1), ו־Y את המשקל. המודל הוא $Y=\mu_A+\epsilon$. את ההשערה $H_0:\mu_0=\mu_1$ למדנו לבדוק באמצעות מבחן־t.

ניתוח שונות חד־ממדי מכליל את דוגמא 4.3.1 משתי קבוצות למספר כלשהו. נתוני המדגם הם זוגות (A,Y), כאשר A מקבל ערכים בקבוצה בגודל k, ו־$Y\sim N(\mu_A,\sigma^2)$ (מקובל להניח שהשונות קבועה). השערת האפס היא

$$H_0:\mu_1=\cdots=\mu_k,$$

ופירושה שכל התוחלות שוות, כלומר Y אינו תלוי ב־A. לכאורה אפשר היה לבדוק את ההשערה הזו באמצעות $\binom{k}{2}$ מבחני־t; אבל התלויות בין הסטטיסטיים של המבחנים האלה הופכים את השיטה הזו ללא מעשית, ואת הפרשנות הנאיבית של התוצאות שלה למטעה.

מתמטית, ניתוח שונות הוא מקרה פרטי של מודל הרגרסיה הרב־ממדית. זאת משום שאפשר להחליף את המשתנה המסביר A בווקטור יחידה מתאים. לדוגמא, אם $k=3$, הנתונים יהיו מהצורה $(1,0,0,Y)$, $(0,1,0,Y)$ או $(0,0,1,Y)$.

4.3.2 ניתוח שונות דו־ממדית ואינטרקציה

נניח שהמשתנה המסובר Y מלווה בשני משתנים מסבירים בדידים, A, B. למשל, Y עשוי להיות משקל התינוק בלידה, ו־A, B ארץ המוצא של האם והאב. המודל במקרה כזה כולל כמה שכבות:

$$Y = \mu + \alpha_A + \beta_B + \gamma_{AB} + N(0, \sigma^2);$$

שפירושו שהנתונים מתפלגים נורמלית סביב מרכז משותף μ, עם תרומה קבועה α_a לקבוצה $A = a$ (מניחים ש־$\sum \alpha_a = 0$), ותרומה קבועה β_b לקבוצה $B = b$ (מניחים ש־$\sum \beta_b = 0$), ואינטרקציה γ_{ab} המטה את הקבוצה $A = a, B = b$ מעבר לתרומתו של כל משתנה בנפרד (כאן נכון לאלץ את התנאים $\sum_a \gamma_{ab} = 0$ לכל b ו־$\sum_a \gamma_{ab} = 0$ לכל a).

לכאורה אפשר היה להניח ש־$\mu = \alpha_a = \beta_b = 0$ ולוותר על האילוצים הלינאריים הכובלים את γ_{ab}; אבל המתכונת שבה בחרנו היא המתאימה להשערות המקובלות, שהן:

1. $H_0 : (\forall a)\alpha_a = 0$. היינו, תרומתו של המשתנה המסביר A אינה מובהקת, והוא אינו נחוץ במודל.
2. $H_0' : (\forall b)\beta_b = 0$. היינו, תרומתו של המשתנה המסביר B אינה מובהקת.
3. $H_0'' : (\forall a, b)\gamma_{ab} = 0$. היינו, לשני המשתנים יש השפעה על התוחלת, אבל השפעות אלה בלתי תלויות זו בזו.

4.4 מבוא לניתוח גורמים

ניתוח גורמים (factor analysis) הוא טכניקה להורדת הממד של מרחב התצפיות. בקצרה, אם התצפיות מאורגנות במטריצה $A \in \mathrm{M}_{n\times m}(\mathbb{R})$, אז מבין כל תת־המרחבים ה־k ממדיים של $\mathbb{R}^n$, זה שסכום ריבועי ההיטלים עליו הוא מינימלי הוא תת־המרחב הנפרש על־ידי k הווקטורים העצמיים הראשונים (כלומר בעלי ערכים עצמיים גדולים ביותר) של AA^t. יתרה מזו, עוצמת התרומה של כל וקטור עצמי פרופורציונית לערך העצמי שלו. עובדה זו מציעה מערכת צירים חדשה למרחב המשתנים, ומאפשרת גם ויזואליזציה טובה (עבור $k = 2$), גם זיהוי משתנים מהותיים, וגם הפרדה בין אוכלוסיות. הניתוח נעזר בכך שאם v וקטור עצמי של AA^t, אז $A^t v$ וקטור עצמי של $A^t A$, עם אותו וקטור עצמי.

מטריצה מלבנית $T \in \mathrm{M}_{n\times r}(\mathbb{R})$, $r \leq n$, היא **אורתוגונלית** אם $T^{\mathrm{t}}T = I_r$. כלומר, העמודות של T הן וקטורי יחידה מאונכים זה לזה.

טענה 4.4.1 תהי $A \in \mathrm{M}_{n\times m}(\mathbb{R})$ מטריצה מדרגה r. אז יש פירוק $A = U\Sigma V^{\mathrm{t}}$, כאשר $U \in \mathrm{M}_{n\times r}(\mathbb{R})$ ו־$V \in \mathrm{M}_{m\times r}(\mathbb{R})$ אורתוגונליות, ואילו $\Sigma \in \mathrm{M}_r(\mathbb{R})$ היא מטריצה אלכסונית שרכיביה חיוביים.

הוכחה. כידוע, $A^{\mathrm{t}}A$ מטריצה נורמלית, ולכן לכסינה אורתוגונלית. אכן, יהיו $v_1,\dots,v_m \in \mathbb{R}^m$ וקטורים עצמיים אורתוגונליים ומנורמלים של $A^{\mathrm{t}}A$, עם ערכים עצמיים $\lambda_1,\dots,\lambda_m$, בהתאמה, המסודרים כך ש־

$$\lambda_1 \geq \cdots\cdots \lambda_r > \lambda_{r+1} = \cdots = \lambda_m = 0.$$

נארוז את $v_1,\dots,v_r$ כעמודות במטריצה $V \in \mathrm{M}_{m\times r}(\mathbb{R})$, שהיא לפיכך אורתוגונלית, היינו $V^{\mathrm{t}}V = I_r$. נסמן $\Sigma = \mathrm{diag}(\sqrt{\lambda_1},\dots,\sqrt{\lambda_r})$. לפי ההנחה, $A^{\mathrm{t}}AV = V\Sigma^2$. כעת נסמן

$$U = AV\Sigma^{-1} \in \mathrm{M}_{n\times r}(\mathbb{R}),$$

ונקבל שגם U אורתוגונלית משום ש־$U^{\mathrm{t}}U = \Sigma^{-1}V^{\mathrm{t}}A^{\mathrm{t}}AV\Sigma^{-1} = \Sigma^{-1}V^{\mathrm{t}}V\Sigma = I_r$. יתרה מזו, הקשר בין U ו־V סימטרי: $A^{\mathrm{t}}U\Sigma^{-1} = A^{\mathrm{t}}AV\Sigma^{-2} = V$, ועמודות U הן הווקטורים העצמיים של AA^{t}, משום ש־$AA^{\mathrm{t}}U = AA^{\mathrm{t}}AV\Sigma^{-1} = AV\Sigma = U\Sigma^2$.
כעת $A = AVV^{\mathrm{t}} = U\Sigma V^{\mathrm{t}}$ (השוויון האחרון דורש נימוק נוסף), כפי שרצינו. □

ההצגה הזו מפרקת את A לסכום של מטריצות מדרגה 1:

$$A = \sum_{i=1}^{r} \lambda_i u_i {v_i}^{\mathrm{t}},$$

כאשר $u_1,\dots,u_r$ הן העמודות של U. מכאן נותר צעד אחד ל**משפט Eckart-Young** (1936): המטריצה מדרגה k הקרובה ביותר ל־A (בנורמה של סכום ריבועי הערכים המוחלטים של ההפרש) היא הסכום החלקי

$$A = \sum_{i=1}^{k} \lambda_i u_i {v_i}^{\mathrm{t}}.$$

נתבונן בקבוצת וקטורי העמודה של A (שהיא קבוצת נקודות ב־$\mathbb{R}^m$). מכאן נובע שתת־המרחב ה־k ממדי שסכום ריבועי ההיטלים עליו הוא מקסימלי, וסכום ריבועי המרחקים ממנו מינימלי, הוא המרחב הנפרש על־ידי k הוקטורים העצמיים הראשונים, $v_1,\dots,v_k$. לצרכי ויזואליזציה, התוצאה הזו שימושית במיוחד כאשר $k = 2$.

תרגיל 4.4.2 קרא על שיטת PCA (Principal Components Analysis).

פרק 5

מבחנים לא פרמטריים

5.1 ההתפלגות המולטינומית

ההתפלגות המולטינומית היא התפלגות בדידה, שלא הוזכרה עד כה משום שהיא מתארת התפלגות משותפת של כמה משתנים תלויים, ולא של משתנה יחיד.

הגדרה 5.1.1 יהי $(p_1,\dots,p_m)$ וקטור הסתברויות, כלומר $p_1+\dots+p_m=1$. יהי $n\geq 1$. הווקטור $(X_1,\dots,X_m)$ מתפלג **התפלגות מולטינומית** $M(n;p_1,\dots,p_m)$ אם $P(\vec{X}=\vec{k})=\binom{n}{k_1,k_2,\dots,k_m}p_1^{k_1}\cdots p_m^{k_m}$ (כרגיל, $\binom{n}{k_1,k_2,\dots,k_m}=\frac{n!}{k_1!\cdots k_m!}$).

התפלגות מולטינומית היא הכללה של ההתפלגות הבינומית: אם $X\sim\mathrm{Bin}(n,p)$, אז $(X,n-X)\sim M(n;p,1-p)$.

הדוגמא הקלאסית להתפלגות מולטינומית היא של וקטור שכיחויות: אם מכונה מייצרת סוכריות בצבעים שונים, בהסתברויות p_i, ו־X_i הוא המשתנה הסופר כמה סוכריות יש בצבע i מתוך מדגם של n סוכריות, אז $(X_1,\dots,X_m)\sim M(n;p_1,\dots,p_m)$. ובאופן פורמלי:

תרגיל 5.1.2 תהי I קוביה מוטה, כלומר $I\sim\Omega=\{1,\dots,m\}$ עם ההסתברויות $p_1,\dots,p_m$ שסכומן $p_1+\dots+p_m=1$. נניח ש־$I_1,I_2,\cdots,I_n$ הן דגימות אקראיות ובלתי תלויות במרחב הזה. אז $X_i=|\{\ell\colon I_\ell=i\}|$ מגדיר וקטור $(X_1,\dots,X_m)$ המתפלג מולטינומית $M(n;p_1,\dots,p_m)$.

אפשר לכתוב את הווקטור $(X_1,\dots,X_m)$ כסכום של n משתנים "מולטינוליים" $(I_1,\dots,I_m)\sim M(1;p_1,\dots,p_m)$ (אם $\vec{X}$ מכליל את ההתפלגות הבינומית, $\vec{I}$ מכליל את התפלגות ברנולי).

תרגיל 5.1.3 אם $(X_1,\dots,X_m)\sim M(n;p_1,\dots,p_m)$ אז

1. הרכיבים מתפלגים לפי $X_i \sim \mathrm{Bin}(n, p_i)$;
2. לכן $\mathbf{E}(X_i) = np_i$, $\mathbf{V}(X_i) = np_i(1-p_i)$;
3. לכל $i \neq j$, $X_i + X_j \sim \mathrm{Bin}(n, p_i + p_j)$;
4. לכן $\mathrm{Cov}(X_i, X_j) = -np_ip_j$;
   הדרכה. $2\mathrm{Cov}(X_i, X_j) = \mathbf{V}(X_i + X_j) - \mathbf{V}(X_i) - \mathbf{V}(X_j)$
5. לכל וקטור מקדמים $\alpha_1, \ldots, \alpha_m$, $\mathbf{V}(\sum \alpha_i X_i) = n \sum \alpha_i^2 p_i - n(\sum \alpha_i p_i)^2$; זו השונות של המשתנה המקרי המקבל את הערך α_i בסיכוי p_i, מוכפלת ב־n.

תרגיל 5.1.4 הראה שאם $\vec{X} \sim M(n; \vec{p})$ אז $\mathrm{Cov}(X)$ אינה הפיכה. מצא מטריצה P כך ש־$PP^t = \mathrm{Cov}(X)$.

תרגיל 5.1.5 נניח ש־$\vec{X} \sim M(n; \vec{p})$. נתבונן בווקטור $\vec{Y}$, בעל התפלגות רב־נורמלית, שהתוחלת שלו $n\vec{p}$ ומטריצת השונות המשותפת שלו היא $\Sigma_{ij} = np_i\delta_{ij} - np_ip_j$. בכל רכיב, Y_i הוא קירוב נורמלי ל־X_i. הראה ש־$\sum Y_i = 0$ (בהסתברות 1). פורמלית, ההתפלגות של $\vec{Y}$ אינה מוגדרת היטב משום ש־Σ לא הפיכה. פתור את הבעיה על־ידי הגדרת הווקטור $\vec{Y}' = (Y_1, \ldots, Y_{n-1})$ שהתפלגותו רב־נורמלית עם השונות המשותפת $(\Sigma_{ij})_{i,j<n}$, ושחזר את המצב הקודם על־ידי הגדרת $Y_n = n - \sum Y_i$.

5.2 מבחני χ^2

מבחן χ^2 (ששמו נגזר מהתפלגות χ^2 שפגשנו בסעיף 1.2.4) הוא מבחן סטטיסטי "לא פרמטרי": למרות שהמודל שלו מערב פרמטרים רבים, הוא אינו מודד ישירות אף פרמטר, אלא תוקף את אופי ההתפלגות באופן ישיר. למבחן זה יש גרסאות רבות; נסתפק כאן בשתי דוגמאות חשובות.

5.2.1 הכרעה בין שתי התפלגויות

שני חוקרים מתווכחים מהי התפלגות מיני הציפורים באי קטן. אחד מהם טוען ש־50% מהציפורים שייכות למשפחת התוכיים, 40% לסבכיים ו־10% לדרוריים. השני טוען שההתפלגות האמיתית היא $14\% : 38\% : 48\%$. כמה ציפורים עליהם לצוד כדי להכריע בין התאוריות? ואם ידוע שהשניים צדו באקראי 200 ציפורים, והתברר שהחלוקה למשפחות היא $(27 : 69 : 104)$, עם מי הצדק?

המודל קובע שהווקטור $(X_1,\ldots,X_m)$ מתפלג מולטינומית, $\vec{X}\sim M(n,\vec{p})$. נשווה שתי השערות פשוטות:

$$H_0:\vec{p}=(p_1^0,\ldots,p_m^0);\qquad H_1:\vec{p}=(p_1^1,\ldots,p_m^1).$$

הנראות של תוצאות המדגם היא

$$L(\vec{X};\vec{p})=\binom{n}{X_1,\,X_2,\,\ldots,\,X_m}p_1^{X_1}\cdots p_m^{X_m}, \tag{5.1}$$

ולכן יחס הנראות הוא

$$\lambda=\frac{\binom{n}{X_1,X_2,\ldots,X_m}(p_1^0)^{X_1}\cdots(p_m^0)^{X_m}}{\binom{n}{X_1,X_2,\ldots,X_m}(p_1^1)^{X_1}\cdots(p_m^1)^{X_m}}=\prod(p_i^0/p_i^1)^{X_i},$$

כלומר $\log\lambda=\sum X_i\log(p_i^0/p_i^1)=\sum\alpha_iX_i$, כאשר $\alpha_i=\log(p_i^0/p_i^1)$. לפי תרגיל 5.1.3,

$$\mathbf{E}(\log\lambda)=n\sum\alpha_ip_i,$$

$$\mathbf{V}(\log\lambda)=n\left[\sum\alpha_i^2p_i-(\sum\alpha_ip_i)^2\right].$$

נסמן ב־$T_1,\ldots,T_n$ משתנים מקריים בלתי תלויים שכל אחד מהם מקבל את הערך α_i בסיכוי p_i. אז $\log\lambda=T_1+\cdots+T_n$. אם n גדול, אפשר לקרב

$$\log\lambda\sim N\left(n\sum\alpha_ip_i,n\left[\sum\alpha_i^2p_i-(\sum\alpha_ip_i)^2\right]\right).$$

אם הווקטורים p^0 ו־p^1 רחוקים זה מזה, ההכרעה בין שתי האפשרויות תהיה קלה ומהירה. ננתח את המקרה שבו ההפרדה קשה יותר, כלומר $p_i^1/p_i^0=1+\epsilon_i$, כאשר ϵ_i קטנים. נסמן $\Delta_k=\sum\epsilon_i^kp_i^0$, ונזניח את הגורמים $\Delta_4\approx\Delta_5\approx\cdots\approx0$.

תרגיל 5.2.1 מן האילוצים $\sum p_i^0=\sum p_i^1=1$ נובע שהווקטורים $\vec{p}^0$ ו־$\vec{\epsilon}$ מאונכים זה לזה, כלומר $\Delta_1=0$.

לפי קירוב טיילור, $\alpha_i=\log(p_i^0/p_i^1)=-\log(1+\epsilon_i)\approx-\epsilon_i+\frac{1}{2}\epsilon_i^2-\frac{1}{3}\epsilon_i^3$. לכן

$$\sum\alpha_ip_i^0\approx\frac{1}{2}\Delta_2-\frac{1}{3}\Delta_3;$$

$$\sum \alpha_i^2 p_i^0 \approx \Delta_2 - \Delta_3;$$

$$\sum \alpha_i p_i^1 \approx -\frac{1}{2}\Delta_2 + \frac{1}{6}\Delta_3;$$

$$\sum \alpha_i^2 p_i^1 \approx \Delta_2.$$

אם נניח בנוסף ש־$\Delta_2 \gg \Delta_3, \Delta_2^2$, נראה שהמרחק בין מרכזי ההתפלגויות הנורמליות (עבור H_0 ועבור H_1) ביחידות של סטיית התקן הוא $\sqrt{n}\sqrt{\Delta_2}$. כלומר, על־מנת להפריד את שתי ההתפלגויות האלה, ברמת מובהקות סבירה, נדרש מדגם מסדר הגודל של $n \approx \frac{1}{\Delta_2}$.

5.2.2 מבחן לטיב ההתאמה

רוצים לבדוק שקוביית משחק היא הוגנת, על־פי תוצאות של, נאמר, 1000 הטלות. המודל קובע שהתפלגות התוצאות $(X_1, \dots, X_6)$ היא מולטינומית, $\vec{X} \sim M(1000, \vec{p})$. השערת האפס, שלפיה הקוביה הוגנת, היא ההשערה $p_1 = \cdots = p_6 = \frac{1}{6}$. בפרט, לכל i מתקיים $X_i \sim \text{Bin}(1000, p_i)$, ואפשר לבחון האם הערך X_i מתיישב עם ההשערה ש־$p_i = \frac{1}{6}$. אבל הסטטיסטים בששת המבחנים האלה תלויים זה בזה, וזה מסבך מאד את ניתוח התוצאות. אפשר להפעיל שיטות מתוחכמות יותר, כמו חלוקה של שש האפשרויות לשתי קבוצות, ובדיקה שהקוביה אינה מעדיפה אף קבוצה; יש $\frac{1}{2}\binom{6}{3} = 10$ חלוקות אפשריות, ומכיוון שמספיקה חלוקה אחת שבה הקוביה אינה הוגנת כדי להוכיח שהקוביה אינה הוגנת, נצטרך למצוא דרך לחשב את הסיכוי שתוצאה כזו תתקבל (בטעות) בקוביה שהיא כן הוגנת. שוב התלויות מסבכות את הניתוח.

נעבור למקרה הכללי. המודל קובע שהווקטור $(X_1, \dots, X_m)$ מתפלג מולטינומית, $\vec{X} \sim M(n, \vec{p})$. השערת האפס היא

$$H_0 : \vec{p} = (p_1^0, \dots, p_m^0);$$

בניגוד לסעיף הקודם, הפעם אנו בודקים אותה כנגד ההשערה האלטרנטיבית $\vec{p} \neq \vec{p}^0$. הממד של מרחב ההשערות הוא $m-1$, בשל האילוץ האפיני $\sum p_i = 1$.

הרעיון הוא לרכז את כל אי־ההתאמות בנוסחה אחת, שתופיע גם בשאר הסעיפים בפרק הזה. מקובל לסמן ב־$O_i = X_i$ את הערך המתקבל בתא ה־i (זהו ה־**O**bserved); וב־$E_i = np_i^0$ את הערך הצפוי לפי השערת האפס (זהו ה־**E**xpected).

טענה 5.2.2 הסטטיסטי

$$W = \sum \frac{(O_i - E_i)^2}{E_i}$$

מתפלג בקירוב χ^2_{m-1}.

איננו מוכיחים את הטענה באופן מלא. במקום זה, ניתן שני נימוקים משכנעים:

1. קירוב ל־$-2\log\lambda$.
2. סכום ריבועים.

חישוב λ דורש את התרגיל השימושי הבא:

תרגיל 5.2.3 המקסימום של המכפלה $\prod p_i^{x_i}$, תחת האילוץ $\sum x_i = c$, מתקבל כאשר הווקטורים $(x_1,\dots,x_m),(p_1,\dots,p_m)$ פרופורציונליים.

הערה 5.2.4 (הסבר ראשון) יחס הנראות חושב ב־(5.1). לפי תרגיל 5.2.3, המקסימום של המכפלה $L(\vec{X};\vec{p})$ מתקבל כאשר $\hat{p}_i = \frac{1}{n}X_i$. לכן יחס הנראות המוכלל שווה ל־

$$\lambda = \frac{\binom{n}{X_1,X_2,\dots,X_m}(p_1^0)^{X_1}\cdots(p_m^0)^{X_m}}{\sup_{\vec{p}}\binom{n}{X_1,X_2,\dots,X_m}(p_1)^{X_1}\cdots(p_m)^{X_m}} = \prod(np_i^0/X_i)^{X_i},$$

כלומר

$$-2\log\lambda = -2\sum X_i\log(np_i^0/X_i).$$

לפי משפט *Wilks*, תחת השערת האפס $-2\log\lambda \sim \chi^2_{m-1}$. כדי להביא את המבחן לצורה מעט פשוטה ומקובלת יותר, נכתוב שוב ש־$\frac{X_i/n}{p_i^0} = 1+\epsilon_i$; תחת הנחת האפס, הסטטיסטיים ϵ_i קטנים, ומתקיים $\sum p_i^0\epsilon_i = 0$, ולכן אפשר לקרב:

$$-2\log\lambda = 2\sum np_i^0(1+\epsilon_i)\log(1+\epsilon_i) \approx \sum np_i^0\epsilon_i^2 = \sum\frac{(X_i-np_i^0)^2}{np_i^0} = W$$

הערה 5.2.5 (הסבר שני) תחת השערת האפס, לכל i מתקיים

$$X_i \sim \mathrm{Bin}(np_i^0, np_i^0(1-p_i^0)),$$

ובקירוב, לפי משפט הגבול המרכזי $\frac{X_i-np_i^0}{\sqrt{np_i^0(1-p_i^0)}} \sim N(0,1)$. הבעיה היא שהמשתנים האלה תלויים, כי סכומם המשוקלל קבוע. כהערת־אגב נאיבית נציין שאלמלא התלות, ואם היינו לוקחים במכנה $np_i^0(1-p_i^0)$, W היה סכום של m ריבועי משתנים נורמליים. האילוץ הלינארי $\sum X_i = n$ מוריד את מספר דרגות החופש וגורם ש־$W \sim \chi^2_{m-1}$.

לפי מסקנה 1.2.31, אם $\vec{X}$ רב־נורמלי, אז המשתנה הנותר אחרי k אילוצים גם הוא רב נורמלי; לאחר נרמול מתאים (שאיננו מסבירים כאן לפרטיו), סכום הריבועים מתפלג χ^2, ומספר דרגות החופש שווה לממד של X פחות מספר האילוצים הלינאריים.

הערה 5.2.6 כלל האצבע המקובל קובע שהקירוב $\sum \frac{(O_i-E_i)^2}{E_i} \sim \chi^2_{m-1}$ סביר בתנאי שכל ה־$E_i > 5$. אם המדגם אינו מספיק גדול לשם כך, אפשר לאחד תאים בעלי הסתברות נמוכה ולבדוק השערה מצומצמת, המתייחסת לתאים המאוחדים.

תרגיל 5.2.7 תרנגולת מטילה ביצים בארבעה גדלים. בדוק את ההשערה שהגדלים מתפלגים לפי הפרופורציות $4:3:2:1$, אם בפועל נאספו ביצים בחלוקה $25:25:16:16$.

תרגיל 5.2.8 למשחק מזל מסויים יש חמש תוצאות סופיות אפשריות, וההסתברויות שלהן (בקירוב של ארבע ספרות אחרי הנקודה) 0.6784, 0.2103, 0.0754, 0.0274 ו־0.0085. באחד מאתרי האינטרנט הופיעו ההסתברויות 0.6764, 0.2133, 0.0751, 0.0277, 0.0089, המבוססות כפי הנראה על סימולציה. כמה משחקים נערכו לצורך הסימולציה הזו?

תרגיל 5.2.9 אדם מטיל קוביה, בהחבא, 600 פעמים. כדי לשכנע שהקוביה הוגנת, הוא מדווח על התפלגות $102:102:98:99:104:97$. מה דעתך (המנומקת סטטיסטית) על התוצאה?

תרגיל 5.2.10 התפלגות ה־10^{12} ספרות הראשונות של π בבסיס הקסהדצימלי היא כדלקמן:

0	62499881108
1	62500212206
2	62499924780
3	62500188844
4	62499807368
5	62500007205
6	62499925426
7	62499878794
8	62500216752
9	62500120671
A	62500266095
B	62499955595
C	62500188610
D	62499613666
E	62499875079
F	62499937801

.

האם יש די ראיות לקבוע ש־π אינו מספר נורמלי? (מספר **נורמלי** הוא מספר שבהצגה שלו לפי כל בסיס b, גבול השכיחות של כל ספרה הוא $1/b$.)

5.2.3 תלות בין משתנים בינאריים

נתונים שני משתנים מקריים I, J, המקבלים ערכים בקבוצות סופיות. לכל נקודת מדגם יש ערך I וערך J (למשל: מין וצבע עיניים). הנתונים נאספים בטבלה בגודל $m \times m'$, כאשר X_{ij} הוא מספר הנתונים עם $I = i$ ו־$J = j$.

המודל מתאר התפלגות מולטינומית (שאנו מחזיקים במערך דו־ממדי מטעמי נוחות), כלומר $(X_{ij}) \sim M(m \times m', (p_{ij}))$, כאשר

$$P(I = i, J = j) = p_{ij}.$$

השערת האפס שאנו מבקשים לבדוק קובעת כי I, J בלתי תלויים. כלומר, יש התפלגויות $I \sim M(1; p_1, \dots, p_m)$ ו־$J \sim M(1; p'_1, \dots, p'_{m'})$, המגדירות עקב האי־תלות את ההתפלגות המשותפת לפי $P(I = i, J = j) = p_i p'_j$. אם כך,

$$H_0 : (\exists p_i, p'_j)\ p_{ij} = p_i p'_j.$$

אם השערת האפס נכונה, אז כדי לאמוד את המטריצה p_{ij} די לדעת מהם סכומי השורה והעמודה, ובמקרה זה:

תרגיל 5.2.11 הראה שתחת השערת האפס, בהנתן סכומי השורות והעמודות $X_{i\cdot}, X_{\cdot j}$, הערך של X_{ij} מתפלג היפרגאומטרית, עם התוחלת $E_{ij} = \frac{X_{i\cdot} X_{\cdot j}}{n}$.

הסטטיסטי למבחן זה מחושב לפי הנוסחה

$$T = \sum \frac{(O_{ij} - E_{ij})^2}{E_{ij}} \tag{5.2}$$

ושווה ל־

$$T = \sum \frac{(X_{ij} - \frac{X_{i\cdot} X_{\cdot j}}{n})^2}{\frac{X_{i\cdot} X_{\cdot j}}{n}}.$$

תרגיל 5.2.12 הראה שהסטטיסטי T עבור טבלה בגודל 2×2 הוא $\frac{n \det(X)^2}{X_{0\cdot} X_{1\cdot} X_{\cdot 0} X_{\cdot 1}}$, כאשר $X = \begin{pmatrix} X_{00} & X_{01} \\ X_{10} & X_{11} \end{pmatrix}$ היא מטריצת השכיחויות של המדגם. ערוך ניסוי המראה שההתפלגות של T תחת השערת האפס היא אכן χ^2_1.

מבחן לאי תלות

הלקוחות הנכנסים לבנק ניגשים לאחד משלושה דלפקי שירות. ספרי ההדרכה של הבנק טוענים שאין קשר בין סוג החשבון של הלקוח (ארבעה סוגים אפשריים) אל הדלפק שאליו הוא ניגש. מנהלת הסניף חושדת שאין זה כך. איך היא יכולה לבדוק את הטענה? (מן הסתם היא תקרא להתפלגות שמציע הספר 'השערת האפס', ותבדוק אותה במגמה להוכיח שהיא שגויה).

באופן פורמלי, בהמשך להערה 5.1.2, מדובר במרחב זוגות (סופי), ולכל לקוח מוצמד זוג משתנים מקריים (I_ℓ, J_ℓ) מאותה התפלגות. המשתנים המונים $X_{ij} = |\{\ell\colon I_\ell = i, J_\ell = j\}|$ מתפלגים, יחד, מולטינומית $(X_{ij}) \sim M(n; (p_{ij}))$.

נסמן ב־X_{ij} את מספר הלקוחות מטיפוס i המבצעים פעולה j. לפי המודל הכללי ביותר, יש מטריצה (p_{ij}) של הסתברויות, כך ש־$\sum p_{ij} = 1$, ו־ $(X_{ij}) \sim M(n; (p_{ij}))$. לפי השערת האפס, יש וקטורי הסתברות $(\alpha_i), (\beta_j)$ כך ש־$p_{ij} = \alpha_i\beta_j$. נסמן $X_{i\cdot} = \sum_j X_{ij}$, $X_{\cdot j} = \sum_i X_{ij}$, סכומי השורות והעמודות.

מתברר שתחת השערת האפס, בקירוב, $W = \sum \frac{(X_{ij} - \frac{X_{i\cdot}X_{\cdot j}}{n})^2}{\frac{X_{i\cdot}X_{\cdot j}}{n}} \sim \chi^2_{(a-1)(b-1)}$; ומכאן, כמקודם, אזור דחיה להשערה ברמת מובהקות α.

5.2.4 האם משתנים בלתי תלויים הם שווי התפלגות

נניח שמשתני האינדקס I, J מקבלים רק שני ערכים, נאמר $0, 1$. בסעיף 5.2.3 (ובעיקר בתרגיל 5.2.12) בנינו מבחן להשערה ש־I, J בלתי תלויים. **מבחן מק'נמר** בודק השערה עדינה יותר: בהנחה ש־I, J בלתי תלויים, האם הם שווי התפלגות?

אם כך, המודל קובע ש־$X_{ij} \sim M(n; (p_{ij}))$, והשערת האפס היא

$$H_0 : p_{01} = p_{10}.$$

תחת השערת האפס, הערכים הצפויים הם $\begin{pmatrix} X_{00} & \frac{X_{01}+X_{10}}{2} \\ \frac{X_{01}+X_{10}}{2} & X_{11} \end{pmatrix}$, ולכן הסטטיסטי של המבחן, המחושב גם כאן לפי הנוסחה (5.2), הוא

$$T' = \frac{(X_{01} - X_{10})^2}{X_{01} + X_{10}};$$

תחת השערת האפס $T' \sim \chi^2_1$.

5.3 ניתוח אשכולות

הסעיף האחרון עוסק בניתוח אשכולות (cluster analysis).

ניתוח אשכולות מקבל קבוצה של נקודות $x_1, \ldots, x_n \in \mathbb{R}^d$, ובאופן יותר כללי - מרחב מטרי בן n נקודות, כלומר פונקציית מרחק סימטרית $d(x_i, x_j)$ המציייתת לאי־שוויון המשולש. מטרתו של **אישכול קשיח** היא להחזיר **חלוקה** של הנקודות לאשכולות, כך שהנקודות באותו אשכול קרובות זו לזו, והנקודות באשכולות שונים רחוקות זו מזו. **אישכול רך** מתיימר לקבוע עבור כל נקודה את התפלגות השייכות שלה לאשכולות השונים.

5.3.1 מה אי־אפשר לעשות

נפתח בתוצאה שלילית חשובה, המראה ששלוש דרישות טבעיות מפונקציית אישכול הן חזקות מדי במשותף. נקבע $n \geq 2$. **פונקציית אישכול** היא פונקציה המקבלת

מרחב מטרי בגודל n ומחזירה חלוקה של הנקודות לאשכולות. נאמר שפונקציה כזו היא **עשירה** אם היא עשויה (א־פריורי) לקבל כל חלוקה. הפונקציה **אדישה למתיחה בקבוע** אם $f(\alpha\cdot d)=f(d)$ לכל קבוע חיובי α ומטריקה. המעבר ממטריקה d למטריקה d' הוא **עקבי** ביחס לפונקציית החלוקה f, אם לכל x_i,x_j באותו אשכול ביחס ל־$f(d)$ מתקיים $d'(x_i,x_j)\le d(x_i,x_j)$, ולכל x_i,x_j שאינם באותו אשכול מתקיים $d'(x_i,x_j)\ge d(x_i,x_j)$. פונקציית החלוקה **עקבית** אם שינוי עקבי אינו משנה את החלוקה.

משפט 5.3.1 (Kleinberg) *יהי $n\ge 2$. לא קיימת פונקציית אישכול (קשיח) עשירה, אדישה למתיחה בקבוע, ועקבית.*

הוכחה. **נניח ש־f היא פונקציה כזו. תהי d מטריקה כזו ש־$f(d)$ היא החלוקה לנקודונים, ותהי d' מטריקה כלשהי. אז יש קבוע גדול מספיק כך ש־$\alpha d'(x_i,x_j)>d(x_i,x_j)$ לכל i,j, ואז $f(d)=f(\alpha d')=f(d)$ לפי האדישות והעקביות. כלומר, f יודעת להחזיר רק את החלוקה הדיסקרטית.** □

לכן יש ליישם בכל מקרה שיטות שונות המבוססות על הנחות יסוד או היוריסטיקה שונה.

תרגיל 5.3.2 תהי f הפונקציה המקבלת מטריקה d, ומחזירה את החלוקה שבה $x_i\sim x_j$ אם ורק אם המרחק $d(x_i,x_j)$ קטן או שווה לכל מרחק אחר. הראה שהפונקציה הזו אדישה ועשירה, ושינוי עקבי של המטריקה מוביל לעידון של החלוקה.

5.3.2 ממוצעי־k

שיטת **ממוצעי־k** (k-means) מחזירה חלוקה ל־k אשכולות, כאשר k קבוע מראש. האלגוריתם עבור נקודות $x_1,\dots,x_n\in\mathbb{R}^d$ פשוט:

1. בחר k נקודות מרכז כלשהן, $c_1,\dots,c_k$.
2. חלק את הנקודות כך שהנקודה x_i שייכת לאשכול C_{j_0} אם $d(x_i,c_{j_0})\le d(x_i,c_j)$ לכל j.
3. החלף את c_j בממוצע כל הנקודות x_i השייכות לאשכול C_j.
4. חזור לשלב 2, עד שהשינוי בסכום המרחקים $\sum_{x_i\in C_j}d(x_i,c_j)$ נעשה זניח.

תהליך זה מוצא מינימום מקומי של הפונקציה $\sum_{x_i \in C_j} d(x_i, c_j)$ (עבור $x_1, \ldots, x_n$ הנתונים). עם זאת, החלוקה המתקבלת תלויה מאד במרכזים המקוריים. חסרון נוסף הוא שלא ברור א־פריורי באיזה k לבחור, ויש להריץ את האלגוריתם על ערכים הולכים וגדלים של k, עד שהשיפור בפונקציית המטרה נעשה זניח. החסרון העיקרי הוא רגישות יתר למבנה האפיני של המרחב, שממנו נגזרת חוסר יכולת לזהות מבנים מורכבים (כגון מעגל העוטף אשכול מובהק).

5.3.3 שיטת GMM

גם שיטה זו מטפלת בנקודות במרחב האוקלידי $\mathbb{R}^d$, אלא שהיא מספקת אישכול רך. נתבונן במודל הבא. קיימים $\mu_1, \ldots, \mu_k$ ומטריצות חיוביות $\Sigma_1, \ldots, \Sigma_k$. כדי לבחור את הנקודה x_i, בוחרים אינדקס j אקראי, ומגרילים $x_i \sim N(\mu_j, \Sigma_j)$. שחזור האשכולות נעשה במשולב עם אמידת הפרמטרים של ההתפלגויות הרב־נורמליות:

1. בחר $\mu_1 \ldots, \mu_k$ כלשהם, ומטריצות חיוביות $\Sigma_1, \ldots, \Sigma_k$.
2. לכל נקודה x_i חשב את הסיכויים לקבל x_i מכל אחת מההתפלגות $N(\mu_j, \Sigma_j)$; נרמל את וקטור הצפיפויות, לקבלת $w_{i1}, \ldots, w_{ik}$. מתייחסים אל המספרים w_{ij} כאל ההסתברות לכך ש־x_i מגיעה מהתפלגות $\#j$ דווקא.
3. אמוד מחדש את הפרמטרים μ_j, Σ_j, לפי הנקודות x_i וההסתברויות w_{ij}.
4. חזור לשלב 2, עד שהשינוי בנראות הכוללת נעשה זניח.

5.3.4 אישכול היררכי

אישכול היררכי אינו מספק חלוקה של הנקודות לאשכולות, אלא עץ מונוטוני של קבוצות, כך שהשורש הוא $\{x_1, \ldots, x_n\}$ והעלים הם יחידונים. אפשר לבנות עץ כזה בדרכים שונות:

1. **מלמעלה־למטה**: בכל שלב מפרקים את הנקודות הנתונות לשתי קבוצות (על־ידי ממוצעי־2 או שיטה אחרת), וממשיכים לפרק כל אשכול בנפרד.
2. **מלמטה־למעלה**: כאן מתחילים מן החלוקה ליחידונים, ומאחדים אשכולות בהדרגה. יתרונה של שיטה זו הוא בכך שהיא דטרמיניסטית. הכלל הוא שמאחדים בכל פעם את זוג האשכולות הקרובים ביותר, כאשר המרחק $d(C, C')$ מוגדר באחת הדרכים הבאות:
   (א) $d(C, C') = \min_{x \in C, y \in C'} d(x, y)$;
   (ב) $d(C, C') = \max_{x \in C, y \in C'} d(x, y)$;
   (ג) $d(C, C') = \frac{1}{|C| \cdot |C'|} \sum_{x \in C, y \in C'} d(x, y)$.

5.3.5 עץ פורש מינימלי

אלגוריתם זה מטפל במרחב מטרי כללי: רואים במרחב המטרי גרף ממושקל, ומוצאים בו את העץ הפורש המינימלי. העץ מגדיר אישכול הירכי.

5.3.6 אלגוריתמים מבוססי צפיפות

ע"ע DBSCAN, HDBSCAN*.