

# מבוא לאותות אקראיים

ד"ר דימה בחובסקי

5 באוקטובר 2026

This work is licensed under a [Creative Commons "Attribution-NonCommercial-ShareAlike 4.0 International"](https://creativecommons.org/licenses/by-nc-sa/4.0/) license.

# תוכן העניינים

|     |                                               |
|-----|-----------------------------------------------|
| 4   | 1. הסתברות וסטטיסטיקה                         |
| 5   | 1.1. משתנה אקראי בדיד                         |
| 14  | 2. משתנה אקראי רציף                           |
| 27  | 3. זוג משתנים אקראיים                         |
| 44  | 4. חיזוי לינארי                               |
| 51  | 5. Descriptive Statistics (סטטיסטיקה תיאורית) |
| 68  | 6. גרפים סטטיסטיים                            |
| 82  | 7. ריבועים פחותים לינאריים חד-משתניים         |
| 101 | א'. פתרונות לשאלות                            |
| 128 | ב'. מילון                                     |

## חלק I.

### הסתברות וסטטיסטיקה

# 1. משתנה אקראי בדיד

**מטרה:** לאפיין בצורה מתמטית תוצאה של ניסוי אקראי בעל תוצאות בדידות.

## תוכן הפרק

|    |                            |
|----|----------------------------|
| 5  | 1.1. הקדמה                 |
| 6  | 1.2. תכונות                |
| 6  | 1.2.1. אפיון               |
| 7  | 1.2.2. תוחלת (Expectation) |
| 8  | 1.2.3. שונות (Variance)    |
| 8  | 1.2.4. חישוב מספרי         |
| 9  | 1.3. התפלגויות נפוצות      |
| 9  | 1.3.1. התפלגות Bernoulli   |
| 10 | 1.3.2. התפלגות בינומית     |
| 11 | 1.3.3. התפלגות גאומטרית    |
| 12 | 1.4. שאלות                 |

## 1.1. הקדמה

**מטרה:** מושגי יסוד.

**הסתברות:** שכיחות יחסית של תופעה כלשהי אחרי אינסוף ניסויים אקראיים<sup>1</sup>.

**משתנה אקראי (random variable):** משתנה אקראי (נקרא גם משתנה מקרי) הוא מודל שמתאר קשר בין תוצאה של ניסוי אקראי למספר ממשי<sup>2</sup>. הסימון הוא  $X, Y$  ע"י אות גדולה, לדוגמה  $X, Y$  וכד'.

**משתנה אקראי בדיד (discrete random variable):** עבור משתנים אקראיים בדידים, תוצאות אפשריות של הניסוי  $x_k$  הן בדידות (= ניתנות למספור), כאשר מספר ערכי  $k$  אפשריים יכול להיות סופי או אינסופי.

**משתנה אקראי רציף (continuous random variable):** עבור משתנים אקראיים רציפים, תוצאות אפשריות של הניסוי מוגדרות על מקטע או אוסף של מקטעים.

**דוגמה 1.1:** (משתנה אקראי בדיד עם מספר תוצאות ניסוי סופי) משתנה אקראי המקשר בין תוצאת זריקת קוביה,  $\{1, 2, 3, 4, 5, 6\}$ , לבין ההסתברות לתוצאה זו.

**דוגמה 1.2:** (משתנה אקראי בדיד עם מספר תוצאות ניסוי אינסופי) משתנה אקראי המקשר בין מספר זריקות קוביה לבין ההסתברות לקבל את התוצאה "6".

<sup>1</sup>תופעות אחרות, שאינן קשורות לניסויים אקראיים, כגון סיכויי זכייה בבחירות וכד', אינן נכללות בהגדרה זאת ולא יטופלו במסגרת חוברת זו.

<sup>2</sup>קיימת הגדרה פורמלית מתמטית יותר, שלא תוצג בחוברת זו.

**דוגמה 1.3:** (משתנה אקראי רציף) משתנה אקראי המקשר בין זמן העבודה של מכונה מסוימת לבין הסתברות התקלה שלה.

## 1.2. תכונות

### 1.2.1. אפיון

**מטרה:** לאפיין משתנה אקראי בדיד.

**Probability density function (PDF):** קשר בין תוצאה של הניסוי  $x_k$  לבין ההסתברות של תוצאה זו:

$$(1.1) \quad p_X[x_k] = \Pr[X = x_k]$$

השימוש בסוגריים  $[\cdot]$  נועד להדגיש שמדובר בערכים בדידים, לעומת  $(\cdot)$  המשמש בהמשך לערכים רציפים.

בחוברת זו המושג PDF מתייחס גם למשתנים בדידים וגם למשתנים רציפים. בספרות לעתים מופיע מושג אחר, **probability mass function (PMF)**, שמטרתו להדגיש שמדובר במשתנה בדיד. באותם ספרים המושג PDF שמור למשתנים רציפים בלבד.

תכונות היסוד של משתנים אקראיים הן כדלהלן.

**תחום ערכי PDF:** תחום ערכי PDF הוא

$$(1.2) \quad 0 \leq p_X[x_i] \leq 1 \quad \forall i$$

**סכום ערכי PDF:** סכום ערכי PDF הוא

$$(1.3) \quad \sum_i p_X[x_i] = 1$$

**Cumulative distribution function (CDF):** ההסתברות שערך משתנה אקראי  $X$  קטן או שווה לערך נתון  $x$ ,

$$(1.4) \quad F_X(x) = \Pr(X \leq x), \quad x \in \mathbb{R}.$$

ערך  $x$  הוא ממשי ורציף.

**קשר בין PDF ל-CDF:** הקשר מתבסס על סכום של  $p_X[x_k]$  עבורם  $x_k \leq x$ :

$$(1.5) \quad F_X(x) = \sum_{k: x_k \leq x} p_X[x_k]$$

**דוגמה 1.4:** חשב את  $F_X(3.5)$ , כאשר המשתנה האקראי  $X$  מתאר תוצאה של זריקת קוביה.

פתרון: ע"י הצבה בנוסחה שלהלן מתקבל

$$F_X(x) = \sum_{k: x_k \leq x} p_X[x_k] = p_X[1] + p_X[2] + p_X[3] = \frac{1}{6} + \frac{1}{6} + \frac{1}{6} = \frac{1}{2}$$

**התנהגות פונקציונלית:** פונקציית CDF היא לא יורדת, ורציפה מימין.

הוכחה. עבור  $x_1 \leq x_2$  מתקיים  $\{X \leq x_1\} \subseteq \{X \leq x_2\}$ , לכן  $F_X(x_1) \leq F_X(x_2)$ .

הסתברות על מקטע: עבור  $x \in (a, b]$  מתקיים

$$(1.6) \quad \Pr(a < X \leq b) = F_X(b) - F_X(a)$$

**ערכי קצה (קיצון):** תחום ערכי CDF הוא  $0 \leq F_X(x) \leq 1$  (בהתאם לתכונה 1.2.1), כאשר עבור ערכי קיצון של  $x$  מתקיים

$$(1.7) \quad F_X(-\infty) = 0, \quad F_X(\infty) = 1$$

**הסתברות משלימה:** מאחר ומתקיים  $\Pr(X > x) + \Pr(X \leq x) = 1$ ,

$$(1.8) \quad \Pr(X > x) = 1 - F_X(x)$$

## 1.2.2. תוחלת (Expectation)

**מטרה:** לאפיין ניסוי אקראי ע"י מספר בודד.

**תוחלת:**

$$(1.9) \quad \mathbb{E}[X] = \sum_i x_i p_X[x_i]$$

הסימון ע"י האות  $E$  מבוסס על המילה Expectation (צפי). הסבר על הבחירה דווקא בשם הזה מופיע בהמשך, בהקשר של חיזוי ע"י קבוע.

**תכונות התוחלת:** תוחלת היא פעולה לינארית, והתוחלת של קבוע היא אותו הקבוע.

$$(1.10a) \quad \mathbb{E}[aX] = a\mathbb{E}[X]$$

$$(1.10b) \quad \mathbb{E}[b] = b$$

**תוחלת של פונקציה של משתנה אקראי:**

$$(1.11) \quad \mathbb{E}[g(X)] = \sum_i g(x_i) p_X[x_i]$$

**דוגמה 1.5:** בהתאם להגדרה,

$$(1.12) \quad \mathbb{E}[X^2] = \sum_i x_i^2 p_X[x_i]$$

**תוחלת של תוחלת:**

$$(1.13) \quad \mathbb{E}[\mathbb{E}[X]] = \mathbb{E}[X]$$

הוכחה. מאחר ותוחלת היא מספר קבוע, ניתן להציב  $\mathbb{E}[X] = b$  ולהיעזר בנוסחה (1.10).

**דוגמה 1.6:** בהתאם לנוסחה (1.10) מתקיים,

$$(1.14) \quad \mathbb{E}[X\mathbb{E}[X]] = \mathbb{E}[X] \mathbb{E}[X] = (\mathbb{E}[X])^2$$

### 1.2.3. שונות (Variance)

**מטרה:** לאפיין פיזור של תוצאות ניסוי אקראי סביב לתוחלת.

**שונות:** שונות של משתנה אקראי היא

$$(1.15) \quad \text{Var}[X] = \mathbb{E} \left[ \left( X - \mathbb{E}[X] \right)^2 \right]$$

**שונות - הגדרה חלופית:**

$$(1.16) \quad \text{Var}[X] = \mathbb{E} \left[ X^2 \right] - \left( \mathbb{E}[X] \right)^2$$

הוכחה.

$$(1.17) \quad \begin{aligned} \mathbb{E} \left[ \left( X - \mathbb{E}[X] \right)^2 \right] &= \mathbb{E} \left[ X^2 - 2X\mathbb{E}[X] + \left( \mathbb{E}[X] \right)^2 \right] \\ &= \mathbb{E} \left[ X^2 \right] - 2\mathbb{E}[X] \mathbb{E}[X] + \left( \mathbb{E}[X] \right)^2 \\ &= \mathbb{E} \left[ X^2 \right] - \left( \mathbb{E}[X] \right)^2 \end{aligned}$$

**תכונות השונות:** הכפלה בקבוע ושונות של קבוע:

$$(1.18) \quad \text{Var}[aX + b] = a^2 \text{Var}[X]$$

$$(1.18) \quad \text{Var}[b] = 0$$

הוכחה. ההוכחה היא במסגרת התרגיל 1.1.

תמיד מתקיים  $\text{Var}[X] \geq 0$ .

באופן כללי<sup>א</sup>,

$$\begin{aligned} \mathbb{E}[g(X)] &\neq g(\mathbb{E}[X]) \\ \text{Var}[g(X)] &\neq g(\text{Var}[X]) \end{aligned}$$

לדוגמה,

$$\begin{aligned} \mathbb{E}[X^2] &\neq \left( \mathbb{E}[X] \right)^2 \\ \text{Var}[X^2] &\neq \left( \text{Var}[X] \right)^2 \end{aligned}$$

<sup>א</sup>ניתוח פורמלי של הקשרים בין  $\mathbb{E}[g(X)]$  לבין  $g(\mathbb{E}[X])$  נקרא Jensen's inequality.

### 1.2.4. חישוב מספרי

**חישוב מתוך תוצאות הניסוי:** נתונות  $N$  תוצאות ניסוי אקראי  $x_1, \dots, x_N$ . במקרה הזה חישוב של תוחלת מוחלף ע"י חישוב ממוצע, באופן הבא.

$$(1.19) \quad \mathbb{E}[X] = \frac{1}{N} \sum_{i=1}^N x_i$$

$$(1.20) \quad \text{Var}[X] = \frac{1}{N} \sum_{i=1}^N (x_i - \mathbb{E}[X])^2$$

אין הבדל בחישוב עבור תוצאות ניסוי, שהן רציפות או בדידות.

ישנן שיטות נוספות לחישוב שונות.

### 1.3. התפלגויות נפוצות

**התפלגות:** מודל אקראי ייחודי שימושי/נפוץ, שקיבל שם בגלל חשיבותו. משמש לתיאור/אפיון של משתנה אקראי.

#### 1.3.1. התפלגות Bernoulli

**מטרה:** לתאר ניסוי בעל 2 תוצאות אפשריות:

- תוצאה "0" עם הסתברות  $1 - p$ .
- תוצאה "1" עם הסתברות  $p$ .

PDF: עבור  $X \sim \text{Ber}(p)$  מתקיים (איור 1.1א)

$$(1.21) \quad \begin{aligned} p_X[k] &= \begin{cases} 1 - p & k = 0 \\ p & k = 1 \end{cases} \\ &= p^k (1 - p)^{1-k} \quad k = 0, 1 \\ &= (1 - p)\delta[k] + p\delta[k - 1] \end{aligned}$$

CDF: עבור  $X \sim \text{Ber}(p)$  מתקיים (איור 1.1ב)

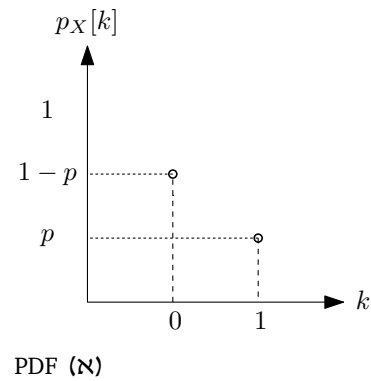
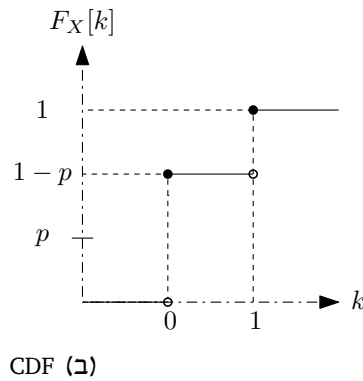
$$(1.22) \quad F_X(x) = \begin{cases} 0 & x < 0 \\ 1 - p & 0 \leq x < 1 \\ 1 & 1 \leq x \end{cases}$$

**דוגמה 1.7:** עבור "מטבע הוגן",  $p = 0.5$ .

**דוגמה 1.8:** חשב תוחלת ושונות של התפלגות Bernoulli.

פתרון: חישוב השונות נעזר ב- $\mathbb{E}[X^2]$ , המחושב לפי הנוסחה לתוחלת של פונקציה של משתנה אקראי, משוואה (1.11).

$$\begin{aligned} \mathbb{E}[X] &= \sum_i x_i p_X[x_i] \\ &= (k=0) p_X[0] + (k=1) p_X[1] \\ &= 0 \cdot (1 - p) + 1 \cdot p = p \\ \text{Var}[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 \end{aligned}$$



איור 1.1: התפלגות Bernoulli.

$$\begin{aligned}
 &= \left[ (0)^2 p_X[0] + (1)^2 p_X[1] \right] - p^2 \\
 &= 0^2 \cdot (1-p) + 1^2 \cdot p - p^2 = p - p^2 \\
 &= p(1-p)
 \end{aligned}$$

### 1.3.2. התפלגות בינומית

**מטרה:** אפיון ניסוי אקראי, כאשר:

- נתונים  $n$  ניסויים בלתי תלויים.
- כל אחד בעל התפלגות Bernoulli עם סיכוי הצלחה  $p$ ,  $\text{Ber}(p)$ .
- נדרשת ההסתברות ל- $k$  תוצאות "1" בדיוק, מתוך סה"כ  $n$  תוצאות.

PDF: עבור  $X \sim \text{Bin}(n, p)$  ניתן לפרק את החישוב ל-3 חלקים:

1. ישנם  $\binom{n}{k} = \frac{n!}{k!(n-k)!}$  אפשרויות ל- $k$  הצלחות, מתוך  $n$  ניסויים.

2. הסתברות של  $k$  הצלחות (תוצאות "1") היא  $p^k$ .

3. הסתברות של  $n-k$  כשלונות (תוצאות "0") היא  $(1-p)^{n-k}$ .

מתוך איחוד של שלושת ההסתברויות,

$$(1.23) \quad p_X[k] = \binom{n}{k} p^k (1-p)^{n-k}$$

CDF: אין ביטוי "סגור", אלא רק חישוב סכום של PDF ע"פ תכונה 1.2.1,

$$(1.24) \quad F_X(x) = \begin{cases} 0 & x < 0 \\ \sum_{k=0}^{\lfloor x \rfloor} p_X[k] & 0 \leq x < n \\ 1 & n \leq x \end{cases}$$

הסימון  $\lfloor \cdot \rfloor$  מעל הסכום הוא עבור עיגול כלפי מטה.

**תוחלת ושונות:** עבור  $X \sim \text{Bin}(n, p)$ ,

$$(1.25) \quad \mathbb{E}[X] = np$$

$$(1.26) \quad \text{Var}[X] = np(1 - p)$$

**קשר להתפלגות Bernoulli:** עבור ניסוי יחיד ( $n = 1$ ), מתקיים הקשר  $X \sim \text{Bin}(1, p) \Leftrightarrow X \sim \text{Ber}(p)$ .

**דוגמה 1.9:** זורקים קובייה 6 פעמים. חשב הסתברות לבדיק שתי תוצאות "6".

פתרון: ניתן לנסח את הבעיה כסיכוי להצלחה של  $k = 2$  ניסויים בדיק מתוך  $n = 6$ , כאשר סיכוי להצלחה בניסוי הוא  $p = \frac{1}{6}$ .

$$X \sim \text{Bin}\left(6, \frac{1}{6}\right)$$

$$p_X[2] = \binom{6}{2} \left(\frac{1}{6}\right)^2 \left(1 - \frac{1}{6}\right)^{6-2} \approx 0.20$$

### 1.3.3. התפלגות גאומטרית

**מטרה:** עבור ניסויים אקראיים בעלי התפלגות Bernoulli, ההסתברות להצלחה ("1") בניסוי  $k$ , אחרי רצף  $(k - 1)$  פעמים של "0".

**PDF, CDF, תוחלת ושונות:** עבור התפלגות גאומטרית,  $X \sim \text{Geo}(p)$ , מתקיים

$$(1.27) \quad p_X[k] = p(1 - p)^{k-1}, \quad k = 1, 2, \dots$$

$$(1.28) \quad F_X[k] = 1 - (1 - p)^k$$

$$(1.29) \quad \mathbb{E}[X] = \frac{1}{p}$$

$$(1.30) \quad \text{Var}[X] = \frac{1 - p}{p^2}$$

**דוגמה 1.10:** מהי ההסתברות לא לקבל תוצאה "6" ב-6 הזריקות הראשונות של קובייה?

פתרון: נגדיר משתנה אקראי  $X$  שהוא מספר זריקות של הקובייה עד להופעת ה-"6" הראשון.

- סיכוי ל-"6" דווקא בזריקה ראשונה הוא  $p_X[1] = \frac{1}{6}$ .
- סיכוי ל-"6" דווקא בזריקה שניה הוא סיכוי של לא "6" בזריקה ראשונה כפול סיכוי של "6" בזריקה השניה,  $p_X[2] = \frac{5}{6} \cdot \frac{1}{6}$ .
- סיכוי ל-"6" דווקא בזריקה שלישית הוא  $p_X[3] = \left(\frac{5}{6}\right)^2 \cdot \frac{1}{6}$ .
- סיכוי ל-"6" דווקא בזריקה שישית הוא  $p_X[6] = \left(\frac{5}{6}\right)^5 \cdot \frac{1}{6}$ .

לסיכום,

$$(1.31) \quad \begin{aligned} \Pr(X > 6) &= 1 - \Pr(X \leq 6) \\ &= 1 - (p_X[1] + p_X[2] + \dots + p_X[6]) \\ &= 1 - F_X(6) = \left(1 - \frac{1}{6}\right)^6 \approx 0.3349 \end{aligned}$$

**Mode:** ערך  $x_k$ , עבורו PDF מקבל ערך מקסימלי,

$$(1.32) \quad x_k = \underset{x_k}{\operatorname{argmax}} p_X[x_k]$$

**חציון (median):** ערך  $m$  עבורו מתקיים  $\Pr(X \leq m) \geq \frac{1}{2}$  וגם  $\Pr(X \geq m) \geq \frac{1}{2}$  בו-זמנית, נקרא ערך חציון. **קשר בין תוחלת לחציון:** ישנו אי שוויון מעניין, שמגדיר הפרש מקסימלי בין תוחלת לחציון,

$$(1.33) \quad |\mathbb{E}[X] - m| \leq \sqrt{\operatorname{Var}[X]},$$

**דוגמה 1.11:** ניתן להתרשם מחדשות, מה הערך של משכורת ממוצעת, ומה הערך של משכורת חציונית.

## 1.4. שאלות

1.1 הוכח את המשוואה (1.18א).

1.2 הסיכוי למספר טבעי מסויים  $k$  מתוך אוסף כל מספרים טבעיים  $\{1, 2, 3, \dots\}$  נתון ע"י הסתברות

$$p[k] = \begin{cases} \frac{4}{7} & k = 1 \\ \frac{2}{7} & k = 2 \\ \left(\frac{1}{8}\right)^{k-2} & k \geq 3 \end{cases}$$

מהו הסיכוי עבור  $p[K \geq 4]$  ?

1.3 (סימלציה) יש לעשות סימולציה לחישוב של תוצאת זריקה של זוג קוביות.

1.4 ניסוי אקראי מוגדר כזריקת מטבע הוגן  $n$  פעמים ברציפות.

(א) מהו הסיכוי ל- $n$  אחדות ברצף?

(ב) חוזרים על הניסוי  $N = 2^n$  פעמים. מהי הסתברות, שמתוך  $N$  תוצאות יהיה לפחות פעם אחת  $n$  אחדות ברצף.

(ג) חשב ערך מספרי של הסתברות זו עבור  $n = 10$ .

1.5 מערכת תקשורת מסויימת אמורה לעבוד עם סיכוי שגיאה של  $10^{-3}$ , משמעותה שגיאה אחת ל-1000 ביט. לקוחות שונים החליטו לבדוק את הביצועים של המערכת:

(א) לקוח שידר 1000 ביט וקיבל 0 שגיאות.

(ב) לקוח שידר 1000 ביט וקיבל 1 שגיאות.

(ג) לקוח שידר 1000 ביט וקיבל 3 שגיאות.

שאלות:

(א) מהי הסתברות עבור התוצאה של לקוח א'?

(ב) מהי הסתברות עבור התוצאה של לקוח ב'?

(ג) מהי הסתברות עבור 3 שגיאות או יותר עבור לקוח ג'?

1.6 נתונה מערכת תקשורת המשדרת הודעות באורך 2 ביטים. כדי לזהות שגיאות, המערכת משתמשת בפרוטוקול חזרה (Repetition Code): כל הודעה משודרת פעמיים ברצף (סך הכל 4 ביטים לשידור). בצד המקלט, מתבצעת בדיקת תקינות על ידי השוואת זוג הביטים הראשון (ההודעה המקורית) לזוג הביטים השני (החזרה):

- אם הזוגות זהים, המערכת מניחה שהשדר תקין ומקבלת אותו.
- אם הזוגות שונים, המערכת מזהה שיש טעות ודוחה את השדר.

לדוגמה: עבור המידע "00", המערכת תשדר "0000".

נתון כי הסתברות השגיאה לביט בודד היא  $p$ , והשגיאות בביטים השונים הן בלתי תלויות.

1. מהי הסתברות השגיאה בשדר, במערכת המשדרת 2 ביטים בלבד ללא מנגנון החזרה (כלומר, השדר נחשב שגוי אם לפחות ביט אחד הגיע משובש)?

2. מהי ההסתברות שהשדר הכפול (4 ביטים) יתקבל באופן תקין לחלוטין? (כלומר, המצב שבו כל 4 הביטים ששודרו נקלטו בצד השני ללא שום שגיאה).

3. מהי ההסתברות לשגיאה בלתי מזוהה במערכת עם מנגנון החזרה (4 ביטים)? (כלומר, המצב בו המקלט מקבל את השדר כ"תקין" כי הזוגות זהים, אך המידע שהתקבל שגוי ביחס למידע המקורי).

1.7 בשנת 2003 נתוני השכר בישראל הם:

- שכר ממוצע של 8,027 ש"ח
- שכר חציוני של 6,543 ש"ח.

המשמעות היא:

- (א) רוב האנשים מקבלים שכר יותר נמוך משכר ממוצע.
- (ב) רוב האנשים מקבלים שכר יותר גבוה משכר ממוצע.
- (ג) בקירוב, חצי האנשים מקבלים שכר מעל הממוצע, וחצי מתחת.
- (ד) רוב האנשים מקבלים שכר יותר נמוך משכר חציוני.
- (ה) רוב האנשים מקבלים שכר יותר גבוה משכר חציוני.
- (ו) בקירוב, חצי האנשים מקבלים שכר מעל ערך חציוני, וחצי מתחת.

1.8 נתונות זוג קוביות. מהי ההסתברות לתוצאה (6, 6) בנסיבות הבאות:

(א) רק אחרי לפחות 10 זריקות קוביה, או יותר?

(ב) לפחות פעם אחת (פעם אחת או יותר), מתוך 10 זריקות.

(ג) רק פעם אחת בלבד, מתוך 10 זריקות.

## 2. משתנה אקראי רציף

**מטרה:** לאפיין בצורה מתמטית תוצאה של ניסוי אקראי בעל תוצאות רציפות.

### תוכן הפרק

|    |                                        |    |
|----|----------------------------------------|----|
| 14 | 2.1. אפיון                             | 14 |
| 15 | 2.2. התפלגויות                         | 15 |
| 15 | 2.2.1. התפלגות אקספוננציאלית (מעריכית) | 15 |
| 17 | 2.2.2. התפלגות אחידה                   | 17 |
| 19 | 2.2.3. התפלגות גאוסית                  | 19 |
| 21 | 2.2.4. התפלגות chi-square              | 21 |
| 23 | 2.2.5. התפלגות Student t               | 23 |
| 25 | 2.3. שאלות                             | 25 |

### 2.1. אפיון

**מטרה:** לחדד הבדלים בין משתנים בדידים ורציפים.

CDF: הגדרה של CDF (הגדרה 1.2.1) זהה בין משתנים בדידים ורציפים: ההסתברות שערך משתנה אקראי  $X$  קטן או שווה לערך נתון  $x$ ,

$$(2.1) \quad F_X(x) = \Pr(X \leq x), \quad x \in \mathbb{R}.$$

PDF רציף: עבור PDF רציף מתקיים

$$(2.2) \quad \Pr(X \leq x) = \Pr(X < x)$$

(PDF) Probability density function: נניח שפונקציית CDF גזירה עבור כל ערכי  $x$ . נוכל להגדיר,

$$(2.3) \quad f_X(x) = \frac{dF_X(x)}{dx}$$

$$(2.3) \quad F_X(x) = \int_{-\infty}^x f_X(p) dp$$

הסתברות על מקטע: בהתבסס על הגדרת PDF/CDF,

$$(2.4) \quad \Pr(a < X \leq b) = \int_a^b f_X(x) dx = F_X(b) - F_X(a)$$

תכונות PDF: תחום ערכים ו"סכום" ערכים,

$$(2.5) \quad f_X(x) \geq 0$$

$$(2.5) \quad \int_{-\infty}^{\infty} f_X(x) dx = 1 = F_X(\infty)$$

PDF של משתנה אקראי אינו מייצג הסתברות באופן ישיר, ואילו CDF כן מייצג הסתברות. עבור משתנה אקראי רציף יכול אף להתקיים  $f_X(x) > 1$ .

**הסתברות לערך בודד:** מאחר ומדובר באינסוף ערכים,

$$(2.6) \quad \Pr(X = a) = 0 \quad \forall a \in \mathbb{R}.$$

ניתן לרשום PDF של משתנה בדיד ע"י פונקציה רציפה מהצורה

$$(2.7) \quad f_X(x) = \sum_k p_X[x_k] \delta(x - x_k).$$

רישום זה מקשר בין  $F_X(x)$  קבועה למקוטעין של משתנה אקראי בדיד, לבין הנגזרת שלה, שהיא ה-PDF.

**תוחלת:** תוחלת של משתנה אקראי רציף היא

$$(2.8) \quad \mathbb{E}[X] = \int_{-\infty}^{\infty} x f_X(x) dx$$

**תוחלת של פונקציה של משתנה אקראי:**

$$(2.9) \quad \mathbb{E}[g(X)] = \int_{-\infty}^{\infty} g(x) f_X(x) dx$$

מאחר וחשוב השונות (סעיף 1.2.3) מתבסס על נוסחת התוחלת, הוא זהה עבור משתנים בדידים ורציפים.

**Mode:** ערך  $x_0$ , עבורו PDF מקבל ערך מקסימלי,

$$(2.10) \quad x_0 = \operatorname{argmax}_x f_X(x)$$

הגדרה של ערך חציוני זהה עבור משתנים בדידים ורציפים.

## 2.2. התפלגויות

### 2.2.1. התפלגות אקספוננציאלית (מעריכית)

**מטרה:** לאפיין תופעות עבורן ההסתברות משתנה באופן אקספוננציאלי (עם הזמן).

PDF: עבור  $X \sim \text{Exp}(\lambda)$  מתקיים<sup>1</sup>

$$(2.11) \quad f_X(x) = \begin{cases} \lambda \exp(-\lambda x) & x \geq 0 \\ 0 & \text{אחרת} \end{cases}$$

**דוגמה 2.1:** חשב תוחלת של  $X$ , כאשר  $X \sim \text{Exp}(\lambda)$ .

פתרון:

$$(2.12) \quad \mathbb{E}[X] = \lambda \int_0^\infty x \exp(-\lambda x) dx = \frac{1}{\lambda}$$

**שונות:** כאשר  $X \sim \text{Exp}(\lambda)$ ,

$$(2.13) \quad \text{Var}[X] = \frac{1}{\lambda^2}$$

הוכחה. ע"י ההצבה  $y = \lambda x$ ,

$$\begin{aligned} \mathbb{E}[X^2] &= \int_0^\infty \lambda x^2 \exp(-\lambda x) dx \\ &= \frac{1}{\lambda^2} \left[ -2e^{-y} - 2ye^{-y} - y^2 e^{-y} \right]_0^\infty = \frac{2}{\lambda^2} \end{aligned}$$

**Mode:**  $x = 0$

**חציון:**  $x = \frac{\ln(2)}{\lambda}$

**דוגמה 2.2:** חשב CDF של  $X$ , כאשר  $X \sim \text{Exp}(\lambda)$ .

פתרון:

$$(2.14) \quad \begin{aligned} F_X(x) &= \int_{-\infty}^x f_X(x) dx \\ &= \begin{cases} 1 - \exp(-\lambda x) & x \geq 0 \\ 0 & x < 0 \end{cases} \end{aligned}$$

**דוגמה 2.3:** מודל של הופעת תקלות במערכת מסויימת מתואר ע"י ההתפלגות  $X \sim \text{Exp}(\lambda)$ , כאשר  $\lambda = \frac{1}{1000} \left[ \frac{1}{\text{שעה}} \right]$  (יש יחידות פיזיקליות ל- $\lambda$ ). שאלות:

(א) מהו זמן ממוצע לתקלה?

<sup>1</sup>בספרות (וגם ב-MATLAB) מופיעה הגדרה נוספת:

$$f_X(x) = \begin{cases} \frac{1}{\mu} \exp\left(-\frac{x}{\mu}\right) & x \geq 0 \\ 0 & \text{אחרת} \end{cases},$$

כאשר  $\mu = \frac{1}{\lambda}$ .

(ב) מהו סיכוי לתקלה במרווח זמן  $[1000, \infty]$ ?

(ג) מהו סיכוי לתקלה ב-100 השעות הראשונות של הפעילות?

(ד) מהו סיכוי לתקלה אחרי 100 שעות ולפני 1000 שעות (מרווח זמן  $[100, 1000]$ )?

(ה) מהו משך הזמן עבורו הסיכוי לתקלה הוא  $\frac{1}{2}$ ?

פתרון: משתנה אקראי  $X$  מתאר זמן תקלה.

$$\mathbb{E}[X] = \frac{1}{\lambda} = 1000 \text{ [שעות]}$$

$$\Pr(X > 1000) = 1 - F_X(1000) = 1 - \left[ 1 - \exp\left(-\frac{1000}{1000}\right) \right] \approx 0.36$$

$$\Pr(X \leq 100) = F_X(100) \approx 0.0953$$

$$\begin{aligned} \Pr(100 \leq X \leq 1000) &= F_X(1000) - F_X(100) \\ &= 0.6321 - 0.0953 \approx 0.537 \end{aligned}$$

$$\Pr(X \leq x) = \Pr(X > x) = \frac{1}{2}$$

$$\Rightarrow F_X(x) = 1 - F_X(x) = \frac{1}{2}$$

$$\Rightarrow x = \frac{\ln\left(\frac{1}{2}\right)}{-\frac{1}{1000}} = \frac{\ln(2)}{\lambda} \approx 693 \text{ [שעות]}$$

**קשר להתפלגות גאומטרית:** אם  $X \sim \text{Exp}(\lambda)$  ומתקיים הקשר

$$Y = \lfloor X \rfloor,$$

אז  $Y \sim \text{Geo}(1 - \exp(-\lambda))$ , כאשר  $\lfloor \cdot \rfloor$  מסמל עיגול למספר השלם כלפי מטה.

## 2.2.2. התפלגות אחידה

**מטרה:** מודל של תופעות בעלות אותה הסתברות על מקטע רציף.

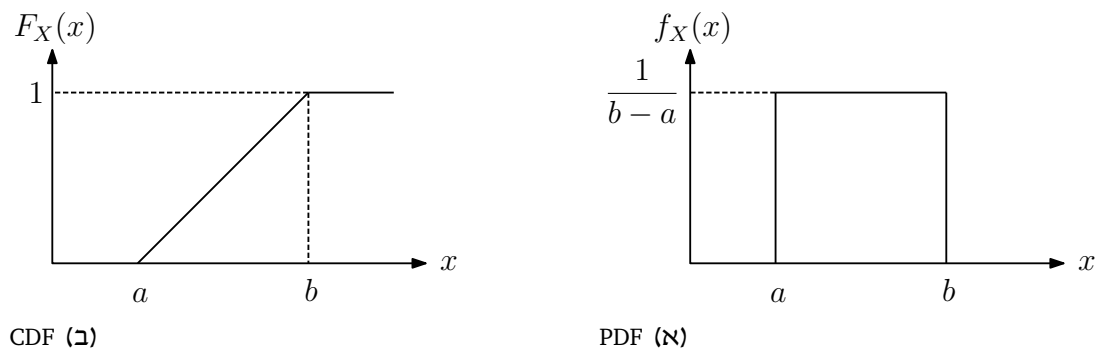
**PDF:** עבור  $X \sim U[a, b]$  מתקיים (איור 2.1א)

$$(2.15) \quad f_X(x) = \begin{cases} \frac{1}{b-a} & a \leq x \leq b \\ 0 & \text{אחרת} \end{cases}$$

עבור  $X \sim U(a, b)$  יש להחליף  $\leq$  ל- $<$ .

**CDF:** עבור  $X \sim U[a, b]$  מתקיים (איור 2.1ב)

$$(2.16) \quad F_X(x) = \int_{-\infty}^x f_X(x) dx = \begin{cases} 0 & x < a \\ \int_a^x \frac{1}{b-a} dx = \frac{x-a}{b-a} & a \leq x < b \\ 1 & b \leq x \end{cases}$$



איור 2.1: התפלגות אחידה.

**דוגמה 2.4:** יש לחשב תוחלת של  $X$ , כאשר  $X \sim U[a, b]$ .

פתרון: בהתאם לחישוב ע"פ ההגדרה,

$$(2.17) \quad \mathbb{E}[X] = \int_a^b x \frac{1}{b-a} dx = \frac{1}{b-a} \cdot \frac{x^2}{2} \Big|_a^b = \frac{a+b}{2}$$

**דוגמה 2.5:** יש לחשב  $\mathbb{E}[X^2]$ , כאשר  $X \sim U[a, b]$ .

פתרון:

$$(2.18) \quad \mathbb{E}[X^2] = \int_a^b x^2 \frac{1}{b-a} dx = \frac{1}{b-a} \cdot \frac{x^3}{3} \Big|_a^b = \frac{1}{3} \frac{b^3 - a^3}{b-a} = \frac{b^2 + ab + a^2}{3}$$

**דוגמה 2.6:** יש לחשב  $\text{Var}[X]$ , כאשר  $X \sim U[a, b]$ .

פתרון:

$$(2.19) \quad \text{Var}[X] = \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{b^2 + ab + a^2}{3} - \frac{(b+a)^2}{4} = \frac{(b-a)^2}{12}$$

קשר בין הדוגמאות, שהוא קיצור דרך בהרבה תרגילים, ע"פ ההגדרה של השונות:

$$\mathbb{E}[X^2] = \text{Var}[X] + (\mathbb{E}[X])^2$$

**דוגמה 2.7:** יש לחשב ערך חציון.

פתרון: ניתן לזהות בקלות, שמתקיים

$$(2.20) \quad \Pr(X \geq \mathbb{E}[X]) = \Pr(X \leq \mathbb{E}[X]) = \frac{1}{2}$$

דוגמאות לעיל מסכמות את התכונות של התפלגות אחידה.

**דוגמה 2.8:** נתון

$$X \sim U[0, 2]$$

$$Y = X^2$$

יש לחשב  $\mathbb{E}[Y]$ ,  $\text{Var}[Y]$ .

פתרון:

$$\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}[X^2] = \text{Var}[X] + (\mathbb{E}[X])^2 \\
&= \int_0^2 x^2 \frac{1}{2-0} dx = \frac{4}{3} \\
\mathbb{E}[Y^2] &= \mathbb{E}[X^4] = \int_0^2 x^4 \frac{1}{2-0} dx = \frac{1}{2} \frac{x^5}{5} \Big|_0^2 = \frac{16}{5} \\
\text{Var}[Y] &= \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = \frac{16}{5} - \left(\frac{4}{3}\right)^2 = \frac{64}{45}
\end{aligned}$$

להתפלגות אחידה קיימת גם גרסה בדידה (כגון תיאור של תוצאת זריקת קוביה).

### 2.2.3. התפלגות גאוסית

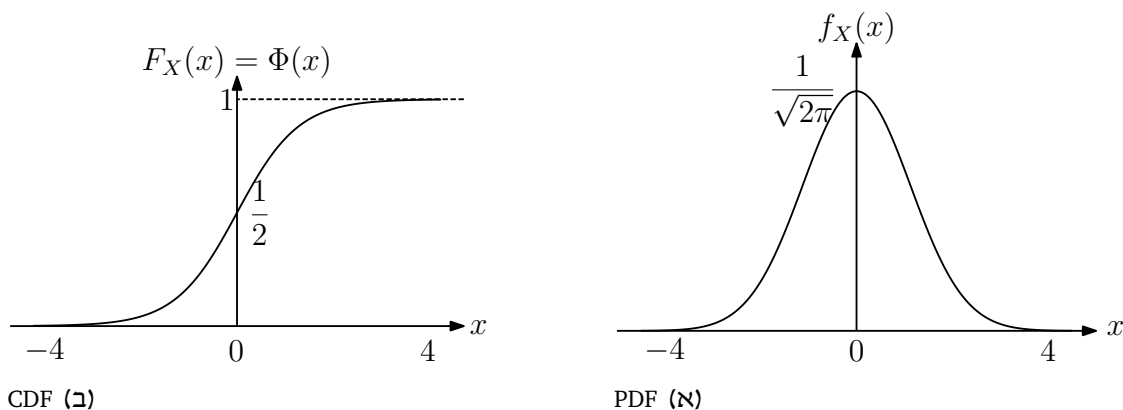
**מטרה:** מודל של תופעות רבות בטבע. שימושי ביותר בזכות משפט הגבול המרכזי, הגורס כי סכום של מספר מספיק גדול של משתנים אקראיים מתפלג בקירוב גאוסית.

PDF: עבור משתנה אקראי גאوسی  $Y \sim \mathcal{N}(\mu, \sigma^2)$ , בעל תוחלת  $\mu$  ושונות  $\sigma^2$ , מתקיים (איור 2.2א, עבור  $\mu = 0, \sigma = 1$ ).

$$(2.21) \quad f_X(x) = \frac{1}{\sigma\sqrt{2\pi}} \exp \left[ -\frac{(x-\mu)^2}{2\sigma^2} \right]$$

:CDF

- ניתן לחישוב נומרי (מספרי) בלבד.
- עבור מקרה פרטי של  $\mathcal{N}(0, 1)$ , מסומן ע"י  $\Phi(x)$  (איור 2.2ב).



איור 2.2: התפלגות גאוסית עבור  $X \sim \mathcal{N}(0, 1)$

הוספת קבוע וכפל בקבוע:

$$(2.22) \quad X \sim \mathcal{N}(0, b) \Rightarrow \sigma X \sim \mathcal{N}(0, b\sigma^2)$$

$$(2.22) \quad X \sim \mathcal{N}(a, 1) \Rightarrow \mu + X \sim \mathcal{N}(a + \mu, 1)$$

$$(2.22) \quad X \sim \mathcal{N}(0, 1) \Rightarrow \mu + \sigma X \sim \mathcal{N}(\mu, \sigma^2)$$

מעבר ל- $\mathcal{N}(0, 1)$ :

$$(2.23) \quad Y \sim \mathcal{N}(\mu, \sigma^2) \Rightarrow \frac{Y - \mu}{\sigma} \sim \mathcal{N}(0, 1), \quad \Pr(Y < a) = \Phi\left(\frac{a - \mu}{\sigma}\right)$$

הוכחה.

$$(2.24) \quad Y - \mu \sim \mathcal{N}(0, \sigma^2)$$

$$(2.25) \quad \frac{1}{\sigma}(Y - \mu) \sim \mathcal{N}(0, 1)$$

Q-function: עבור  $X \sim \mathcal{N}(0, 1)$ , ניתן להגדיר (חלופה שימושית מאוד ל-CDF)

$$(2.26) \quad Q(x) = \Pr(x < X) = \frac{1}{\sqrt{2\pi}} \int_x^\infty \exp\left(-\frac{s^2}{2}\right) ds.$$

מאפיינים של Q-function:

$$(2.27) \quad Q(x) = 1 - \Phi(x) = \Phi(-x)$$

$$(2.27) \quad Q(-x) = 1 - Q(x)$$

$$(2.27) \quad Q(0) = \frac{1}{2}$$

עבור תוחלת ושונות שרירותיים: בהינתן משתנה אקראי גאوسی  $Y \sim \mathcal{N}(\mu, \sigma^2)$

$$(2.28) \quad \Pr(Y > y) = Q\left(\frac{y - \mu}{\sigma}\right)$$

דוגמה 2.9: נתון הקשר

$$Y = X + a,$$

כאשר  $a$  הוא קבוע ו- $X \sim \mathcal{N}(0, \sigma^2)$ . מהי ההסתברות  $\Pr(Y \leq 0)$ ?

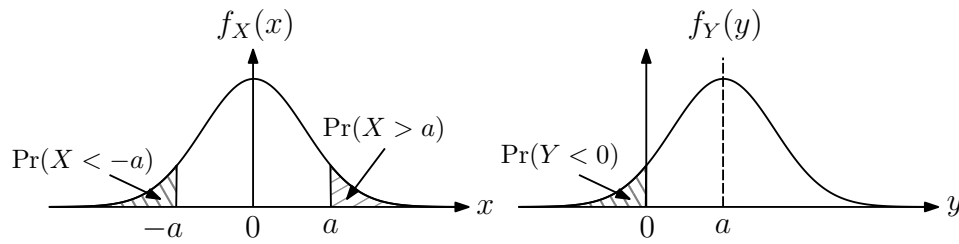
פתרון: בהתאם לאיור 2.3,

$$Y \sim \mathcal{N}(a, \sigma^2)$$

$$\begin{aligned} \Pr(Y < 0) &= F_Y(0) = \Phi\left(\frac{0 - a}{\sigma}\right) = \Phi\left(-\frac{a}{\sigma}\right) \\ &= Q\left(\frac{a}{\sigma}\right) \end{aligned}$$

הפתרון המספרי החלופי הוא ע"י ביצוע של מספר מספיק גדול של ניסויים אקראיים. שלבים:

- להגדיל  $N$  ערכים של משתנה אקראי  $X$ .
- לחשב  $Y$ .



איור 2.3: ציור נלווה לדוגמה 2.9.

- לספור מספר תוצאות ניסויים העומדים בקריטריון  $Y < 0$ . נסמן את התוצאה ב- $n$ .
- $\Pr(Y < 0) \approx \frac{n}{N}$

**משפט הגבול המרכזי:** נתונים  $n$  משתנים אקראיים בלתי תלויים בעלי אותה התפלגות כלשהי  $X_1, X_2, \dots, X_n$  בעלי תוחלת  $\mu$  ושונות  $\sigma^2$ . בהתאם למשפט, הממוצע מתפלג גאוסית

$$(2.29) \quad \lim_{n \rightarrow \infty} \frac{1}{n} \sum_{k=1}^n X_k \sim \mathcal{N}(\mu, \sigma^2)$$

בפועל, עבור  $n \gtrsim 30$ , המשפט מתקיים בקירוב סביר עבור כל סוג התפלגות, לרבות התפלגויות בדידות.

## 2.2.4. התפלגות chi-square

**מטרה:** לאפיין סכום של ריבועים של משתנים גאוסיים סטנדרטיים בלתי תלויים, הגודל שעומד מאחורי כל הערכת שונות וכל סף של מרחק ריבועי.

**משתנה chi-square:** יהיו  $Z_1, Z_2, \dots, Z_k$  משתנים נורמליים סטנדרטיים בלתי תלויים,  $Z_i \sim \mathcal{N}(0, 1)$ . סכום הריבועים שלהם

$$(2.30) \quad Q = \sum_{i=1}^k Z_i^2$$

מתפלג chi-square עם  $k$  דרגות חופש, ומסומן  $Q \sim \chi_k^2$ .

**תחום ערכים ודרגות חופש:** שתי תכונות נובעות ישירות ממשוואה (2.30): מתקיים  $Q \geq 0$ , מאחר ומדובר בסכום של ריבועים, והפרמטר  $k$  אינו אלא מספר האיברים הריבועיים הבלתי תלויים המחוברים.

**PDF:** פונקציית צפיפות ההסתברות היא (איור 2.4ב)

$$(2.31) \quad f(q; k) = \frac{1}{2^{k/2} \Gamma(k/2)} q^{k/2-1} e^{-q/2}, \quad q > 0,$$

כאשר  $\Gamma(\cdot)$  היא פונקציית גמא. הצפיפות אסימטרית ימינה, והאסימטריה קטנה ככל ש- $k$  גדל.

**מומנטים:**

$$(2.32) \quad \mathbb{E}[Q] = k,$$

$$(2.33) \quad \text{Var}[Q] = 2k.$$

**אדיטיביות:** עבור  $Q_1 \sim \chi_{k_1}^2$  ו- $Q_2 \sim \chi_{k_2}^2$  בלתי תלויים,

$$(2.34) \quad Q_1 + Q_2 \sim \chi_{k_1+k_2}^2.$$

דרגות החופש מתחברות בדיוק כפי שמתחברים האיברים הריבועיים של משוואה (2.30).

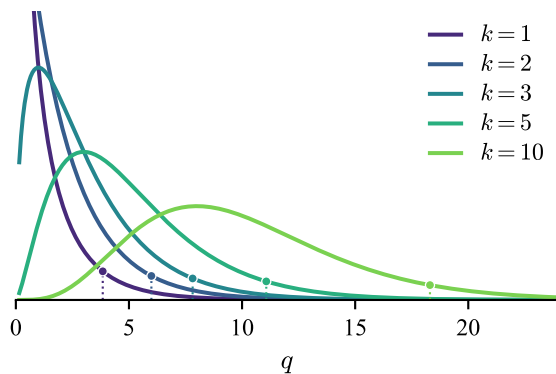
יחד עם משפט הגבול המרכזי (תכונה 2.2.3), האדיטיביות מסבירה גם את ההתנהגות עבור  $k$  גדול: משתנה  $\chi_k^2$  הוא סכום של  $k$  תרומות בלתי תלויות, ולכן

$$(2.35) \quad Q \approx \mathcal{N}(k, 2k), \quad k \gg 1.$$

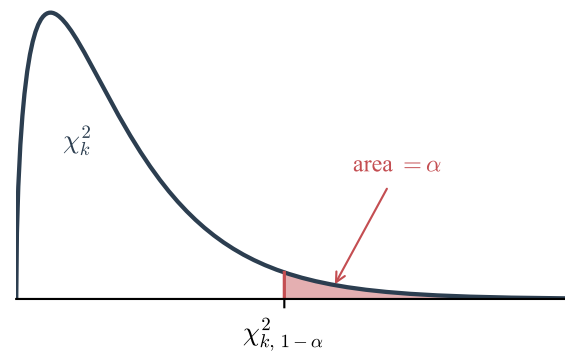
**quantile עליון:** ה-quantile העליון  $\chi_{k,1-\alpha}^2$  הוא הערך שמשתנה  $\chi_k^2$  עולה עליו בהסתברות  $\alpha$  (איור 2.4א),

$$(2.36) \quad \Pr(Q > \chi_{k,1-\alpha}^2) = \alpha, \quad Q \sim \chi_k^2,$$

כאשר הממד התחתון מציין את הרמה המצטברת  $1 - \alpha$ , ולא את הזנב שנותר מעליה. באופן כללי הוא מחושב נומרית, ובשני מקרים קיים לו ביטוי סגור.



(ב) תלות ב- $k$



(א) ה-quantile והזנב שלו

איור 2.4: התפלגות chi-square. (א) ה-quantile העליון  $\chi_{k,1-\alpha}^2$  חותך זנב ששטחו  $\alpha$ . (ב) הצפיפות עבור מספר ערכי  $k$ , כאשר כל אחת מסומנת בנקודה ב- $\chi_{k,0.95}^2$  שלה: הסף נע ימינה והאסימטריה הימנית דועכת ככל ש- $k$  גדל. הצפיפויות עבור  $k=1$  ו- $k=2$  אינן חסומות בראשית, ולכן העלייה שלהן יוצאת מגבולות האיור.

**ביטוי סגור עבור  $k=1$ :** ע"פ משוואה (2.30) משתנה  $\chi_1^2$  הוא  $Z^2$  כאשר  $Z \sim \mathcal{N}(0, 1)$ , ולכן ה-quantile שלו הוא סף נורמלי דו-צדדי בריבוע,

$$(2.37) \quad \chi_{1,1-\alpha}^2 = \left[ \Phi^{-1}(1 - \alpha/2) \right]^2,$$

שנותן  $3^2 = 9$  עבור  $\alpha = 0.0027$ .

**ביטוי סגור עבור  $k=2$ :** עבור  $k=2$  הצפיפות מצטמצמת ל- $f(q; 2) = \frac{1}{2}e^{-q/2}$ , ולכן הזנב וההופכי שלו הם אלמנטריים,

$$(2.38) \quad \Pr(\chi_2^2 > \tau) = e^{-\tau/2}, \quad \chi_{2,1-\alpha}^2 = -2 \ln \alpha.$$

**סכום סטיות ריבועיות:** עבור מדגם גאוסי בלתי תלוי ושווה התפלגות  $X_1, \dots, X_n \sim \mathcal{N}(\mu, \sigma^2)$ , סכום הסטיות הריבועיות סביב הממוצע המדגמי  $\bar{X}$  מקיים

$$(2.39) \quad \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 \sim \chi_{n-1}^2.$$

הוכחה. נוסף נחסיר  $\bar{X}$  בתוך כל ריבוע בסכום  $\sum_i (X_i - \mu)^2$ . האיבר המעורב מתאפס, מאחר והסטיות סביב הממוצע המדגמי מקיימות  $\sum_i (X_i - \bar{X}) = 0$ , ונותרת הזהות

$$\sum_{i=1}^n (X_i - \mu)^2 = \sum_{i=1}^n (X_i - \bar{X})^2 + n(\bar{X} - \mu)^2.$$

חלוקה של כל האיברים ב- $\sigma^2$  מעבירה את שלושתם לצורה מנורמלת,

$$\sum_{i=1}^n \left( \frac{X_i - \mu}{\sigma} \right)^2 = \frac{1}{\sigma^2} \sum_{i=1}^n (X_i - \bar{X})^2 + \left( \frac{\bar{X} - \mu}{\sigma/\sqrt{n}} \right)^2.$$

שניים מתוך השלושה כבר ידועים. אגף שמאל הוא סכום של  $n$  ריבועים של משתנים נורמליים סטנדרטיים, ולכן ע"פ משוואה (2.30) הוא  $\chi_n^2$ . האיבר הימני ביותר הוא ריבוע של משתנה נורמלי סטנדרטי בודד, מאחר ומתקיים  $\bar{X} \sim \mathcal{N}(\mu, \sigma^2/n)$ , ולכן הוא  $\chi_1^2$ . שני האיברים באגף ימין בלתי תלויים, מאחר ועבור מדגם גאוסי הממוצע המדגמי  $\bar{X}$  והסטיות  $X_i - \bar{X}$  בלתי תלויים.

כעת נקרא את משוואה (2.34) מהסוף להתחלה. משתנה  $\chi_n^2$  נרשם כסכום של שני איברי chi-square בלתי תלויים, שאחד מהם הוא  $\chi_1^2$ . דרגות החופש מתחברות, ולכן האיבר הנותר חייב לשאת  $n-1$  מהן, וזוהי משוואה (2.39).

## 2.2.5. התפלגות Student t

**מטרה:** לנרמל ממוצע מדגמי כאשר השונות אינה ידועה, אלא מוערכת מאותם נתונים עצמם.

**התפלגות Student t:** יהיו  $Z \sim \mathcal{N}(0, 1)$  ו- $Q \sim \chi_k^2$  בלתי תלויים. היחס

$$(2.40) \quad \frac{Z}{\sqrt{Q/k}}$$

מתפלג Student t עם  $k$  דרגות חופש, ומסומן  $t_k$ .

**מאפיינים:** הצפיפות סימטרית סביב אפס, כמו  $\mathcal{N}(0, 1)$ , אך בעלת זנבות כבדים יותר (איור 2.5). המשקל הנוסף מגיע מהמכנה של משוואה (2.40), שהוא עצמו אקראי: ערך  $Q$  קטן במיוחד מנפח את היחס גם כאשר  $Z$  רגיל לחלוטין. ככל ש- $k$  גדל המכנה מתרכז סביב 1, ומתקיים

$$(2.41) \quad t_k \longrightarrow \mathcal{N}(0, 1), \quad k \rightarrow \infty.$$

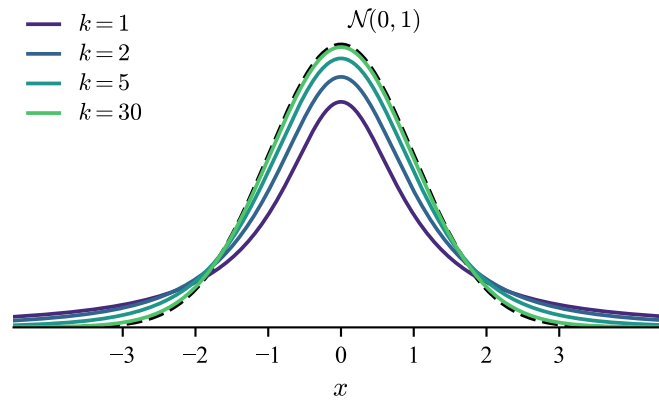
**quantile עליון:** כמו עבור chi-square, הערך  $t_{k, 1-\alpha/2}$  הוא הערך שמשתנה  $t_k$  עולה עליו בהסתברות  $\alpha/2$ ,

$$(2.42) \quad \Pr(t_k > t_{k, 1-\alpha/2}) = \frac{\alpha}{2}.$$

מאחר והצפיפות סימטרית, אותו ערך משמש את שני הזנבות: מקטע שההסתברות הכוללת שלו היא  $1 - \alpha$  משתרע מ- $-t_{k, 1-\alpha/2}$  ועד  $+t_{k, 1-\alpha/2}$ .

**ממוצע מדגמי מנורמל:** יהיו  $x_1, \dots, x_n$  בלתי תלויים ושווי התפלגות  $\mathcal{N}(\mu, \sigma^2)$ , בעלי ממוצע מדגמי  $\bar{x}$  וסטיית תקן מדגמית חסרת הטיה  $s_x$ . אז

$$(2.43) \quad \frac{\bar{x} - \mu}{s_x/\sqrt{n}} \sim t_{n-1}.$$



איור 2.5: התפלגות Student t עבור מספר ערכי  $k$ , מול  $\mathcal{N}(0, 1)$  (מקווקו). ככל ש- $k$  קטן יותר, השיא נמוך יותר והזנבות כבדים יותר; כבר עבור  $k = 30$  קשה להבחין בין השתיים.

הוכחה. שני המרכיבים של משוואה (2.40) כבר נמצאים בידינו.

המונה. הממוצע המדגמי של  $n$  תצפיות גאוסיות הוא עצמו גאוסי,  $\bar{x} \sim \mathcal{N}(\mu, \sigma^2/n)$ , ולכן נרמול שלו ע"י סטיית התקן שלו עצמו נותן

$$Z = \frac{\bar{x} - \mu}{\sigma/\sqrt{n}} \sim \mathcal{N}(0, 1).$$

המכנה. ע"פ משוואה (2.39),

$$Q = \frac{1}{\sigma^2} \sum_{i=1}^n (x_i - \bar{x})^2 \sim \chi_{n-1}^2.$$

השונות חסרת ההטיה היא  $s_x^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$ , ולכן אותו  $Q$  עצמו שווה ל- $(n-1)s_x^2/\sigma^2$ , כלומר

$$s_x = \sigma \sqrt{\frac{Q}{n-1}}.$$

אי תלות. עבור מדגם גאוסי הממוצע המדגמי  $\bar{x}$  והסטיות  $x_i - \bar{x}$  בלתי תלויים, ולכן גם  $Z$  ו- $Q$  בלתי תלויים. זהו בדיוק התנאי שמשוואה (2.40) דורשת משני המרכיבים שלה.

חיבור התוצאות. נחלק את המונה ואת המכנה של אגף שמאל של משוואה (2.43) ב- $\sigma$  ונציב את שתי התוצאות,

$$\frac{\bar{x} - \mu}{s_x/\sqrt{n}} = \frac{(\bar{x} - \mu) / (\sigma/\sqrt{n})}{s_x/\sigma} = \frac{Z}{\sqrt{Q/(n-1)}},$$

וזהי משוואה (2.40) עבור  $k = n - 1$ . הערך  $\sigma$  הלא ידוע הצטמצם, וזו כל המטרה: אגף שמאל ניתן לחישוב מתוך המדגם בלבד.

דרגת החופש שאבדה כאן היא אותה דרגת חופש שהופכת את  $s_x$  לחסר הטיה.

הערכת  $\sigma$  במקום ידיעתו עולה אפוא בדיוק, והמחיר נראה ב-quantile. עבור  $\alpha = 0.05$  של  $t_{k, 0.975}$  הוא 2.571 עבור  $k = 5$ , 2.228 עבור  $k = 10$ , 2.042 עבור  $k = 30$  ו-1.984 עבור  $k = 100$ , מול  $\Phi^{-1}(0.975) = 1.960$  כאשר  $\sigma$  ידוע. מעבר ל- $k = 30$  בקירוב הפער קטן מ-5%, ומכאן מגיע כלל האצבע המקובל " $n \gtrsim 30$ ".

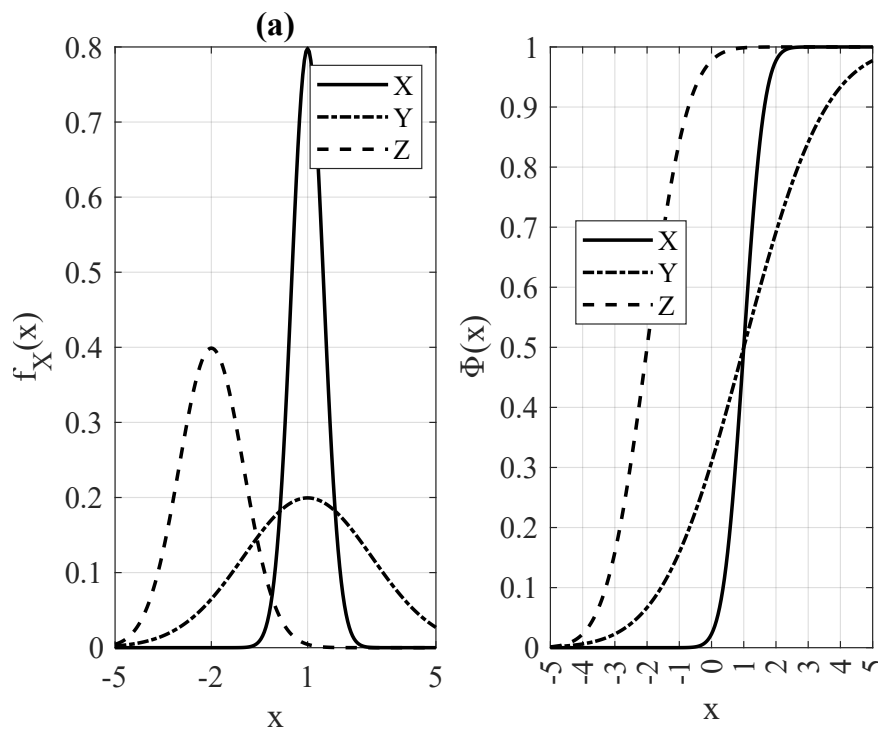
## 2.3. שאלות

2.1 נתונים גרפים של CDF של משתנים אקראיים גאומיים.

$$X \sim \mathcal{N}(\mu_X, \sigma_X^2)$$

$$Y \sim \mathcal{N}(\mu_Y, \sigma_Y^2)$$

$$Z \sim \mathcal{N}(\mu_Z, \sigma_Z^2)$$



(א) ציין מהי התוחלת של כל אחד מהמשתנים,  $\mu_X, \mu_Y, \mu_Z$ . יש לנמק את תשובתך!

(ב) סדר את הגרפים לפי ערך סטיית תקן שלהם  $\sigma_i > \sigma_j > \sigma_k, i, j, k \in \{X, Y, Z\}$  בסדר יורד. יש לנמק את תשובתך!

2.2 (תאוריה ו/או סימולציה) נתון רצף ניסויים  $X_1, X_2, \dots, X_{200}$ , כאשר הניסויים הם בלתי תלויים  $X_i \sim \mathcal{N}(1/2, 1/2^2)$  נגדיר

$$Y = X_1 + X_2 + \dots + X_{200}$$

יש לחשב  $\Pr(90 \leq Y \leq 110)$

2.3 (כולל סימולציה) נתונים משתנים אקראיים,

$$(2.44) \quad X \sim \text{Exp}(\lambda)$$

$$(2.45) \quad Y = X^2 + 1$$

(א) חשבו  $\mathbb{E}[Y], \text{Var}[Y]$

(ב) עבור  $\lambda = 1/2$ , שרטטו  $\mathbb{E}[Y], \text{Var}[Y]$  מחושבים ע"י סימולציה נומרית, כפונקציה של מספר ניסויים אקראיים  $N \leq 10^6$ . השוו ערכים המתקבלים לתוצאות אנליטיות שהתקבלו בסעיף הקודם.

(ג) שרטטו PDF המתקבל של  $Y$  ע"י היסטוגרמה.

2.4 (סימולציה) הגרל רצף מספרים בעלי התפלגות Bernoulli,  $X \sim \text{Ber}(p)$ . ספור מספר הגרלות ברצף עד להופעת ה-"1" הראשון. שרטט התפלגות (PDF) של מספר ההגרלות המתקבל. השווה את התוצאה ל-PDF של התפלגות גאומטרית.

2.5 (סימולציה) נתון  $X \sim U[0, 1]$ . יש להראות ע"י סימולציה את הקשר

$$\Pr\left(X^{\frac{1}{4}} < \frac{1}{2}\right) = \frac{1}{16} = 0.0625$$

## 3. זוג משתנים אקראיים

**מטרה:** לאפיין ניסוי אקראי בעל שתי תוצאות (בדידות או רציפות), והקשר בין התוצאות האלה.

### תוכן הפרק

|    |                               |
|----|-------------------------------|
| 27 | 3.1. משתנים אקראיים בדידים    |
| 27 | 3.1.1. אפיון משותף            |
| 28 | 3.1.2. התפלגות שולית          |
| 30 | 3.1.3. אי תלות סטטיסטית       |
| 31 | 3.2. משתנים אקראיים רציפים    |
| 31 | 3.2.1. אפיון משותף            |
| 33 | 3.2.2. תוחלת                  |
| 34 | 3.2.3. התפלגות אחידה דו-ממדית |
| 35 | 3.3. שונות משותפת             |
| 35 | 3.3.1. הגדרה                  |
| 36 | 3.3.2. מקדם קורלציה           |
| 38 | 3.4. התפלגות גאוסית דו-ממדית  |
| 40 | 3.5. שאלות                    |
| 43 | נספחים                        |
| 43 | 3.א'. בניית זוג גאוסי דו-ממדי |

### 3.1. משתנים אקראיים בדידים

**מטרה:** להרחיב את ה-PDF, ה-CDF והתוחלת ממשתנה בדיד יחיד לזוג משתנים.

#### 3.1.1. אפיון משותף

**PDF משותף:** ההסתברות שתוצאת הניסוי היא זוג הערכים  $(X = x_k, Y = y_j)$  היא

$$(3.1) \quad p_{XY}[x_k, y_j] = \Pr[X = x_k, Y = y_j],$$

והיא מוגדרת עבור כל זוג אינדקסים אפשרי  $(j, k)$ . בעצם מדובר בחיתוך מאורעות  $\{X = x_k\} \cap \{Y = y_j\}$ .

**תכונות PDF משותף:** תחום ערכים ו"סכום" ערכים,

$$(3.2) \quad 0 \leq p_{XY}[x_k, y_j] \leq 1 \quad \forall j, k$$

$$(ב3.2) \quad \sum_{j,k} p_{XY}[x_k, y_j] = 1$$

כאשר הסכימה היא עבור כל זוג אינדקסים אפשרי  $(j, k)$ .

**דוגמה 3.1:** עבור תוצאה של זריקת 2 קוביות הוגנות, אף זריקה אינה משפיעה על רעותה, ולכן ההסתברות לזוג פאות מסוים היא מכפלת שתי ההסתברויות של קוביה בודדת,

$$(3.3) \quad p_{XY}[x_k, y_j] = \Pr[X = x_k] \cdot \Pr[Y = y_j] = \frac{1}{6} \cdot \frac{1}{6} = \frac{1}{36}$$

**CDF משותף:** ה-CDF של התפלגות משותפת של שני משתנים אקראיים בדידים הוא

$$(3.4) \quad F_{XY}(x, y) = \Pr(X \leq x, Y \leq y)$$

**קשר בין PDF משותף ל-CDF משותף:** ה-CDF המשותף צובר את ה-PDF המשותף, בדיוק כמו במקרה של משתנה יחיד (תכונה 1.2.1), אלא שהצבירה מתבצעת כעת על שני אינדקסים,

$$(3.5) \quad F_{XY}(x, y) = \sum_{\substack{j,k \\ x_k \leq x \cap y_j \leq y}} p_{XY}[x_k, y_j]$$

**תוחלת של פונקציה של שני משתנים אקראיים:** עבור כל פונקציה ממשית  $g(x, y)$ , הצורה הדו-ממדית של משוואה (1.11) היא

$$(3.6) \quad \mathbb{E}[g(X, Y)] = \sum_j \sum_k g(x_k, y_j) p_{XY}[x_k, y_j]$$

**דוגמה 3.2:** המקרה הנפוץ ביותר בפרק זה הוא המומנט המעורב,  $g(x, y) = xy$ ,

$$(3.7) \quad \mathbb{E}[XY] = \sum_j \sum_k x_k y_j p_{XY}[x_k, y_j]$$

### 3.1.2. התפלגות שולית

**מטרה:** לקבל את ההתפלגות של כל אחד מהמשתנים בנפרד, מתוך ההתפלגות המשותפת,

$$\begin{aligned} p_{XY}[x_k, y_j] &\longrightarrow p_X[x_k], \quad p_Y[y_j], \\ F_{XY}(x, y) &\longrightarrow F_X(x), \quad F_Y(y). \end{aligned}$$

**התפלגות שולית:** ההתפלגות השולית של משתנה אחד מתקבלת ע"י סכימה של ההתפלגות המשותפת על כל ערכי המשתנה השני,

$$(א3.8) \quad p_X[x_k] = \sum_j p_{XY}[x_k, y_j]$$

$$(ב3.8) \quad p_Y[y_j] = \sum_k p_{XY}[x_k, y_j]$$

$$(ג3.8) \quad F_X(x) = F_{XY}(x, \infty)$$

$$(ד3.8) \quad F_Y(y) = F_{XY}(\infty, y)$$

**דוגמה 3.3:** לניסוי 5 זוגות של תוצאות אפשריות, בעלי הסתברות זהה,

| $x_k$ | $y_j$ | $p_{XY}[x_k, y_j]$ |
|-------|-------|--------------------|
| 1     | 1     | 1/5                |
| 1     | 2     | 1/5                |
| 1     | 3     | 1/5                |
| 2     | 1     | 1/5                |
| 2     | 2     | 1/5                |

לכל זוג שאינו מופיע בטבלה, כגון  $(2, 3)$ , ההסתברות היא אפס.

(א) יש למצוא את ההסתברויות השוליות  $p_X[1]$  ו- $p_Y[3]$ .

(ב) יש למצוא  $F_{XY}(1.5, 2.5)$ .

(ג) יש לחשב  $\mathbb{E}[XY]$  ו- $\mathbb{E}[X]$ ,  $\mathbb{E}[X^2]$ ,  $\text{Var}[X]$ ,  $\mathbb{E}[Y]$ ,  $\mathbb{E}[Y^2]$ ,  $\text{Var}[Y]$ .

פתרון: בהתאם לנוסחה של הסתברות שולית, סכימה של שורה או עמודה בטבלה נותנת את ההתפלגויות השוליות,

$$\begin{aligned} p_X[1] &= \sum_j p_{XY}[1, y_j] \\ &= p_{XY}[1, 1] + p_{XY}[1, 2] + p_{XY}[1, 3] = \frac{3}{5} \\ p_Y[3] &= \sum_k p_{XY}[x_k, 3] \\ &= p_{XY}[1, 3] + p_{XY}[2, 3] = \frac{1}{5} + 0 = \frac{1}{5} \end{aligned}$$

שאר הערכים השוליים מתקבלים באותו אופן,  $p_X[2] = 2/5$  וכן  $p_Y[1] = p_Y[2] = 2/5$ .

ה-CDF המשותף בנקודה  $(1.5, 2.5)$  צובר את שני הזוגות המופיעים בטבלה המקיימים גם  $x_k \leq 1.5$  וגם  $y_j \leq 2.5$ , כלומר  $(1, 1)$  ו- $(1, 2)$ ,

$$F_{XY}(1.5, 2.5) = p_{XY}[1, 1] + p_{XY}[1, 2] = \frac{2}{5}$$

המומנטים של  $X$  ושל  $Y$  מחושבים מתוך ההתפלגויות השוליות שלהם, והמומנט המעורב מתוך ה-PDF המשותף,

$$\begin{aligned} \mathbb{E}[X] &= \sum_k x_k p_X[x_k] \\ &= 1 \cdot \frac{3}{5} + 2 \cdot \frac{2}{5} = \frac{7}{5} \\ \mathbb{E}[X^2] &= \sum_k x_k^2 p_X[x_k] \\ &= 1 \cdot \frac{3}{5} + 4 \cdot \frac{2}{5} = \frac{11}{5} \\ \text{Var}[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{11}{5} - \left(\frac{7}{5}\right)^2 = \frac{6}{25} \\ \mathbb{E}[Y] &= 1 \cdot \frac{2}{5} + 2 \cdot \frac{2}{5} + 3 \cdot \frac{1}{5} = \frac{9}{5} \\ \mathbb{E}[Y^2] &= 1 \cdot \frac{2}{5} + 4 \cdot \frac{2}{5} + 9 \cdot \frac{1}{5} = \frac{19}{5} \end{aligned}$$

$$\begin{aligned}\text{Var}[Y] &= \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = \frac{19}{5} - \left(\frac{9}{5}\right)^2 = \frac{14}{25} \\ \mathbb{E}[XY] &= \sum_j \sum_k x_k y_j p_{XY}[x_k, y_j] \\ &= \frac{1}{5} (1 \cdot 1 + 1 \cdot 2 + 1 \cdot 3 + 2 \cdot 1 + 2 \cdot 2) \\ &= \frac{12}{5}\end{aligned}$$

### 3.1.3. אי תלות סטטיסטית

**מטרה:** לאפיין את התנאי שבו שני המשתנים אינם נושאים מידע זה על זה.

**משתנים בלתי תלויים סטטיסטית:** משתנים  $X, Y$  נקראים בלתי תלויים סטטיסטית אם ורק אם ההתפלגות המשותפת מתפרקת למכפלה של ההתפלגויות השוליות,

$$\begin{aligned}(א3.9) \quad p_{XY}[x_k, y_j] &= p_X[x_k] p_Y[y_j] \\ (ב3.9) \quad F_{XY}(x, y) &= F_X(x) F_Y(y)\end{aligned}$$

ההתפלגויות השוליות התקבלו מתוך ההתפלגות המשותפת. השאלה הטבעית היא, האם ניתן לבצע את הגזירה בכיוון ההפוך, כלומר האם

$$p_{XY}[x_k, y_j] \stackrel{?}{\leftarrow} p_X[x_k], p_Y[y_j].$$

#### התפלגות משותפת והתפלגויות שוליות

שני הכיוונים אינם סימטריים:

- ההתפלגויות השוליות נובעות תמיד מההתפלגות המשותפת, ללא הנחה נוספת.
- ההתפלגות המשותפת נובעת מההתפלגויות השוליות רק כאשר המשתנים בלתי תלויים סטטיסטית. אחרת ההתפלגויות השוליות פשוט אינן מספיקות: הן מתארות כל משתנה בנפרד, ואינן אומרות דבר על האופן שבו השניים נעים יחד.

**דוגמה 3.4:** זוג המשתנים בדוגמה 3.3 הוא תלוי. לדוגמה, מתקיים  $p_{XY}[2, 3] = 0$ , ואילו מכפלת ההתפלגויות השוליות היא  $p_X[2] p_Y[3] = \frac{2}{5} \cdot \frac{1}{5} = \frac{2}{25} \neq 0$ .

**דוגמה 3.5:** נתונים זוג ניסויי Bernoulli בלתי תלויים,  $X, Y \sim \text{Ber}(p)$ . יש למצוא את ה-PDF המשותף  $p_{XY}[x_k, y_j]$  ואת ה-CDF המשותף  $F_{XY}(x, y)$ .

פתרון: מאחר והניסויים בלתי תלויים, ה-PDF המשותף הוא מכפלת ההתפלגויות השוליות (טבלה 3.1),

$$\begin{aligned}(א3.10) \quad p_{XY}[0, 0] &= p_X[0] p_Y[0] = (1 - p)^2, \\ (ב3.10) \quad p_{XY}[0, 1] &= p_X[0] p_Y[1] = (1 - p)p, \\ (ג3.10) \quad p_{XY}[1, 0] &= p_X[1] p_Y[0] = p(1 - p), \\ (ד3.10) \quad p_{XY}[1, 1] &= p_X[1] p_Y[1] = p^2.\end{aligned}$$

סכימה לפי שורה או לפי עמודה בטבלה מחזירה את ההתפלגויות השוליות, וכל אחת מהן היא Bernoulli, כמצופה,

$$\begin{aligned}(א3.11) \quad p_X[0] &= p_{XY}[0, 0] + p_{XY}[0, 1] \\ &= (1 - p)^2 + (1 - p)p = 1 - p,\end{aligned}$$

$$\begin{aligned}
 p_X[1] &= p_{XY}[1, 0] + p_{XY}[1, 1] \\
 &= (1-p)p + p^2 = p.
 \end{aligned}
 \tag{3.11}$$

ה-CDF המשותף צובר את ארבעת הערכים האלה,

$$F_{XY}(x, y) = \begin{cases} 0 & x < 0, y < 0, \\ p_{XY}[0, 0] & 0 \leq x < 1, 0 \leq y < 1, \\ p_{XY}[0, 0] + p_{XY}[0, 1] & 0 \leq x < 1, 1 \leq y, \\ p_{XY}[0, 0] + p_{XY}[1, 0] & 1 \leq x, 0 \leq y < 1, \\ 1 & 1 \leq x, 1 \leq y, \end{cases}
 \tag{3.12}$$

וכמובן, הפונקציה רציפה מימין בשני המשתנים, כפי שנדרש מ-CDF.

|   |   | X         |          |
|---|---|-----------|----------|
|   |   | 0         | 1        |
| Y | 0 | $(1-p)^2$ | $p(1-p)$ |
|   | 1 | $(1-p)p$  | $p^2$    |

טבלה 3.1: ה-PDF המשותף של שני משתני Bernoulli בלתי תלויים, בצורת טבלה.

## 3.2. משתנים אקראיים רציפים

**מטרה:** לחזור על האפיון עבור ניסוי אקראי בעל שתי תוצאות רציפות.

**דוגמה 3.6:** דוגמאות לניסויים אקראיים בעלי שתי תוצאות רציפות:

- גובה ומשקל של הנוכחים בכיתה.
- קריאות "רועשות" של מתח וזרם על אותו נגד.

### 3.2.1. אפיון משותף

**CDF משותף:** ההגדרה של ה-CDF המשותף, משוואה (3.4), זהה עבור משתנים רציפים ובדידים,

$$F_{XY}(x, y) = \Pr(X \leq x, Y \leq y)
 \tag{3.13}$$

**PDF משותף:** ה-PDF המשותף  $f_{XY}(x, y)$  הוא הנגזרת המעורבת של ה-CDF המשותף,

$$f_{XY}(x, y) = \frac{\partial^2 F_{XY}(x, y)}{\partial x \partial y}
 \tag{3.14}$$

**קשר בין PDF משותף ל-CDF משותף:** האינטגרציה מחליפה את הסכימה של המקרה הבדיד,

$$F_{XY}(x, y) = \int_{-\infty}^x \int_{-\infty}^y f_{XY}(s, t) dt ds
 \tag{3.15}$$

**תכונות PDF משותף:** תחום ערכים ו"סכום" ערכים,

$$(א3.16) \quad f_{XY}(x, y) \geq 0$$

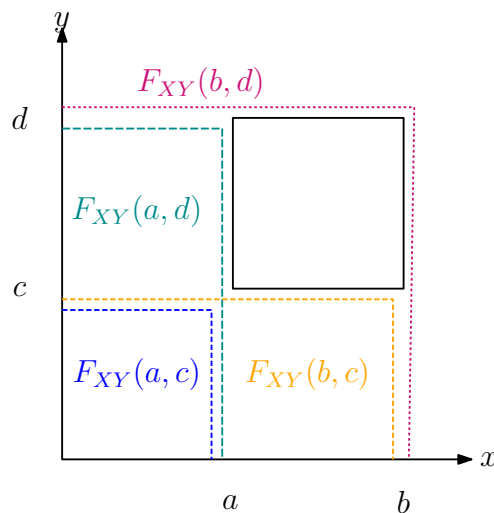
$$(ב3.16) \quad \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} f_{XY}(x, y) dx dy = 1 = F_{XY}(\infty, \infty)$$

**דוגמה 3.7:** נתון מלבן

$$B = (a, b] \times (c, d]$$

המופיע באיור 3.1. בהינתן ה-CDF המשותף  $F_{XY}(x, y)$ , יש למצוא את ההסתברות שהזוג  $(X, Y)$  ייפול בתוך  $B$ . פתרון: כל ערך של ה-CDF המשותף הוא הסתברות של רביע אינסופי. המלבן מורכב מארבעה רביעים כאלה, בחיבור ובחיסור, כך שכל אזור מחוץ ל- $B$  נספר בדיוק כמספר הפעמים שהוא מוחסר (איור 3.1),

$$(3.17) \quad \begin{aligned} \Pr((X, Y) \in B) &= \Pr(a < X \leq b, c < Y \leq d) \\ &= F_{XY}(b, d) - F_{XY}(a, d) \\ &\quad - F_{XY}(b, c) + F_{XY}(a, c) \end{aligned}$$



איור 3.1: איור עבור דוגמה 3.7.

**התפלגות שולית:** ההתפלגויות השוליות מתקבלות ע"י אינטגרציה על המשתנה השני,

$$(א3.18) \quad f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy,$$

$$(ב3.18) \quad f_Y(y) = \int_{-\infty}^{\infty} f_{XY}(x, y) dx,$$

$$(ג3.18) \quad F_X(x) = F_{XY}(x, \infty),$$

$$(ד3.18) \quad F_Y(y) = F_{XY}(\infty, y)$$

**אי תלות סטטיסטית:** שני משתנים אקראיים רציפים  $X, Y$  הם בלתי תלויים סטטיסטית אם ורק אם מתקיים

$$(א3.19) \quad f_{XY}(x, y) = f_X(x) f_Y(y),$$

$$(ב3.19) \quad F_{XY}(x, y) = F_X(x) F_Y(y)$$

במקרה זה בלבד ניתן לקבל את ההתפלגות המשותפת מתוך ההתפלגויות השוליות.

### 3.2.2. תוחלת

**תוחלת של פונקציה של שני משתנים אקראיים:** המקבילה הרציפה של משוואה (3.6) היא

$$(3.20) \quad \mathbb{E}[g(X, Y)] = \int \int_{-\infty}^{\infty} g(x, y) f_{XY}(x, y) dx dy$$

**דוגמה 3.8:** גם כאן, המקרה הנפוץ ביותר הוא המומנט המעורב,

$$(3.21) \quad \mathbb{E}[XY] = \int \int_{-\infty}^{\infty} x y f_{XY}(x, y) dx dy$$

**לינאריות של התוחלת:** תוחלת של סכום היא סכום התוחלות, עבור כל זוג משתנים אקראיים,

$$(3.22) \quad \mathbb{E}[aX + bY] = a\mathbb{E}[X] + b\mathbb{E}[Y]$$

הוכחה. ההוכחה מוצגת עבור המקרה המייצג  $a = b = 1$  ומשתנים רציפים. נפצל את האינטגרל לשני איברים, ונשים לב שהאינטגרל הפנימי של כל אחד מהם הוא בדיוק ההגדרה של התפלגות שולית,

$$\begin{aligned} \mathbb{E}[X + Y] &= \int \int_{-\infty}^{\infty} (x + y) f_{XY}(x, y) dx dy \\ &= \int \int_{-\infty}^{\infty} x f_{XY}(x, y) dx dy \\ &\quad + \int \int_{-\infty}^{\infty} y f_{XY}(x, y) dx dy \\ &= \int_{-\infty}^{\infty} x f_X(x) dx + \int_{-\infty}^{\infty} y f_Y(y) dy \\ &= \mathbb{E}[X] + \mathbb{E}[Y] \end{aligned}$$

אותו נימוק עם סכומים במקום אינטגרלים מוכיח את המקרה הבדיד.

**תוחלת של מכפלה של משתנים בלתי תלויים:** עבור זוג משתנים בלתי תלויים ופונקציה מהצורה  $g(x, y) = g_1(x) g_2(y)$ , התוחלת מתפרקת,

$$(3.23) \quad \mathbb{E}[g_1(X)g_2(Y)] = \mathbb{E}[g_1(X)] \mathbb{E}[g_2(Y)]$$

המקרה הפרטי החשוב, התקף עבור משתנים בלתי תלויים בלבד, הוא

$$(3.24) \quad \mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y]$$

הוכחה. אי התלות, משוואה (3.19), מאפשרת להחליף את הצפיפות המשותפת במכפלה  $f_X(x)f_Y(y)$ , ולאחר מכן האינטגרל הכפול מתפרק,

$$\begin{aligned} \mathbb{E}[XY] &= \int \int_{-\infty}^{\infty} x y f_X(x) f_Y(y) dx dy \\ (3.25) \quad &= \left( \int_{-\infty}^{\infty} x f_X(x) dx \right) \left( \int_{-\infty}^{\infty} y f_Y(y) dy \right) \end{aligned}$$

זוהי משוואה (3.24). המקרה הבדיד מתקבל באותו אופן.

### נדרשת אי תלות

משוואה (3.22) מתקיימת עבור כל זוג משתנים, אך משוואה (3.24) מתקיימת רק עבור משתנים בלתי תלויים. הפער בין  $\mathbb{E}[XY]$  לבין  $\mathbb{E}[X] \mathbb{E}[Y]$  הוא בדיקת הגודל המוגדר בסעיף הבא.

### 3.2.3. התפלגות אחידה דו-ממדית

**התפלגות אחידה דו-ממדית:** ה-PDF המשותף של התפלגות אחידה דו-ממדית על המלבן  $[a, b] \times [c, d]$  נתון ע"י

$$(3.26) \quad f_{XY}(x, y) = \begin{cases} \frac{1}{b-a} \frac{1}{d-c}, & a \leq x \leq b, c \leq y \leq d, \\ 0, & \text{אחרת} \end{cases}$$

**דוגמה 3.9:** עבור התפלגות אחידה דו-ממדית:

(א) יש לחשב  $\mathbb{E}[X]$  ע"פ הנוסחה הכללית לתוחלת, משוואה (3.20).

(ב) יש להוכיח שמדובר בזוג משתנים בלתי תלויים,  $X \sim U[a, b]$  ו- $Y \sim U[c, d]$ .

פתרון: נציב  $g(x, y) = x$  במשוואה (3.20). האינטגרל על  $y$  תורם גורם אחד,

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^{\infty} \int_{-\infty}^{\infty} x f_{XY}(x, y) dx dy \\ &= \int_a^b \frac{x}{b-a} dx \int_c^d \frac{dy}{d-c} = \frac{a+b}{2} \end{aligned}$$

וזו אותה תוצאה כמו עבור משתנה אחיד יחיד.

עבור הסעיף השני, נבצע אינטגרציה של הצפיפות המשותפת על  $y$ ,

$$f_X(x) = \int_{-\infty}^{\infty} f_{XY}(x, y) dy = \begin{cases} \frac{1}{b-a}, & a \leq x \leq b, \\ 0, & \text{אחרת} \end{cases}$$

וחישוב סימטרי נותן את  $f_Y(y)$ . מאחר ומתקיים  $f_{XY}(x, y) = f_X(x)f_Y(y)$ , התנאי (3.19) מתקיים ושני המשתנים בלתי תלויים. כתוצאה מכך, ע"פ משוואה (3.24),

$$\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y] = \frac{a+b}{2} \cdot \frac{c+d}{2}$$

**דוגמה 3.10:** עבור המקרה הפרטי  $a = c = 0, b = d = 1$ , יש למצוא את ה-CDF המשותף.

פתרון: הצפיפות המשותפת שווה לאחד על ריבוע היחידה, ולכן ה-CDF הוא השטח של אותו חלק של הריבוע הנמצא מתחת ל- $(x, y)$  ומשמאלו,

$$(3.27) \quad F_{XY}(x, y) = \begin{cases} 0 & x < 0 \text{ או } y < 0, \\ xy & 0 \leq x \leq 1, 0 \leq y \leq 1, \\ x & 0 \leq x \leq 1, 1 < y, \\ y & 1 < x, 0 \leq y \leq 1, \\ 1 & 1 < x, 1 < y \end{cases}$$

### 3.3. שונות משותפת

**מטרה:** לאפיין את הקשר הלינארי בין שני משתנים אקראיים ע"י מספר בודד.

#### 3.3.1. הגדרה

**דוגמה 3.11:** יש לפתח את  $\text{Var}[X + Y]$  באמצעות השונויות של  $X$  ושל  $Y$ . פתרון: נתחיל מההגדרה של השונות, ונמרכז כל משתנה בנפרד,

$$\begin{aligned}
 \text{Var}[X + Y] &= \mathbb{E} \left[ \left( X + Y - \mathbb{E}[X + Y] \right)^2 \right] \\
 &= \mathbb{E} \left[ \left( (X - \mathbb{E}[X]) + (Y - \mathbb{E}[Y]) \right)^2 \right] \\
 (3.28) \quad &= \mathbb{E} \left[ \left( X - \mathbb{E}[X] \right)^2 \right] + \mathbb{E} \left[ \left( Y - \mathbb{E}[Y] \right)^2 \right] \\
 &\quad + 2 \mathbb{E} \left[ \left( X - \mathbb{E}[X] \right) \left( Y - \mathbb{E}[Y] \right) \right]
 \end{aligned}$$

שני האיברים הראשונים הם  $\text{Var}[X]$  ו- $\text{Var}[Y]$ . האיבר השלישי אינו שייך לאף אחד מהמשתנים בנפרד: זוהי השונות המשותפת, המוגדרת להלן.

**שונות משותפת (covariance):** השונות המשותפת של שני משתנים אקראיים  $X$  ו- $Y$  היא

$$(3.29) \quad \text{Cov}[X, Y] = \mathbb{E} \left[ \left( X - \mathbb{E}[X] \right) \left( Y - \mathbb{E}[Y] \right) \right]$$

**שונות משותפת, הגדרה חלופית:** הצורה השימושית בפועל, מאחר והיא דורשת רק את שלושת המומנטים  $\mathbb{E}[XY]$ ,  $\mathbb{E}[X]$  ו- $\mathbb{E}[Y]$  היא

$$(3.30) \quad \text{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y]$$

הוכחה. נפתח את המכפלה במשוואה (3.29) ונשתמש בלינאריות של התוחלת יחד עם העובדה ש- $\mathbb{E}[X]$  ו- $\mathbb{E}[Y]$  הם קבועים,

$$\begin{aligned}
 \text{Cov}[X, Y] &= \mathbb{E} \left[ \left( X - \mathbb{E}[X] \right) \left( Y - \mathbb{E}[Y] \right) \right] \\
 (3.31) \quad &= \mathbb{E}[XY] - \mathbb{E}[X\mathbb{E}[Y]] - \mathbb{E}[Y\mathbb{E}[X]] + \mathbb{E}[X]\mathbb{E}[Y] \\
 &= \mathbb{E}[XY] - 2\mathbb{E}[X]\mathbb{E}[Y] + \mathbb{E}[X]\mathbb{E}[Y] \\
 &= \mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]
 \end{aligned}$$

#### שונות משותפת ושונות

בהשוואה בין משוואה (3.29) לבין ההגדרה של השונות בסעיף 1.2.3, השונות המשותפת היא השונות כאשר אחד משני הגורמים הממורכזים הוחלף במשתנה השני. הצבה של  $Y = X$  מחזירה את השונות עצמה.

**תכונות של שונות משותפת:** השונות המשותפת סימטרית, אינה מושפעת מהזזה של אף אחד מהמשתנים, ומשתנה בהתאם לשינוי קנה המידה של שניהם,

$$(3.32) \quad \text{Cov}[X, X] = \text{Var}[X]$$

$$(ב3.32) \quad \text{Cov}[X, Y] = \text{Cov}[Y, X]$$

$$(ג3.32) \quad \text{Cov}[X, a] = 0$$

$$(ד3.32) \quad \text{Cov}[aX, bY] = ab \text{Cov}[X, Y]$$

$$(ה3.32) \quad \text{Cov}[X + a, Y + b] = \text{Cov}[X, Y]$$

$$(ו3.32) \quad \text{Var}[X \pm Y] = \text{Var}[X] + \text{Var}[Y] \pm 2 \text{Cov}[X, Y]$$

### השונות אינה לינארית

משוואה (3.32ו) היא הסיבה לכך שהשונות, בניגוד לתוחלת, אינה עוברת כפשוטה דרך סכום. האיבר המעורב נעלם רק כאשר השונות המשותפת מתאפסת.

**שונות משותפת של משתנים בלתי תלויים:** עבור שני משתנים אקראיים בלתי תלויים סטטיסטית מתקיים תמיד

$$(3.33) \quad \text{Cov}[X, Y] = 0$$

הוכחה. אי התלות נותנת  $\mathbb{E}[XY] = \mathbb{E}[X] \mathbb{E}[Y]$  ע"פ משוואה (3.24), ולכן אגף ימין של משוואה (3.31) מתאפס.

**שונות של סכום משתנים בלתי תלויים:** עבור משתנים בלתי תלויים בלבד, האיבר המעורב במשוואה (3.32) מתאפס והשונות הופכת לאדיטיבית,

$$(3.34) \quad \text{Var}[aX + bY] = a^2 \text{Var}[X] + b^2 \text{Var}[Y]$$

**אי שוויון Cauchy-Schwarz:** המומנט המעורב והשונות המשותפת חסומים שניהם ע"י המומנטים השניים המתאימים,

$$(א3.35) \quad \left( \mathbb{E}[XY] \right)^2 \leq \mathbb{E}[X^2] \mathbb{E}[Y^2],$$

$$(ב3.35) \quad \left( \text{Cov}[X, Y] \right)^2 \leq \text{Var}[X] \text{Var}[Y]$$

הוכחה. נתחיל מכך שתוחלת של ריבוע לעולם אינה שלילית,

$$\mathbb{E}[(aX + Y)^2] = a^2 \mathbb{E}[X^2] + 2a \mathbb{E}[XY] + \mathbb{E}[Y^2] \geq 0,$$

עבור כל  $a$  ממשי. לפולינום ריבועי אי שלילי ב- $a$  יש דיסקרימיננטה אי חיובית,

$$4 \left( \mathbb{E}[XY] \right)^2 - 4 \mathbb{E}[X^2] \mathbb{E}[Y^2] \leq 0,$$

וזהו אי השוויון הראשון. החלפה של  $X \rightarrow X - \mathbb{E}[X]$  ושל  $Y \rightarrow Y - \mathbb{E}[Y]$  הופכת את המומנטים השניים לשונות ואת המומנט המעורב לשונות משותפת, ומתקבל אי השוויון השני.

## 3.3.2. מקדם קורלציה

**מטרה:** להפוך את השונות המשותפת למספר חסר יחידות, שניתן להשוואה בין ניסויים שונים.

**מקדם קורלציה:** מקדם הקורלציה, הידוע גם בשם מקדם מתאם, או בשמו המלא Pearson product-moment correlation coefficient, הוא השונות המשותפת מנורמלת ע"י שתי סטיות התקן,

$$(3.36) \quad \rho_{XY} = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X] \text{Var}[Y]}}$$

**תחום ערכים של מקדם הקורלציה:** מקדם הקורלציה חסום,

$$(3.37) \quad |\rho_{XY}| \leq 1$$

וזה נובע ישירות מחלוקת אי שוויון Cauchy-Schwarz (3.35) ב- $\text{Var}[X] \text{Var}[Y]$ .

**תלות לינארית מלאה:** החסם מושג,  $\rho_{XY} = \pm 1$ , אם ורק אם  $Y$  הוא פונקציה לינארית מדויקת של  $X$ . הסימן של  $\rho_{XY}$  הוא הסימן של השיפוע.

**משתנים חסרי קורלציה:** שני משתנים אקראיים נקראים חסרי קורלציה, כלומר אין ביניהם קשר לינארי, כאשר

$$(3.38) \quad \text{Cov}[X, Y] = \rho_{XY} = 0$$

**משתנים אורתוגונליים:** שני משתנים אקראיים נקראים אורתוגונליים כאשר המומנט המעורב שלהם מתאפס,

$$(3.39) \quad \mathbb{E}[XY] = 0$$

עבור משתנים בעלי תוחלת אפס, אורתוגונליות וחוסר קורלציה מתלכדים.

#### חוסר קורלציה חלש מאי תלות

אי תלות גוררת חוסר קורלציה, ע"פ תכונה 3.3.1. הכיוון ההפוך אינו מתקיים באופן כללי:  $\rho_{XY} = 0$  שולל קשר לינארי בלבד, וקשר לא לינארי חזק יכול להשאיר את השונות המשותפת אפס בדיוק. החרג החשוב היחיד, משתנים גאוסיים במשותף, מטופל בסעיף 3.4.

**דוגמה 3.12:** יש לחשב את מקדם הקורלציה  $\rho_{XY}$  עבור הזוג הבדיד של דוגמה 3.3.

פתרון: המומנטים כבר חושבו שם, ולכן נותרה רק השונות המשותפת,

$$\text{Cov}[X, Y] = \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y] = \frac{12}{5} - \frac{7}{5} \cdot \frac{9}{5} = -\frac{3}{25}$$

ולכן

$$\rho_{XY} = \frac{-3/25}{\sqrt{\frac{6}{25} \cdot \frac{14}{25}}} = -\frac{3}{\sqrt{84}} = -\frac{\sqrt{21}}{14} \approx -0.327$$

שני המשתנים קשורים באופן חלש ושלילי.

**דוגמה 3.13:**

**מטרה:** ניתוח משתנים בעלי קשר לינארי מדויק.

יש לחשב  $\rho_{XY}$  ו- $\text{Cov}[X, Y]$ ,  $\text{Var}[X + Y]$ ,  $\text{Var}[Y]$ ,  $\text{Var}[X]$  עבור

$$X \sim \mathcal{U}[0, 1], \quad Y = -X$$

פתרון:

$$\begin{aligned}
\text{Var}[X] &= \frac{(b-a)^2}{12} = \frac{1}{12} \\
\text{Var}[Y] &= \text{Var}[(-1)X] = (-1)^2 \text{Var}[X] = \text{Var}[X] \\
\text{Var}[X+Y] &= \text{Var}[X-X] = 0 \\
\text{Cov}[X, Y] &= \text{Cov}[X, -X] \\
&= \mathbb{E}[X(-X)] - \mathbb{E}[X] \mathbb{E}[-X] \\
&= -\left(\mathbb{E}[X^2] - (\mathbb{E}[X])^2\right) \\
&= -\text{Var}[X] \\
\rho_{XY} &= \frac{-\text{Var}[X]}{\sqrt{\text{Var}[X] \text{Var}[X]}} = -1
\end{aligned}$$

מקדם הקורלציה מגיע לחסם התחתון שלו, כפי שנדרש עבור קשר לינארי מדויק בעל שיפוע שלילי.

### דוגמה 3.14:

**מטרה:** להציג זוג משתנים תלויים זה בזה באופן חזק, ובכל זאת חסרי קורלציה.

יש לחשב  $\text{Cov}[X, Y]$  ו- $\rho_{XY}$  עבור

$$X \sim U[-1, 1], \quad Y = X^2$$

פתרון: הצפיפות של  $X$  סימטרית סביב אפס, ולכן כל מומנט אי זוגי מתאפס,

$$\begin{aligned}
\mathbb{E}[X] &= \frac{1 + (-1)}{2} = 0 \\
\mathbb{E}[XY] &= \mathbb{E}[X^3] = \int_{-1}^1 x^3 \frac{dx}{2} = 0 \\
\text{Cov}[X, Y] &= \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y] = 0 \\
\rho_{XY} &= 0
\end{aligned}$$

ידעת  $X$  קובעת את  $Y$  במדויק, ולכן השניים תלויים זה בזה עד כמה שניתן, ובכל זאת מקדם הקורלציה אינו מדווח דבר. הקשר הוא לא לינארי גרידא, ומשוואה (3.36) מודדת רק את החלק הלינארי.

## 3.4. התפלגות גאוסית דו-ממדית

**מטרה:** לאפיין זוג משתנים גאוסיים במשותף, המקרה היחיד שבו חוסר קורלציה חזק כמו אי תלות.

**סכום של משתנים גאוסיים בלתי תלויים:** סכום משוקלל של משתנים גאוסיים בלתי תלויים הוא שוב גאوسی. עבור  $X \sim \mathcal{N}(\mu_x, \sigma_x^2)$  ו- $Y \sim \mathcal{N}(\mu_y, \sigma_y^2)$  בלתי תלויים,

$$(3.40) \quad aX + bY \sim \mathcal{N}(a\mu_x + b\mu_y, a^2\sigma_x^2 + b^2\sigma_y^2)$$

התוחלת נובעת ממשוואה (3.22) והשונות ממשוואה (3.34). החידוש הוא שצורת ההתפלגות נשמרת.

**משתנים גאוסיים במשותף:** משתנים אקראיים  $X_1, X_2, \dots, X_N$ , תלויים או בלתי תלויים, הם גאוסיים במשותף אם ורק אם כל קומבינציה לינארית

$$\sum_{m=1}^N a_m X_m, \quad a_m \in \mathbb{R},$$

היא משתנה אקראי גאوسی.

**PDF גאוס דו-ממדי:** זוג משתנים גאוסיים במשותף מתואר במלואו ע"י חמישה מספרים: שתי התוחלות, שתי השונות ומקדם הקורלציה. ה-PDF המשותף שלו הוא

$$(3.41) \quad f_{X_1 X_2}(x_1, x_2) = \frac{1}{2\pi\sigma_1\sigma_2\sqrt{1-\rho^2}} \exp \left( -\frac{1}{2(1-\rho^2)} \times \right. \\ \left. \times \left[ \frac{(x_1 - \mu_1)^2}{\sigma_1^2} + \frac{(x_2 - \mu_2)^2}{\sigma_2^2} - \frac{2\rho(x_1 - \mu_1)(x_2 - \mu_2)}{\sigma_1\sigma_2} \right] \right)$$

כאשר  $\rho = \rho_{X_1 X_2}$  ו- $\sigma_i^2 = \text{Var}[X_i]$ ,  $\mu_i = \mathbb{E}[X_i]$ . הבנייה המייצרת צפיפות זו מופיעה בנספח לפרק זה.

**התפלגויות שוליות:** ההתפלגויות השוליות של זוג גאוסים במשותף הן גאוסיות,  $X_i \sim \mathcal{N}(\mu_i, \sigma_i^2)$ .

#### התפלגות שולית גאוסית אינה התפלגות משותפת גאוסית

ההפך של התכונה לעיל אינו נכון. שני משתנים יכולים להיות גאוסיים כל אחד בנפרד, ועדיין לא להיות גאוסיים במשותף, כפי שמראה דוגמת הנגד להלן. הגדרה 3.4 היא טענה על כל קומבינציה לינארית, ולא על שתי ההתפלגויות השוליות.

**משתנים גאוסיים במשותף חסרי קורלציה:** עבור זוג גאוסים במשותף, ורק עבורו, חוסר קורלציה ואי תלות הם היינו הך,

$$(3.42) \quad \rho_{XY} = 0 \iff X, Y \text{ בלתי תלויים}$$

הצבה של  $\rho = 0$  במשוואה (3.41) אכן מפרקת את הצפיפות המשותפת למכפלה של שתי ההתפלגויות השוליות.

**דוגמה 3.15:** נתונים שני משתנים,  $X \sim \mathcal{N}(0, \sigma^2)$  ו- $Y = 3X$ .

(א) מהו מקדם הקורלציה (בלי לחשב)?

(ב) יש להראות שהמשתנים  $X, Y$  הם גאוסיים במשותף.

(ג) יש לחשב  $\text{Var}[Y]$  ו- $\text{Cov}[X, Y]$ , ולציין מהי ההתפלגות המשותפת.

פתרון: (א) בגלל קשר לינארי מדויק עם שיפוע חיובי,  $\rho_{XY} = 1$ .

(ב) כל קומבינציה לינארית מצטמצמת לכפולה של  $X$ ,

$$a_1 X + a_2 Y = a_1 X + 3a_2 X = (a_1 + 3a_2) X \sim \mathcal{N}\left(0, (a_1 + 3a_2)^2 \sigma^2\right),$$

שהיא גאוסית עבור כל  $a_1, a_2$ . ע"פ הגדרה 3.4 הזוג הוא גאוסים במשותף.

(ג) לשני המשתנים תוחלת אפס, ולכן

$$\begin{aligned} \text{Var}[Y] &= 3^2 \mathbb{E}[X^2] = 9\sigma^2, \\ \text{Cov}[X, Y] &= \mathbb{E}[3X^2] = 3\sigma^2, \end{aligned}$$

ומתקיים  $\rho_{XY} = 3\sigma^2 / \sqrt{\sigma^2 \cdot 9\sigma^2} = 1$ . כפי שנטען בסעיף (א). ההתפלגות המשותפת היא ההתפלגות

הגאוסית הדו-ממדית של משוואה (3.41) עם  $\mu_1 = \mu_2 = 0$ ,  $\sigma_1 = \sigma$ ,  $\sigma_2 = 3\sigma$  ו- $\rho = 1$ .

## דוגמה 3.16:

**מטרה:** דוגמת נגד: התפלגויות שוליות גאוסיות ללא התפלגות משותפת גאוסית.

נתון  $X \sim \mathcal{N}(0, 1)$ , ונתון סימן בלתי תלוי  $W$  המקיים  $\Pr[W = 1] = \Pr[W = -1] = \frac{1}{2}$ . נגדיר  $Y = WX$ . פתרון: בהינתן כל אחד מערכי  $W$ , המשתנה  $Y$  הוא גאוזי סטנדרטי, ולכן ההתפלגות השולית שלו היא  $\mathcal{N}(0, 1)$ . הזוג חסר קורלציה,

$$\begin{aligned}\text{Cov}[X, Y] &= \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y] = \mathbb{E}[X^2 W] \\ &= \mathbb{E}[X^2] \mathbb{E}[W] = \mathbb{E}[X^2] \cdot 0 = 0\end{aligned}$$

אך ברור שהמשתנים תלויים, מאחר ותמיד מתקיים  $|X| = |Y|$ . אילו היה הזוג גאוזי במשותף, תכונה 3.4 הייתה כופה אי תלות, ולכן אינו כזה. נימוק ישיר יותר הוא ש- $\Pr(X + Y = 0) = \frac{1}{2}$ , ואילו כל צפיפות משותפת רציפה מקצה הסתברות אפס לקו.

**דוגמה 3.17:** נתונים  $X, Y$  עם  $\text{Cov}[X, Y] \neq 0$ , ונגדיר  $Z = aX + Y$  כאשר  $a$  הוא קבוע דטרמיניסטי.

(א) האם ניתן לבחור  $a$  כך ש- $X$  ו- $Z$  יהיו חסרי קורלציה?

(ב) יש לחזור על הסעיף הקודם, בהנחה הנוספת ש- $X, Y$  גאוזיים במשותף.

פתרון: נפתח את השונויות המשותפת של  $X$  עם  $Z$ ,

$$\begin{aligned}\mathbb{E}[Z] &= a\mathbb{E}[X] + \mathbb{E}[Y] \\ \mathbb{E}[XZ] &= a\mathbb{E}[X^2] + \mathbb{E}[XY] \\ \text{Cov}[X, Z] &= a\mathbb{E}[X^2] + \mathbb{E}[XY] - a(\mathbb{E}[X])^2 - \mathbb{E}[X] \mathbb{E}[Y] \\ &= a\text{Var}[X] + \text{Cov}[X, Y]\end{aligned}$$

ולכן הבחירה

$$a = -\frac{\text{Cov}[X, Y]}{\text{Var}[X]}$$

הופכת אותם לחסרי קורלציה. אם בנוסף  $X, Y$  גאוזיים במשותף, אז  $Z$  הוא קומבינציה לינארית שלהם והזוג  $(X, Z)$  הוא גאוזי במשותף אף הוא. במקרה זה תכונה 3.4 משדרגת את המסקנה:  $X$  ו- $Z$  אינם רק חסרי קורלציה, אלא בלתי תלויים לחלוטין.

## 3.5. שאלות

3.1 (שונויות) נתונים הקשרים הבאים בין המשתנים  $X, Y$ ,

$$\begin{aligned}\text{Var}[X + 2Y] &= 40 \\ \text{Var}[X - 2Y] &= 20 \\ \text{Var}[X] &= \text{Var}[Y]\end{aligned}$$

יש לחשב מספרית את  $\text{Var}[X]$  ואת  $\text{Cov}[X, Y]$ .

3.2 (מקדם קורלציה של משתנים בדידים) נתונה התפלגות משותפת של משתנים אקראיים בדידים  $X, Y$ ,

$$p_{XY}[x, y] = \begin{cases} \Pr[X = 0, Y = -1] = \frac{1}{3}, \\ \Pr[X = -1, Y = 0] = \frac{1}{3}, \\ \Pr[X = 1, Y = 1] = \frac{1}{6}, \\ \Pr[X = -1, Y = -1] = \frac{1}{6} \end{cases}$$

יש לחשב את מקדם הקורלציה  $\rho_{XY}$ .

3.3 (מכפלה וסכום של משתנים אחידים) נתונים  $n$  משתנים אקראיים בלתי תלויים  $X_i \sim U[0, 1]$ ,  $i = 1, \dots, n$ , נגדיר

$$Y = X_1 \cdot X_2 \cdot \dots \cdot X_n, \\ Z = X_1 + X_2 + \dots + X_n$$

יש לחשב  $\mathbb{E}[Y]$ ,  $\text{Var}[Y]$ ,  $\mathbb{E}[Z]$ ,  $\text{Var}[Z]$ .

3.4 (מקדם קורלציה) נתונה פונקציית צפיפות משותפת,

$$f_{XY}(x, y) = \exp\left(-\frac{4x+y}{2}\right) u(x) u(y),$$

כאשר  $u(\cdot)$  היא פונקציית מדרגה. יש למצוא את מקדם הקורלציה  $\rho_{XY}$ .

3.5 (משתנים בעלי קשר לינארי) נתון הקשר

$$Y = -6X + 22,$$

כאשר  $\mathbb{E}[X] = 3$  ו- $\text{Var}[X] = 2$ :

(א) יש לחשב  $\mathbb{E}[Y]$  ו- $\text{Var}[Y]$ .

(ב) יש לחשב  $\text{Cov}[X, Y]$  ואת מקדם הקורלציה, ולהסביר את התוצאה.

3.6 (מדידות רדיוס רועשות) נתונות מדידות רדיוס של מעגלים. ידוע שהרדיוסים מתפלגים  $r \sim U[a = 1, b = 3]$  ושגיאת המדידה היא  $n \sim \mathcal{N}(\mu = 0, \sigma^2 = 0.01)$ . המדידה ושגיאת המדידה בלתי תלויות, ותוצאת המדידה היא  $X = r + n$ . ההיקף והשטח של המעגל הנמדד הם

$$C = 2\pi X, \quad A = \pi X^2$$

יש לחשב:

(א) מהו ההיקף הממוצע של המעגל הנמדד,  $\mathbb{E}[C]$ ?

(ב) מהי השונות של ההיקף,  $\text{Var}[C]$ ?

(ג) מהו השטח הממוצע של המעגל הנמדד,  $\mathbb{E}[A]$ ?

(ד) יש לחשב  $\text{Cov}[X, r]$ .

3.7 (חסרי קורלציה אך תלויים) נגדיר  $\Theta \sim U[-\pi, \pi]$  וכן

$$X = \cos(\Theta), \quad Y = \sin(\Theta)$$

(א) יש להוכיח שהמשתנים  $X, Y$  חסרי קורלציה.

(ב) יש לחשב  $\mathbb{E}[X^2]$  ו- $\mathbb{E}[Y^2]$ .

(ג) יש להראות ע"י דוגמה נגדית, שמדובר במשתנים שאינם בלתי תלויים. רמז: עבור משתנים בלתי תלויים משוואה (3.23) חייבת להתקיים עבור כל זוג פונקציות  $g_1, g_2$ .

3.8 (שונות של הפרש) נתונים  $X, Y$  משתנים אקראיים כלשהם.

(א) יש להוכיח ביטוי עבור  $\text{Var}[X - Y]$ .

(ב) כעת נתון שהמשתנים  $X \sim \mathcal{N}(\mu_1, \sigma_1^2)$  ו- $Y \sim \mathcal{N}(\mu_2, \sigma_2^2)$  הם גאוסיים במשותף עם מקדם קורלציה  $\rho$ . יש לחשב  $\mathbb{E}[X - Y]$  ו- $\text{Var}[X - Y]$  כפונקציה של פרמטרי ההתפלגות.

3.9 (CDF של זוג גאוס) נתונים זוג משתנים גאוסיים במשותף,  $X \sim \mathcal{N}(\frac{1}{2}, 1)$  ו- $Y \sim \mathcal{N}(\frac{1}{2}, 2^2)$ , בעלי מקדם קורלציה  $\rho = \frac{1}{2}$ . בהינתן פונקציות ה-CDF מהצורה  $F_X(x), F_Y(y), F_{XY}(x, y)$ , יש לבטא באמצעותן את ההסתברויות

$$\Pr(-1 < X < 1, Y < 1), \quad \Pr(X < 1, -1 < Y)$$

רמז: ניתן להיעזר בדוגמה 3.7.

3.10 (משתנים דו-ממדיים גאוסיים) נתונים

$$X_1 \sim \mathcal{N}(m, \sigma^2), \quad X_2 \sim \mathcal{N}(m, \sigma^2)$$

גאוסיים במשותף. לא ניתן להניח אי תלות בין  $X_1, X_2$ ! נגדיר משתנים חדשים,

$$\begin{aligned} V &= X_1 \cos(\theta) + X_2 \sin(\theta), \\ W &= -X_1 \sin(\theta) + X_2 \cos(\theta) \end{aligned}$$

(א) יש למצוא כל  $\theta \in [-\pi, \pi]$  עבורו  $V$  ו- $W$  בלתי תלויים.

(ב) מהי ההתפלגות המשותפת של  $(V, W)$  ומה הפרמטרים שלה? בסעיף זה ניתן להניח  $\mathbb{E}[X_1 X_2] = b$ , ולא ניתן להניח את ערך  $\theta$  מהסעיף הקודם.

3.11 (סכום משתני Bernoulli) נתונים 3 משתנים אקראיים בלתי תלויים בעלי התפלגות Bernoulli,  $X_1 \sim \text{Ber}(p_1)$ ,  $X_2 \sim \text{Ber}(p_2)$ ,  $X_3 \sim \text{Ber}(p_3)$ .

(א) יש לחשב את ההסתברויות  $\Pr(X_1 + X_2 + X_3 = 1)$  ו- $\Pr(X_1 + X_2 + X_3 = 2)$ .

(ב) יש לבדוק את התוצאה ע"י סימולציה עבור ערכי  $p_i \neq 0.5$ .

3.12 (סימולציה) נתונים משתנים אקראיים בלתי תלויים  $X, Y \sim \mathcal{U}[0, 1]$ . יש להראות ע"י סימולציה את הקשר

$$\Pr\left(X^{\frac{1}{3}} > \frac{Y}{2}\right) = \frac{31}{32} = 0.96875$$

3.13 (סימולציה של זוג בדיד) יש לבצע סימולציה של סכום זוג קוביות הוגנות ולשרטט את ההתפלגות האמפירית של הסכום. יש להשוות את התוצאה להתפלגות המדויקת, המתקבלת ע"י ספירת 36 הזוגות שווי ההסתברות.

## נספחים

## 3.א'. בניית זוג גאוסִי דו-ממדי

משוואה (3.41) הוצגה ללא גזירה. הבנייה להלן מראה מהיכן מגיעים חמשת הפרמטרים שלה: כל זוג גאוסִי במשותף הוא תערובת לינארית של שני משתנים גאוסִיים סטנדרטיים בלתי תלויים.

יהיו  $W_1, W_2 \sim \mathcal{N}(0, 1)$  בלתי תלויים, ונבנה

$$(א3.43) \quad X_1 = \mu_1 + \sigma_{11}W_1 + \sigma_{12}W_2$$

$$(ב3.43) \quad X_2 = \mu_2 + \sigma_{21}W_1 + \sigma_{22}W_2$$

כל אחד מ- $X_1, X_2$  הוא סכום משוקלל של משתנים גאוסִיים בלתי תלויים, ולכן גאוסִי ע"פ משוואה (3.40). אותו דבר נכון עבור כל קומבינציה לינארית  $a_1X_1 + a_2X_2$ , מאחר וגם היא מצטמצמת לסכום משוקלל של  $W_1$  ו- $W_2$ . ע"פ הגדרה 3.4 הזוג הוא אפוא גאוסִי במשותף.

התוחלות נובעות מלינאריות התוחלת ומ- $\mathbb{E}[W_i] = 0$ ,

$$(א3.44) \quad \mathbb{E}[X_1] = \mu_1 + \sigma_{11}\mathbb{E}[W_1] + \sigma_{12}\mathbb{E}[W_2] = \mu_1$$

$$(ב3.44) \quad \mathbb{E}[X_2] = \mu_2$$

השונות נובעות ממשוואה (3.34) ומ- $\text{Var}[W_i] = 1$ ,

$$(א3.45) \quad \text{Var}[X_1] = \sigma_{11}^2 \text{Var}[W_1] + \sigma_{12}^2 \text{Var}[W_2] = \sigma_{11}^2 + \sigma_{12}^2 \equiv \sigma_1^2$$

$$(ב3.45) \quad \text{Var}[X_2] = \sigma_{21}^2 + \sigma_{22}^2 \equiv \sigma_2^2$$

עבור השונות המשותפת נפתח את המכפלה. האיברים המעורבים ב- $W_1W_2$  נעלמים מאחר והשניים בלתי תלויים ובעלי תוחלת אפס, והאיברים  $\mathbb{E}[W_i^2] = 1$  שורדים,

$$\mathbb{E}[X_1X_2] = \mu_1\mu_2 + \sigma_{11}\sigma_{21}\mathbb{E}[W_1^2] + \sigma_{12}\sigma_{22}\mathbb{E}[W_2^2]$$

$$(א3.46) \quad = \mu_1\mu_2 + \sigma_{11}\sigma_{21} + \sigma_{12}\sigma_{22}$$

$$\text{Cov}[X_1, X_2] = \mathbb{E}[X_1X_2] - \mathbb{E}[X_1]\mathbb{E}[X_2]$$

$$(ב3.46) \quad = \sigma_{11}\sigma_{21} + \sigma_{12}\sigma_{22}$$

ומקדם הקורלציה הוא

$$(3.47) \quad \rho = \frac{\text{Cov}[X_1, X_2]}{\sqrt{\text{Var}[X_1]\text{Var}[X_2]}} = \frac{\sigma_{11}\sigma_{21} + \sigma_{12}\sigma_{22}}{\sigma_1\sigma_2}$$

ארבעה משקלי ערבוב  $\sigma_{11}, \sigma_{12}, \sigma_{21}, \sigma_{22}$  מייצרים רק שלושה גדלים שונים,  $\sigma_1^2, \sigma_2^2$  ו- $\rho$ . יחד עם שתי התוחלות  $\mu_1, \mu_2$ , אלה הם בדיוק חמשת הפרמטרים המופיעים במשוואה (3.41). שינוי המשקלים תוך שמירה על שלושת הגדלים הנגזרים שקול לסיבוב של  $(W_1, W_2)$ , ואינו משנה את הצפיפות המשותפת.

## 4. חיזוי לינארי

**מטרה:** לחזות משתנה אקראי אחד מתוך משנהו, ולכמת כמה מרוויח החיזוי לעומת התעלמות מוחלטת מהמשתנה השני.

### תוכן הפרק

|    |                           |
|----|---------------------------|
| 44 | 4.1. חיזוי ע"י קבוע       |
| 45 | 4.2. חיזוי לינארי דו-ממדי |
| 47 | 4.3. שאלות                |

### 4.1. חיזוי ע"י קבוע

**מטרה:** למצוא את המספר הבודד המייצג בצורה הטובה ביותר משתנה אקראי.

**חיזוי ע"י קבוע:** חיזוי מהצורה

$$(4.1) \quad \hat{Y} = b,$$

כאשר  $b$  הוא מספר קבוע, המשמש עבור כל מימוש של הניסוי.

**שגיאה ריבועית ממוצעת:** איכות החיזוי נמדדת ע"י השגיאה הריבועית הממוצעת,

$$(4.2) \quad \text{MSE}(b) = \mathbb{E} \left[ (Y - \hat{Y})^2 \right]$$

והחיזוי הטוב ביותר הוא זה הממזער אותה,

$$(4.3) \quad b_{opt} = \arg \min_b \text{MSE}(b)$$

עבור חיזוי ע"י קבוע, פתיחת הסוגריים של השגיאה הריבועית נותנת שלושה מומנטים,

$$(4.4) \quad \begin{aligned} \text{MSE}(b) &= \mathbb{E} \left[ (Y - b)^2 \right] \\ &= \mathbb{E} \left[ Y^2 - 2bY + b^2 \right] \\ &= \mathbb{E} \left[ Y^2 \right] - 2b \mathbb{E}[Y] + b^2 \end{aligned}$$

רק שני האיברים האחרונים תלויים ב- $b$ . השוואת הנגזרת לאפס,

$$(4.5) \quad \frac{\partial \text{MSE}(b)}{\partial b} = -2 \mathbb{E}[Y] + 2b = 0,$$

נותנת את הקבוע האופטימלי,

$$(4.6) \quad b_{opt} = \mathbb{E}[Y]$$

הצבה חזרה במשוואה (4.2) משחזרת את ההגדרה של השונות,

$$(4.7) \quad \text{MSE}_{\min} = \mathbb{E} \left[ \left( Y - \mathbb{E}[Y] \right)^2 \right] = \text{Var}[Y]$$

### מקור המושג "תוחלת"

זוהי התשובה לשאלה שנוטרה פתוחה בסעיף 1.2.3. כאשר יש לסכם משתנה אקראי ע"י מספר בודד, התוחלת היא המספר שיש להשתמש בו: זהו הערך שצפוי לקבל במובן הריבועים הפחותים. המחיר המשולם עבור החלפת המשתנה באותו מספר בודד הוא השונות. שני הגדלים המרכזיים של הפרקים הקודמים הם אפוא זוג תואם: החיזוי הקבוע הטוב ביותר ושגיאתו.

## 4.2. חיזוי לינארי דו-ממדי

**מטרה:** לשפר את החיזוי ע"י קבוע, תוך שימוש במשתנה שני המתואם עמו.

**חיזוי לינארי:** חיזוי מהצורה

$$(4.8) \quad \hat{Y} = ax + b,$$

כאשר  $x$  הוא המימוש הנצפה של  $X$  עבורו נעשה החיזוי.<sup>1</sup>

שני המקדמים נבחרים שוב כך שימזערו את השגיאה הריבועית הממוצעת,

$$(4.9) \quad a_{\text{opt}}, b_{\text{opt}} = \arg \min_{a,b} \text{MSE}(a, b)$$

גזירה לפי כל אחד מהמקדמים בתורו והשוואת שתי הנגזרות לאפס,

$$(4.10) \quad \begin{aligned} \frac{\partial \text{MSE}(a, b)}{\partial a} &= \mathbb{E} [2(Y - aX - b)(-X)] \\ &= -2\mathbb{E}[XY] + 2a\mathbb{E}[X^2] + 2b\mathbb{E}[X] = 0 \end{aligned}$$

$$(4.11) \quad \begin{aligned} \frac{\partial \text{MSE}(a, b)}{\partial b} &= \mathbb{E} [2(Y - aX - b)(-1)] \\ &= -2\mathbb{E}[Y] + 2a\mathbb{E}[X] + 2b = 0 \end{aligned}$$

מייצרות זוג משוואות לינאריות ב- $a$  וב- $b$ ,

$$(4.12) \quad \begin{bmatrix} \mathbb{E}[X^2] & \mathbb{E}[X] \\ \mathbb{E}[X] & 1 \end{bmatrix} \begin{bmatrix} a \\ b \end{bmatrix} = \begin{bmatrix} \mathbb{E}[XY] \\ \mathbb{E}[Y] \end{bmatrix}$$

שפתרון נרשם בצורה הקומפקטית ביותר באמצעות השונות המשותפת והשונות,

$$(4.13) \quad a_{\text{opt}} = \frac{\mathbb{E}[XY] - \mathbb{E}[X]\mathbb{E}[Y]}{\mathbb{E}[X^2] - (\mathbb{E}[X])^2} = \frac{\text{Cov}[X, Y]}{\text{Var}[X]}$$

$$(4.13) \quad b_{\text{opt}} = \mathbb{E}[Y] - a_{\text{opt}}\mathbb{E}[X]$$

<sup>1</sup>מכונה גם קו מגמה לינארי (linear trendline) או רגרסיה לינארית (linear regression).

**חיזוי לינארי בין זוג משתנים אקראיים:** צירוף שני המקדמים נותן את החיזוי בצורתו הקריאה ביותר,

$$\begin{aligned}\hat{Y} &= a_{opt}x + b_{opt} \\ (4.14) \quad &= \mathbb{E}[Y] + \frac{\text{Cov}[X, Y]}{\text{Var}[X]}(x - \mathbb{E}[X])\end{aligned}$$

ניתן לקרוא זאת כתיקון לחיזוי ע"י קבוע של משוואה (4.6): מתחילים ב- $\mathbb{E}[Y]$ , ואז מתרחקים ממנו ביחס ישר למרחק שבו נופל הערך הנצפה  $x$  מ- $\mathbb{E}[X]$ . מקדם היחס הוא השונות המשותפת מנורמלת ע"י השונות של המשתנה הנצפה.

**שגיאת חיזוי:** שגיאת החיזוי הלינארי האופטימלי היא השונות של  $Y$ , מוקטנת בפקטור התלוי במקדם הקורלציה בלבד,

$$(4.15) \quad \text{MSE}_{min} = \mathbb{E}\left[\left(Y - (a_{opt}X + b_{opt})\right)^2\right] = \text{Var}[Y] \left(1 - \rho_{XY}^2\right)$$

#### קריאת פקטור השיפור

- בהשוואה בין משוואה (4.15) לבין משוואה (4.7), תצפית ב- $X$  מקטינה את שגיאת החיזוי בפקטור  $1 - \rho_{XY}^2$ :
- $\rho_{XY} = 0$ : אין כל רווח, ומשוואה (4.14) מתכנסת חזרה לקבוע  $\mathbb{E}[Y]$ .
  - $\rho_{XY} = \pm 1$ : השגיאה מתאפסת, מאחר ו- $Y$  הוא פונקציה לינארית מדויקת של  $X$ .
  - בתווך,  $\rho_{XY}^2$  הוא החלק מתוך השונות של  $Y$  שהחיזוי הלינארי מסלק.
- מאחר ו- $\rho_{XY}$  נכנס בריבוע, הסימן של הקורלציה אינו משנה: קשר שלילי חזק שימושי לחיזוי בדיוק כמו קשר חיובי חזק.

**שגיאה ממוצעת אפס:** החיזוי הלינארי האופטימלי הוא חסר הטיה: בממוצע הוא אינו חורג כלפי מעלה ואינו חורג כלפי מטה,

$$(4.16) \quad \mathbb{E}[Y - (a_{opt}X + b_{opt})] = 0$$

הוכחה. ניקח את התוחלת איבר איבר ונציב את  $b_{opt}$  ממשוואה (4.13),

$$\begin{aligned}(4.17) \quad \mathbb{E}[Y - (a_{opt}X + b_{opt})] &= \mathbb{E}[Y] - a_{opt}\mathbb{E}[X] - b_{opt} \\ &= \mathbb{E}[Y] - a_{opt}\mathbb{E}[X] - (\mathbb{E}[Y] - a_{opt}\mathbb{E}[X]) = 0\end{aligned}$$

**אופטימליות עבור זוג גאוסי במשותף:** עבור זוג גאוסי במשותף (סעיף 3.4), משוואה (4.14) אינה רק החיזוי הלינארי הטוב ביותר: זהו החיזוי הטוב ביותר מכל צורה שהיא. שום פונקציה לא לינארית של  $X$  אינה משיגה שגיאה ריבועית ממוצעת נמוכה יותר.

#### גדלים של ההתפלגות, ולא של הנתונים

כל האמור בפרק זה נרשם באמצעות האופרטורים  $\mathbb{E}[\cdot]$ ,  $\text{Var}[\cdot]$  ו- $\text{Cov}[\cdot]$ , המתארים את ההתפלגות עצמה ואינם זמינים בפועל. בהינתן מערך נתונים, כל אחד מהם מוחלף במקבילה המדגמית שלו: התוחלת בממוצע המדגמי, והשונות המשותפת בשונות המשותפת המדגמית. החיזוי של משוואה (4.14) הופך אז לקו הריבועים הפחותים החד-משתני, ופקטור השיפור  $1 - \rho_{XY}^2$  הופך לגודל המדגמי  $1 - r_{xy}^2$ . השלב הזה, ומה שעלול להשתבש בו, הוא נושאם של פרקי הריבועים הפחותים בהמשך החוברת.

**דוגמה 4.1:** נתונים זוג משתנים אקראיים בלתי תלויים  $W_1, W_2 \sim \mathcal{U}[0, 1]$ , ומתקיים

$$\begin{aligned}X &= W_1, \\ Y &= 0.9W_1 + 0.1W_2\end{aligned}$$

יש לחשב  $\rho_{XY}$ ,  $\text{Cov}[X, Y]$ ,  $\mathbb{E}[XY]$ ,  $\text{Var}[Y]$ ,  $\text{Var}[X]$ ,  $\mathbb{E}[X]$ ,  $\mathbb{E}[Y]$ , את החיזוי הלינארי ואת שגיאתו.

פתרון: שני המשתנים חולקים את האיבר  $W_1$ , וזה מה שיוצר ביניהם את הקורלציה. איור 4.1 מציג את פיזור הנקודות המתקבל יחד עם החיזוי.

$$\mathbb{E}[X] = \mathbb{E}[W_1] = \frac{1}{2}$$

$$\mathbb{E}[Y] = 0.9 \mathbb{E}[W_1] + 0.1 \mathbb{E}[W_2] = \frac{1}{2}$$

$$\text{Var}[X] = \text{Var}[W_1] = \frac{1}{12}$$

$$\begin{aligned} \text{Var}[Y] &= 0.9^2 \text{Var}[W_1] + 0.1^2 \text{Var}[W_2] \\ &= \frac{82}{100} \cdot \frac{1}{12} = \frac{41}{600} \approx 0.06833 \end{aligned}$$

המומנט המעורב דורש את המומנט השני של משתנה אחד,  $\mathbb{E}[W_1^2] = \text{Var}[W_1] + (\mathbb{E}[W_1])^2 = \frac{1}{12} + \frac{1}{4} = \frac{1}{3}$ , ואת העובדה ש- $W_1$  ו- $W_2$  בלתי תלויים,

$$\begin{aligned} \mathbb{E}[XY] &= 0.9 \mathbb{E}[W_1^2] + 0.1 \mathbb{E}[W_1] \mathbb{E}[W_2] \\ &= \frac{9}{10} \cdot \frac{1}{3} + \frac{1}{10} \cdot \frac{1}{2} \cdot \frac{1}{2} = \frac{3}{10} + \frac{1}{40} \\ \text{Cov}[X, Y] &= \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y] \\ &= \frac{13}{40} - \frac{1}{4} = \frac{3}{40} = 0.075 \\ \rho_{XY} &= \frac{3/40}{\sqrt{\frac{1}{12} \cdot \frac{41}{600}}} = \frac{9}{\sqrt{82}} \approx 0.9939 \end{aligned}$$

הצבה במשוואות (4.14) ו-(4.15),

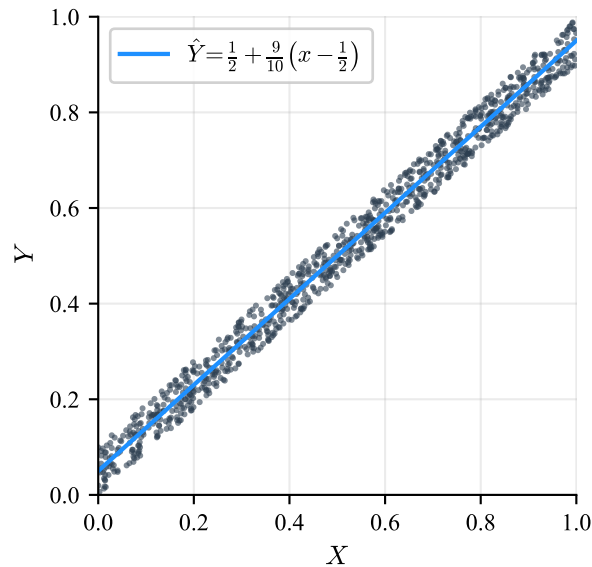
$$\begin{aligned} \hat{Y} &= \frac{1}{2} + \frac{3/40}{1/12} \left( x - \frac{1}{2} \right) = \frac{1}{2} + \frac{9}{10} \left( x - \frac{1}{2} \right) \\ \text{MSE}_{\min} &= \text{Var}[Y] \left( 1 - \rho_{XY}^2 \right) = \frac{41}{600} \left( 1 - \frac{81}{82} \right) \\ &= \frac{1}{1200} \approx 8.33 \times 10^{-4} \end{aligned}$$

השיפוע המשוחזר הוא 0.9, בדיוק המשקל שבו נכנס  $W_1$  לתוך  $Y$ , ושגיאת השארית היא מה שתורם האיבר  $0.1W_2$ .

הפקודות עבור הסימולציה:

```
N = 1e3;
W1 = rand(1,N);
W2 = rand(1,N);
X = W1;
Y = 9/10*W1+1/10*W2;
Yh = 1/2+9/10*(X-1/2);
```

## 4.3. שאלות



איור 4.1: פיזור הנקודות של הזוג המתואם בדוגמה 4.1 יחד עם החיזוי הלינארי האופטימלי.

4.1 (חיזוי) נתון חזאי לינארי של  $Y$  מתוך  $X = x$  מהצורה  $\hat{Y} = ax$ , ללא קבוע חופשי.

(א) יש לחשב ערך אופטימלי של  $a$  במובן של שגיאה ריבועית מינימלית.

(ב) יש לחשב את השגיאה הממוצעת המתקבלת,  $\mathbb{E}[Y - \hat{Y}]$ .

(ג) יש לחשב את השגיאה הריבועית הממוצעת המתקבלת,  $\mathbb{E}[(Y - \hat{Y})^2]$ .

4.2 (חיזוי לא-לינארי) נתון חזאי ממשתנה אקראי  $X$  למשתנה אקראי  $Y$ , מהצורה הריבועית  $\hat{Y} = ax^2$ .

(א) חשב מקדם  $a$  אופטימלי במובן שגיאה ריבועית ממוצעת מינימלית, וביטוי לשגיאה המתקבלת.

(ב) עבור המשתנים

$$\begin{aligned} X &= 0.9W_1 + 0.1W_2, \\ Y &= W_1^2, \end{aligned}$$

כאשר  $W_1, W_2 \sim U[0, 1]$  ובלתי תלויים, חשב ערך מספרי של שגיאת החיזוי המתקבלת.

נוסחאות עזר:

$$(x + y)^2 = x^2 + 2xy + y^2$$

$$(x + y)^3 = x^3 + 3x^2y + 3xy^2 + y^3$$

$$(x + y)^4 = x^4 + 4x^3y + 6x^2y^2 + 4xy^3 + y^4$$

(ג) חשב את שגיאת החיזוי של חזאי לינארי מהצורה  $\hat{Y} = ax + b$  עבור אותם משתנים, והשווה את התוצאה לסעיף הקודם.

4 (חיזוי לינארי של משתנים בדידים) נתון משתנה אקראי דו-ממדי בעל ארבעת הערכים  $\begin{bmatrix} X \\ Y \end{bmatrix} = \left\{ \begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} 1 \\ 1 \end{bmatrix}, \begin{bmatrix} 2 \\ 2 \end{bmatrix}, \begin{bmatrix} 2 \\ 3 \end{bmatrix} \right\}$

בעלי הסתברות זהה. יש לחשב:

(א) את ההתפלגויות השוליות.

(ב) את התוחלת והשונות של כל התפלגות שולית.

(ג) את השונות המשותפת ואת מקדם הקורלציה.

(ד) את החיזוי הלינארי של  $Y$  מתוך  $X$ .

4.4 (חיזוי לינארי של משתנים בדידים) נתונה טבלה של חמש תוצאות ניסוי אפשריות בעלות הסתברות זהה,

$$\begin{bmatrix} X \\ Y \end{bmatrix} = \left\{ \begin{bmatrix} -2 \\ 0 \end{bmatrix}, \begin{bmatrix} -1 \\ 0 \end{bmatrix}, \begin{bmatrix} 0 \\ 3 \end{bmatrix}, \begin{bmatrix} 1 \\ 3 \end{bmatrix}, \begin{bmatrix} 2 \\ 4 \end{bmatrix} \right\}$$

(א) מהו החיזוי הלינארי של  $Y$  עבור  $x = 1$ ?

(ב) מהו מקדם הקורלציה בין  $X$  ל- $Y$ ?

(ג) מהי השגיאה הריבועית הממוצעת המינימלית של החיזוי של  $Y$  מתוך  $x$  שרירותי?

4.5 (חיזוי לינארי מטבלת PDF משותף) נתון זוג משתנים בדידים  $X, Y$ , שה-PDF המשותף שלהם מופיע בטבלה 4.1. לדוגמה,  $\Pr[X = 1, Y = 3] = 0.2$ . יש לחשב:

(א) את מקדם הקורלציה  $\rho_{XY}$ .

(ב) את החיזוי הלינארי של  $Y$  עבור  $x = 3$ .

|   |   | Y   |     |     |
|---|---|-----|-----|-----|
|   |   | 1   | 2   | 3   |
| X | 1 | 0   | 0   | 0.2 |
|   | 2 | 0   | 0.2 | 0.1 |
|   | 3 | 0.1 | 0.2 | 0   |
|   | 4 | 0.2 | 0   | 0   |

טבלה 4.1: PDF משותף עבור שאלה 4.5.

4.6 (חיזוי של משתני Bernoulli) נתונים זוג משתנים אקראיים בלתי תלויים  $X \sim \text{Ber}(p_1)$ ,  $Y \sim \text{Ber}(p_2)$ , בעלי התפלגות משותפת

$$p_{XY}[0, 0] = 0.08$$

$$p_{XY}[0, 1] = 0.32$$

$$p_{XY}[1, 0] = 0.12$$

$$p_{XY}[1, 1] = 0.48$$

(א) מהם ערכי  $p_1$  ו- $p_2$ ?

עבור יתר הסעיפים, יש להשאיר  $p_1, p_2$  כפרמטרים.

(ב) נדרש חיזוי של  $X$  ע"י קבוע  $b_x$  עבורו השגיאה הריבועית מינימלית. מהי שגיאת החיזוי המתקבלת?

(ג) נתון הקשר  $Z = aX + cY$ . נדרש חיזוי של  $Z$  ע"י קבוע  $b_z$  עבורו מתקבלת שגיאה ריבועית מינימלית, ומהי שגיאת החיזוי?

(ד) האם קיימים קבועים  $a \neq 0, c \neq 0$  כדי שהמשתנים  $X$  ו- $Z$  יהיו חסרי קורלציה? אם כן, מהו התנאי המתקבל?

4.7 (קשר בין משתנים) נתון כי  $(X, Y)^T$  הוא וקטור גאוסי משותף בעל תוחלת אפס, כך ש

$$\text{Var}[X] = 4, \quad \text{Var}[Y] = 1, \quad \rho_{XY} = \frac{1}{2}$$

(א) יש לחשב את  $\text{Cov}[X, Y]$ .

(ב) יש לחשב את  $\mathbb{E}[(X + Y)^2]$  ואת  $\text{Var}[X + Y]$ .

(ג) יש למצוא את מעריך ה-MMSE הליניארי הטוב ביותר של  $X$  בהינתן  $Y$ , מהצורה  $\hat{X} = aY$ , ולחשב את שגיאת ה-MSE המתקבלת  $\mathbb{E}[(X - \hat{X})^2]$ .

(ד) מהי ההסתברות  $\Pr(X + Y > 1)$ ?

4.8 (תאוריה) נתונים משתנים אקראיים גאוסיים במשותף  $X_1 \sim \mathcal{N}(0, 1), X_2 \sim \mathcal{N}(1, 2), X_3 \sim \mathcal{N}(0, 1)$  (בסימון  $(\mathcal{N}(\mu, \sigma^2))$ ). ידוע שמקדם הקורלציה בין  $X_1$  ל- $X_2$  הוא  $\rho_{X_1 X_2} = \frac{1}{\sqrt{2}}$ . בנוסף,

$$Y_1 = X_2 + X_3 \sim \mathcal{N}(1, 5)$$

$$Y_2 = -X_1 + 2X_3 \sim \mathcal{N}(0, 3)$$

$$Y_3 = X_1 + 2X_2$$

$$Y_4 = 3X_1 + 2X_2$$

(א) מהו החיזוי האופטימלי במובן שגיאה ריבועית מינימלית של  $X_2$  מתוך  $X_1$ ?

(ב) מהו החיזוי האופטימלי במובן שגיאה ריבועית מינימלית של  $Y_4$  מתוך  $Y_3$ ?

(ג) עבור כל אחד מזוגות המשתנים  $(X_1, X_3)$  ו- $(X_2, X_3)$ , יש להוכיח האם הם בלתי תלויים. רמז: ניתן להיעזר בשונויות הנתונות של  $Y_1$  ו- $Y_2$ .

(ד) מהי ההתפלגות המשותפת של  $Y_3$  ו- $Y_4$ ?

4.9 (היקף ושטח של מלבן) נתונים שני משתנים אקראיים רציפים בלתי תלויים  $X, Y \sim \mathcal{U}[0, 1]$  המגדירים צלעות של מלבן. יהיו  $C$  היקף המלבן ו- $A$  שטח המלבן.

(א) רשמו ביטויים ל- $C$  ול- $A$ .

(ב) חשבו את השונות המשותפת בין  $C$  ל- $A$ .

(ג) הציעו דרך לשערך לינארי של היקף המלבן מתוך שטחו.

(ד) מהי השגיאה הריבועית הממוצעת המינימלית של השערך הלינארי?

# 5. Descriptive Statistics (סטטיסטיקה תיאורית)

מטרה: לתאר את המאפיינים התמציתיים של קבוצת נתונים.

## תוכן הפרק

|    |                                       |       |
|----|---------------------------------------|-------|
| 51 | Population Variance-1 Sample Variance | 5.1   |
| 52 | הגדרות                                | 5.1.1 |
| 52 | אפיון באמצעות Repeated Experiments    | 5.1.2 |
| 54 | Bias                                  | 5.1.3 |
| 56 | שגיאת תקן ורווח סמך                   | 5.2   |
| 56 | שגיאת תקן                             | 5.2.1 |
| 57 | רווח סמך                              | 5.2.2 |
| 59 | Mode, חציון, Skewness ו-Kurtosis      | 5.3   |
| 59 | חציון                                 | 5.3.1 |
| 60 | Mode                                  | 5.3.2 |
| 60 | Skewness                              | 5.3.3 |
| 61 | Kurtosis                              | 5.3.4 |
| 62 | Monte-Carlo Simulation                | 5.4   |
| 62 | מספר הניסויים                         | 5.4.1 |
| 66 | שחזוריות                              | 5.4.2 |
| 66 | שאלות                                 | 5.5   |

## 5.1 Population Variance-1 Sample Variance

### הנחות יסוד

נניח ניסוי אקראי חד-ממדי, המתואר ע"י משתנה אקראי ממשי  $X$  בעל

$$\begin{aligned}\mathbb{E}[X] &= \mu \\ \text{Var}[X] &= \sigma^2.\end{aligned}$$

יהיו

(5.1)  $\mathbf{x} = \{x_1, \dots, x_n\}$

$n$  תצפיות של  $X$ , כאשר  $n$  יכול להיקבע מראש או להיבחר באופן שרירותי.

### 5.1.1. הגדרות

**Sample mean** בהינתן התצפיות  $x_1, \dots, x_n$ , ה-sample mean (ממוצע מדגמי)  $\bar{x}$  נתון ע"י

$$(5.2) \quad \bar{x} = \frac{1}{n} \sum_{i=1}^n x_i$$

**Sample variance** ה-sample variance (שונות מדגמית) נתונה ע"י

$$(5.3) \quad s_{\text{unbiased}}^2 = \frac{1}{n-1} \sum_{i=1}^n (x_i - \bar{x})^2$$

נוסחה זו היא unbiased (חסרת הטיות).

**Population variance** נוסחת ה-population variance (שונות האוכלוסייה), האינטואיטיבית יותר אך biased (מוטת), היא

$$(5.4) \quad s_{\text{biased}}^2 = \frac{1}{n} \sum_{i=1}^n (x_i - \bar{x})^2$$

**סטיית תקן** סטיית התקן (std) של המדגם היא פשוט השורש הריבועי של השונות המדגמית (ה-unbiased או ה-biased),  $s$ .

טבלה 5.1: ההבדל בין שונות לסטיית תקן.

| היבט      | שונות ( $s^2$ )               | סטיית תקן ( $s$ )      |
|-----------|-------------------------------|------------------------|
| יחידות    | (היחידה המקורית) <sup>2</sup> | היחידה המקורית         |
| פירוש     | "סטייה ריבועית ממוצעת"        | "סטייה ממוצעת מהממוצע" |
| נחות הצגה | מופשטת (יחידות בריבוע)        | מוחשית ( $\pm 5$ ק"ג)  |

### 5.1.2. אפיון באמצעות Repeated Experiments

**מטרה:** לקשור בין הגדלים התאורטיים לגדלים המחושבים,  $\mu \stackrel{?}{\leftarrow} \bar{x}$ ,  $\sigma^2 \stackrel{?}{\leftarrow} s_{\text{biased}}^2$  ו-  $\sigma^2 \stackrel{?}{\leftarrow} s_{\text{unbiased}}^2$ .

גישת ה-repeated experiments (ניסויים חוזרים) מבוססת על

•  $k$  ניסויים

– בכל ניסוי  $j = 1, \dots, k$  נדגם מדגם  $x_j$  של  $n$  תצפיות מהמשתנה האקראי הידוע  $X$ .

#### ממוצע

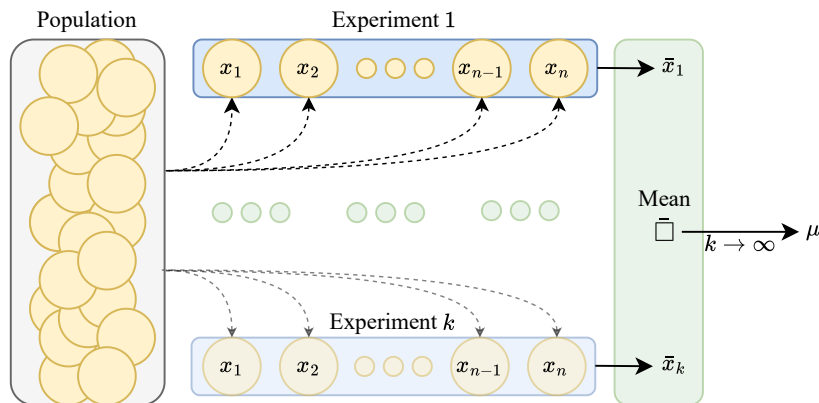
הממוצע של ערכי ה-sample mean הוא עצמו בקירוב  $\mu$ , ולכן בממוצע

$$(5.5) \quad \bar{\mu}_k = \frac{1}{k} \sum_{j=1}^k \bar{x}_j$$

משחזר את התוחלת האמיתית,

$$(5.6) \quad \lim_{k \rightarrow \infty} \bar{\mu}_k \rightarrow \mu$$

עיקרון זה מודגם באיור 5.1.



איור 5.1: המחשה של התכנסות הממוצע של  $k$  ניסויים לתוחלת האמיתית.

התנודות של  $\bar{\mu}_k$ :

- ככל שנאספות יותר תצפיות ( $n$  גדול יותר),  $\bar{x}$  נוטה להתייצב: הוא מתנודד פחות, וההתפלגות האמפירית שלו מתרכזת סביב המרכז האמיתי  $\mu$ .
- יתרה מזאת, עבור  $n$  גדול הפיזור של  $\bar{x}$  קטן כמו  $\sigma/\sqrt{n}$  (או  $\text{Var}[\bar{x}] \approx \frac{\sigma^2}{n}$ ), כלומר ככל שנאסוף יותר נתונים, ההערכה של מרכז התהליך תהיה הדוקה יותר.

## שונות

במקרה של שונות biased, הממוצע של הערכים  $s_1^2, \dots, s_k^2$

$$(5.7) \quad \bar{s}_k^2 = \frac{1}{k} \sum_{j=1}^k s_j^2$$

מעריך בחסר, באופן שיטתי, את השונות האמיתית  $\sigma^2$ ,

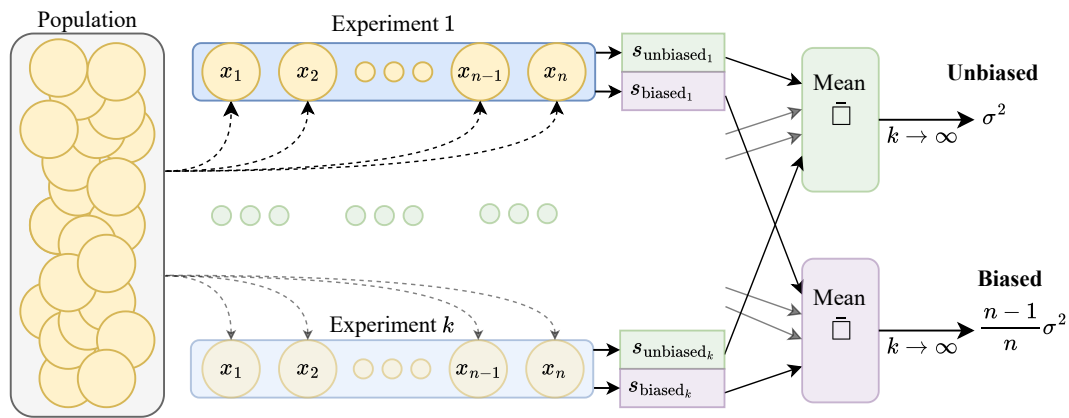
$$(5.8) \quad \lim_{k \rightarrow \infty} \bar{s}_{k,\text{biased}}^2 \rightarrow \frac{n-1}{n} \sigma^2$$

הפרש מובנה זה נקרא bias (הטיה). לעומת זאת, עבור השונות ה-unbiased,

$$(5.9) \quad \lim_{k \rightarrow \infty} \bar{s}_{k,\text{unbiased}}^2 \rightarrow \sigma^2$$

עיקרון זה מודגם באיור 5.2.

ההפרש בין  $s_{\text{biased}}$  ל- $s_{\text{unbiased}}$  קטן ככל ש- $n$  גדל.



איור 5.2: המחשה של התכנסות הממוצע של  $k$  ניסויים לשונות ה-biased ולשונות ה-unbiased.

## MSE

הסיבה שהמערך ה-biased בכל זאת שימושי היא כאשר מעניינת השגיאה הריבועית הממוצעת (MSE) האמפירית. ה-MSE מוגדרת ע"י

$$(5.10) \quad \overline{\text{MSE}}_{\text{unbiased}} = \frac{1}{k} \sum_{j=1}^k \left( s_{\text{unbiased},j}^2 - \sigma^2 \right)^2 \rightarrow \frac{2}{n-1} \sigma^4$$

$$(5.11) \quad \overline{\text{MSE}}_{\text{biased}} = \frac{1}{k} \sum_{j=1}^k \left( s_{\text{biased},j}^2 - \sigma^2 \right)^2 \rightarrow \frac{2n-1}{n^2} \sigma^4$$

ומתקיים

$$(5.12) \quad \overline{\text{MSE}}_{\text{unbiased}} \geq \overline{\text{MSE}}_{\text{biased}}$$

ה-MSE של הנוסחה ה-biased נמוכה במעט מזו של הנוסחה ה-unbiased.

**דוגמה 5.1 (היעדר bias ושונות של ה-sample mean, ומעריכי שונות biased מול unbiased):** ה-sample mean הוא unbiased, והשונות שלו דועכת כמו  $\sigma^2/n$ . מעריכי השונות המדגמית המקובלים יכולים להיות biased או unbiased. נדגים את שלוש התכונות באמצעות סימולציה:

1. פרמטרים.

- ההתפלגות האמיתית:  $X \sim \mathcal{N}(0, 1)$ , ולכן  $\mu = 0$ ,  $\sigma^2 = 1$ .
- גודלי מדגם:  $n \in \{5, 20, 100\}$ .
- מספר חזרות:  $k = 5000$ .

2. תוצאות. התוצאות בטבלה 5.2 מתאימות היטב לערכים התאורטיים.

### 5.1.3 Bias

ה-bias של סטטיסטי מדגמי הוא ההפרש השיטתי בין הממוצע שלו לטווח ארוך, על פני ניסויים חוזרים, לבין הערך האמיתי המתאים של המשתנה האקראי. סטטיסטי נקרא unbiased כאשר הפרש זה הוא אפס, ו-biased אחרת.

טבלה 5.2: תוצאות הסימולציה: ה-sample mean, הערכות השונות, וה-MSE שלהן, עבור גודלי מדגם שונים.

| $\overline{MSE}_{unbiased}$ | $\overline{MSE}_{biased}$ | $\bar{s}_{unbiased}^2$ | $\bar{s}_{biased}^2$ | $\text{Var}[\bar{x}] = 1/n$<br>תאורטי | $\text{Var}[\bar{x}]$<br>אמפירי | $\bar{\mu}_n$ | $n$ |
|-----------------------------|---------------------------|------------------------|----------------------|---------------------------------------|---------------------------------|---------------|-----|
| 0.50                        | 0.36                      | 0.99                   | 0.79                 | 0.200                                 | 0.198                           | 0.001         | 5   |
| 0.11                        | 0.10                      | 1.00                   | 0.95                 | 0.050                                 | 0.049                           | -0.000        | 20  |
| 0.02                        | 0.02                      | 1.01                   | 0.99                 | 0.010                                 | 0.010                           | 0.000         | 100 |

לדוגמה, עבור השונות המדגמית,

$$(5.13) \quad \text{bias}(s^2) = \lim_{k \rightarrow \infty} \bar{s}_k^2 - \sigma^2.$$

לכן  $s_{unbiased}^2$  הוא unbiased, בעוד של  $s_{biased}^2$  יש bias של  $-\sigma^2/n$ .

הוכחה. יהיו  $n$  הדגימות בלתי תלויות. נכתוב כל סטייה סביב  $\bar{x}$  כסטייה סביב  $\mu$  פחות השגיאה של הממוצע,  $\sum_i (x_i - \mu) = n(\bar{x} - \mu)$  ו- $x_i - \bar{x} = (x_i - \mu) - (\bar{x} - \mu)$  ונפתח את הריבוע.

$$\begin{aligned} \sum_{i=1}^n (x_i - \bar{x})^2 &= \sum_{i=1}^n (x_i - \mu)^2 - 2(\bar{x} - \mu) \sum_{i=1}^n (x_i - \mu) + n(\bar{x} - \mu)^2 \\ &= \sum_{i=1}^n (x_i - \mu)^2 - n(\bar{x} - \mu)^2. \end{aligned}$$

זוהי זהות, המתקיימת עבור כל מדגם בנפרד: הפיזור הנמדד סביב  $\bar{x}$  קטן תמיד מהפיזור סביב  $\mu$ , בדיוק ב- $n(\bar{x} - \mu)^2$ .

לשני האיברים באגף ימין יש תוחלות ידועות. הראשון הוא  $n\sigma^2$ , לפי הגדרת  $\sigma^2$ . השני הוא  $n\text{Var}[\bar{x}]$ , ו- $\text{Var}[\bar{x}] = \sigma^2/n$  היא התנודה של ה-sample mean שתוארה לעיל. לכן

$$\begin{aligned} \mathbb{E} \left[ \sum_i (x_i - \bar{x})^2 \right] &= n\sigma^2 - n \cdot \frac{\sigma^2}{n} = (n-1)\sigma^2 \\ \mathbb{E} \left[ s_{biased}^2 \right] &= \frac{n-1}{n} \sigma^2 \\ \text{bias} \left( s_{biased}^2 \right) &= \frac{n-1}{n} \sigma^2 - \sigma^2 = -\frac{\sigma^2}{n}. \end{aligned}$$

חלוקת אותו סכום ב- $n-1$  במקום ב- $n$  מכפילה את התוצאה ב- $n/(n-1)$ , וזה מבטל בדיוק את הגורם  $(n-1)/n$ , ולכן  $s_{unbiased}^2$  הוא unbiased. הסיבה היא ש- $\bar{x}$  עצמו מותאם לנתונים: הוא קרוב יותר לדגימות מאשר  $\mu$ , ודרגת חופש אחת מנוצלת למיקומו.

**דוגמה 5.2 (מתי ה-bias משנה):** עבור איזה  $n$  ה-bias היחסי של  $s_{biased}^2$  הוא לכל היותר 1%?

פתרון: ה-bias היחסי הוא  $|\sigma^2/n|/\sigma^2 = 1/n$ , יהא  $\sigma$  אשר יהא, ולכן

$$\frac{1}{n} \leq 0.01 \quad \Rightarrow \quad n \geq 100.$$

טבלה 5.2 מראה את אותו גורם בפעולה. התוחלות החזיוניות  $\frac{n-1}{n}\sigma^2 = 0.80$ , ו-0.95 ו-0.99 עבור  $n = 5$ , 20 ו-100 מתאימות לעמודה המסומלצת  $\bar{s}_{biased}^2$ , שבה 0.79, 0.95 ו-0.99. מתחת לכמאה דגימות יש לציין את הנרמול; מעליהן, הבחירה משנה את הערך המדויק בפחות מ-1%.

השימושים במעריכים biased ו-unbiased:

- Biased: במשימות למידת מכונה (למשל, פונקציית הפסד), ואופטימלי (שערוך maximum-likelihood, נראות מרבית) עבור התפלגות גאוסית.
- Unbiased: סטטיסטיקה.

ברירת המחדל במימושי קוד משתנה:

- Python משתמשת בביטוי ה-biased, למשל `numpy.std`.
- Matlab משתמשת בביטוי ה-unbiased, למשל `std`.

## 5.2. שגיאת תקן ורווח סמך

**מטרה:** להצמיד מספר למידה שבה הערכה הייתה זזה אילו הניסוי היה חוזר על עצמו, ולהפוך מספר זה לתחום.

### 5.2.1. שגיאת תקן

לתנודות של  $\bar{x}$  שתוארו לעיל יש שם וביטוי סגור. שגיאת התקן (standard error) היא סטיית התקן של הערכה, בניגוד לסטיית התקן של הנתונים שמהם ההערכה חושבה.

עבור ה-sample mean של  $n$  דגימות בלתי מתואמות בעלות שונות  $\sigma^2$ , השונות של  $n$  האיברים מתחברות, בעוד שהגורם  $1/n$  שלפני הסכום עולה בריבוע,

$$\text{Var}[\bar{x}] = \frac{1}{n^2} \sum_{i=1}^n \text{Var}[x_i] = \frac{n\sigma^2}{n^2} = \frac{\sigma^2}{n}, \quad (5.14)$$

כך ש-

$$\text{SE}(\bar{x}) = \sqrt{\text{Var}[\bar{x}]} = \frac{\sigma}{\sqrt{n}}. \quad (5.15)$$

השורש מופיע מאחר והשונות הן שמתחברות, ולא האמפליטודות.

את משוואה (5.15) יש לקרוא כתקציב: הדיוק של הממוצע משתפר רק כשורש של גודל המדגם, ולכן הקטנת שגיאת התקן בחצי עולה פי ארבעה נתונים. ה- $\sigma$  של האוכלוסייה ידוע לעיתים רחוקות, ובפועל מחליפים אותו בסטיית התקן המדגמית  $s_x$ ,

$$\widehat{\text{SE}}(\bar{x}) = \frac{s_x}{\sqrt{n}}. \quad (5.16)$$

#### $\sigma$ או SE

השניים עונים על שאלות שונות.  $\sigma$  הוא הפיזור של הנתונים, והוא אינו משתנה כאשר נאספים נתונים נוספים. שגיאת התקן היא הפיזור של ההערכה, והיא קטנה כמו  $1/\sqrt{n}$ . דיווח על  $\sigma$  במקום שבו הכוונה ל- $\text{SE}(\bar{x})$  מגזים באי הוודאות של הממוצע, והמקרה ההפוך מסתיר את הפיזור של האוכלוסייה.

טבלה 5.2 מציגה את הכלל על נתונים מסומלצים. כשגיאות תקן, כלומר לאחר הוצאת שורש, שלוש השונות האמפיריות של  $\bar{x}$  הן 0.445, 0.221 ו-0.100, לעומת  $\sigma/\sqrt{n} = 0.447$ , 0.224 ו-0.100. הגדלת  $n$  פי ארבעה, מ-5 ל-20, מקטינה את שגיאת התקן בחצי, בדיוק כפי שמשוואה (5.15) דורשת.

### דגימות בלתי מתואמות

משוואה (5.15) מניחה ש- $n$  הדגימות בלתי מתואמות. דגימות מתואמות חוזרות על מידע שהדגימות הקודמות כבר נשאו, ולכן הן שוות פחות ממה שמספרן מרמז: שגיאת התקן האמיתית גדולה אז מזו שמשוואה (5.16) מדווחת, וכל טענת דיוק המבוססת עליה אופטימית מדי.

## 5.2.2. רווח סמך

שגיאת התקן היא קנה מידה, ועדיין לא טענה על מיקום הערך האמיתי. רווח סמך (confidence bound) הופך אותה לתחום: טווח המחושב מתוך הנתונים, שמכסה את הערך האמיתי בחלק  $1 - \alpha$  מהניסויים החוזרים, כאשר  $\alpha$  הוא הסיכון המקובל להחטיא אותו. הבחירה הרגילה היא  $\alpha = 0.05$ , רווח של 95%.

**תוחלת** יהיו  $n$  התצפיות בלתי תלויות ושוות התפלגות  $\mathcal{N}(\mu, \sigma^2)$ , כאשר שני הפרמטרים אינם ידועים. זהו הקושי: טענה על  $\mu$  נראית כדורשת את  $\sigma$ , שאינו זמין יותר מ- $\mu$  עצמו. ה-sample mean המנורמל (2.43) הוא הדרך לעקוף זאת. ב- $(\bar{x} - \mu) / (s_x / \sqrt{n}) \sim t_{n-1}$  כל גודל באגף שמאל, פרט ל- $\mu$  עצמו, מחושב מתוך המדגם, בעוד שההתפלגות באגף ימין אינה תלויה לא ב- $\mu$  ולא ב- $\sigma$ . לכן התפלגות זו ידועה עוד לפני שנאספה תצפית אחת, וזה בדיוק מה שהצעד הבא דורש.

מאחר ו- $t_{n-1}$  סימטרית סביב אפס, השארת  $\alpha/2$  מההסתברות בכל זנב פירושה הישארות בין  $\pm t_{n-1, 1-\alpha/2}$ ,

$$\Pr \left( -t_{n-1, 1-\alpha/2} \leq \frac{\bar{x} - \mu}{s_x / \sqrt{n}} \leq t_{n-1, 1-\alpha/2} \right) = 1 - \alpha.$$

זוהי עדיין טענה על הסטטיסטי, ולא על  $\mu$ , אך ניתן לפתור את אי השוויון עבור  $\mu$ . נכפיל את שלושת האגפים ב- $s_x / \sqrt{n}$ , נחסיר  $\bar{x}$ , ולבסוף נהפוך את הסימנים, מה שהופך את כיווני אי השוויון ומשאיר את  $\mu$  לבדו באמצע,

$$(5.17) \quad \Pr \left( \bar{x} - t_{n-1, 1-\alpha/2} \frac{s_x}{\sqrt{n}} \leq \mu \leq \bar{x} + t_{n-1, 1-\alpha/2} \frac{s_x}{\sqrt{n}} \right) = 1 - \alpha.$$

שני הקצוות ניתנים לחישוב מתוך המדגם בלבד, ולכן משוואה (5.17) היא רווח הסמך בכתב מפורש. בכתב מקוצר,

$$(5.18) \quad \bar{x} \pm t_{n-1, 1-\alpha/2} \frac{s_x}{\sqrt{n}},$$

כאשר  $t_{n-1, 1-\alpha/2}$  הוא ה-quantile העליון מסעיף 2.2.5.

הרווח סימטרי סביב  $\bar{x}$ , והגורם שממנו הוא בנוי,  $s_x / \sqrt{n}$ , הוא בדיוק שגיאת התקן המוערכת (5.16). לכן הוא מצטמצם כמו  $1 / \sqrt{n}$ , באותו קצב ומאותה סיבה כמו שגיאת התקן עצמה. רוחבו נקבע ע"י ה-quantile, וה-quantile נלקח מ- $t_{n-1}$  ולא מההתפלגות הגאוסית מאחר ו- $\sigma$  הוערך ולא היה ידוע. סעיף 2.2.5 מראה כיצד מחיר זה קטן ככל ש- $n$  גדל.

**שוונות** השונות נובעת מאותו טיעון, עם (2.39) במקום (2.43). תוצאה זו קובעת  $(n-1)s_x^2 / \sigma^2 \sim \chi_{n-1}^2$ , והיא מתאימה מאותה סיבה: ה- $\sigma^2$  הלא ידוע נמצא באגף שמאל, בעוד שההתפלגות באגף ימין אינה כוללת אף גודל לא ידוע. הצפיפות של chi-square אינה סימטרית, ולכן כאן כל אחד משני הזנבות דורש quantile משלו,

$$\Pr \left( \chi_{n-1, \alpha/2}^2 \leq \frac{(n-1)s_x^2}{\sigma^2} \leq \chi_{n-1, 1-\alpha/2}^2 \right) = 1 - \alpha.$$

פתרון עבור  $\sigma^2$  פירושו מעבר להופכיים, מה שהופך שוב את כיווני אי השוויון ולכן מחליף בין שני הקצוות: ה-quantile הגדול מגיע לגבול התחתון,

$$(5.19) \quad \left[ \frac{(n-1)s_x^2}{\chi_{n-1, 1-\alpha/2}^2}, \frac{(n-1)s_x^2}{\chi_{n-1, \alpha/2}^2} \right].$$

התוצאה אינה סימטרית סביב  $s_x^2$ . האסימטריה הימנית של הצפיפות מציבה את שני ה-quantiles במרחקים שונים מהמרכז, כך שהקצה העליון הוא הרחוק יותר. הוצאת שורש משני הקצוות נותנת את הרווח עבור  $\sigma$ .

**דוגמה 5.3 (רווח סמך מתוך עשר תצפיות):** מדגם של  $n = 10$  תצפיות נותן  $\bar{x} = 4.2$  ו- $s_x = 1.3$  (unbiased). מצאו את רווחי הסמך של 95% עבור התוחלת, השונות וסטיית התקן.

פתרון: שגיאת התקן המוערכת (5.16) ושלושת ה-quantiles עבור  $\alpha = 0.05$  עם  $n - 1 = 9$  דרגות חופש הם

$$\widehat{SE}(\bar{x}) = \frac{1.3}{\sqrt{10}} = 0.411, \quad t_{9, 0.975} = 2.262, \quad \chi_{9, 0.975}^2 = 19.02, \quad \chi_{9, 0.025}^2 = 2.70.$$

התוחלת, לפי (5.18),

$$4.2 \pm 2.262 \cdot 0.411 = 4.2 \pm 0.93 = [3.27, 5.13].$$

השונות, לפי (5.19), כאשר  $(n-1)s_x^2 = 9 \cdot 1.69 = 15.21$ ,

$$\left[ \frac{15.21}{19.02}, \frac{15.21}{2.70} \right] = [0.80, 5.63],$$

וסטיית התקן היא השורש של קצוות אלו,  $[0.89, 2.37]$ .

כדאי לקרוא את שתי התוצאות זו לצד זו. עשר תצפיות ממקמות את התוחלת בדיוק של  $\pm 0.93$ , בעוד שהשונות נקבעת רק עד כדי פקטור של שבע בין הקצוות,  $5.63/0.80 \approx 7$ . זוהי הצורה המוחשית של מה שהשוואת ה-MSE בסעיף 5.1 אומרת באופן מופשט: אותם נתונים שממקמים את המרכז בצורה סבירה בקושי מגבילים את הפיזור, ושונות המדווחת מתוך מדגם קטן נושאת הרבה פחות מידע ממה שהספרות שלה מרמזות.

#### מה אקראי כאן

משוואה (5.17) מזמינה קריאה שגויה. הרווח הוא העצם האקראי, ולא הפרמטר:  $\mu$  הוא מספר קבוע, ושני הקצוות הם שזים מניסוי לניסוי. רווח של 95% פירושו שאילו כל הניסוי היה חוזר פעמים רבות, 95% מהרווחים שנבנו כך היו מכילים את הערך האמיתי. עבור הרווח היחיד שחושב בפועל, הערך האמיתי נמצא בתוכו או שאינו נמצא בו; ה-95% מתאר את השיטה, ולא את זוג המספרים המסוים.

#### הנחת גאוסיות

הנחת הגאוסיות שהוצגה לעיל אינה משפיעה באותה מידה על שני הרווחים. רווח התוחלת (5.18) עמיד לנתונים שאינם גאוסיים עבור  $n$  בינוני, מאחר ומשפט הגבול המרכזי הופך את  $\bar{x}$  לכמעט גאוס, תהא ההתפלגות של  $X$  אשר תהא. רווח השונות (5.19) אינו עמיד: הוא נשען על (2.39), המתקיימת עבור נתונים גאוסיים בלבד, ונתונים בעלי זנבות כבדים הופכים אותו לאופטימי מאוד.

**דוגמה 5.4 (רווח סמך, והמחיר של הקטנתו בחצי):** נמדדת אצווה של  $n = 16$  חלקים, ומתקבלים  $\bar{x} = 12.0$  מ"מ ו- $s_x = 3.2$  מ"מ (unbiased).

(א) העריכו את שגיאת התקן של הממוצע.

(ב) מצאו את רווח הסמך של 95% עבור התוחלת.

(ג) מצאו את רווחי הסמך של 95% עבור השונות ועבור סטיית התקן.

(ד) כמה חלקים יש למדוד כדי להקטין בחצי את רוחב הרווח של סעיף (ב)?

$$\text{Quantiles: } \chi_{15,0.025}^2 = 6.262, \chi_{15,0.975}^2 = 27.488, t_{15,0.975} = 2.131$$

פתרון: (א) לפי (5.16),

$$\widehat{\text{SE}}(\bar{x}) = \frac{s_x}{\sqrt{n}} = \frac{3.2}{\sqrt{16}} = \frac{3.2}{4} = 0.8 \text{ mm.}$$

כאן  $s_x$  הוא ההערכה ה-unbiased, עם  $n - 1$  במכנה, מאחר וה-quantiles של  $t$  ושל  $\chi^2$  מחושבים עבור נרמול זה.

(ב) לפי (5.18), עם  $n - 1 = 15$  דרגות חופש,

$$t_{15,0.975} \cdot \widehat{\text{SE}}(\bar{x}) = 2.131 \cdot 0.8 = 1.705$$

$$\bar{x} \pm 1.705 = [10.30, 13.71] \text{ mm.}$$

(ג) הרווח (5.19) אינו סימטרי, מאחר והתפלגות  $\chi^2$  אינה סימטרית. עם  $(n-1)s_x^2 = 15 \cdot 3.2^2 = 15 \cdot 10.24 = 153.6$ , רוחב השונות הוא

$$\left[ \frac{153.6}{27.488}, \frac{153.6}{6.262} \right] = [5.59, 24.53] \text{ mm}^2,$$

ורוחב סטיית התקן הוא

$$\left[ \sqrt{5.59}, \sqrt{24.53} \right] = [2.36, 4.95] \text{ mm.}$$

ההערכה הנקודתית  $s_x = 3.2$  אינה נמצאת במרכז  $[2.36, 4.95]$ : הזנב העליון ארוך הרבה יותר.

(ד) חצי הרוחב הוא  $t_{n-1,0.975} s_x / \sqrt{n}$ . אם מתייחסים ל- $t$  ול- $s_x$  כקבועים, הרוחב קטן כמו  $1/\sqrt{n}$ , ולכן הקטנתו בחצי עולה פי ארבעה,

$$n = 4 \cdot 16 = 64.$$

ההערכה שמרנית:  $t_{63,0.975} = 1.998 < 2.131$ , ולכן  $n = 64$  משיג בפועל מעט יותר מהקטנה בחצי.

## 5.3 Mode, חציון, Kurtosis-1 Skewness

### 5.3.1 חציון

חציון הוא ערך הנמצא באמצע התפלגות השכיחויות של הערכים הנצפים, כך שהסתברות ליפול מעליו שווה להסתברות ליפול מתחתיו.

באופן שקול, עבור מדגם  $x_1, \dots, x_n$  בעל סטטיסטי סדר  $x_{(1)} \leq \dots \leq x_{(n)}$ , החציון המדגמי הוא

$$\text{median}(x) = \begin{cases} x_{(\frac{n+1}{2})}, & n \text{ odd,} \\ \frac{x_{(\frac{n}{2})} + x_{(\frac{n}{2}+1)}}{2}, & n \text{ even.} \end{cases}$$

### Mode 5.3.2

עבור מדגם סופי  $x_1, \dots, x_n$ , ה-mode (שכיח) של המדגם הוא כל ערך (או ערכים) המופיע בתדירות הגבוהה ביותר מבין התצפיות.

**Multimodality**: התפלגות נקראת multimodal (רב-שיאית) אם יש לה יותר ממקסימום מקומי אחד (mode) בפונקציית הצפיפות או בהיסטוגרמה שלה. מקרים נפוצים:

- Unimodal (חד-שיאית): שיא יחיד (למשל, התפלגות נורמלית).
  - Bimodal (דו-שיאית): שני שיאים נפרדים, שלרוב מעידים שהנתונים הם תערובת של שתי תת-אוכלוסיות.
  - Multimodal: שלושה שיאים או יותר.
- Multimodality (רב-שיאיות) מרמזת שהנתונים עשויים לנבוע מתערובת של תהליכים או קבוצות שונים.

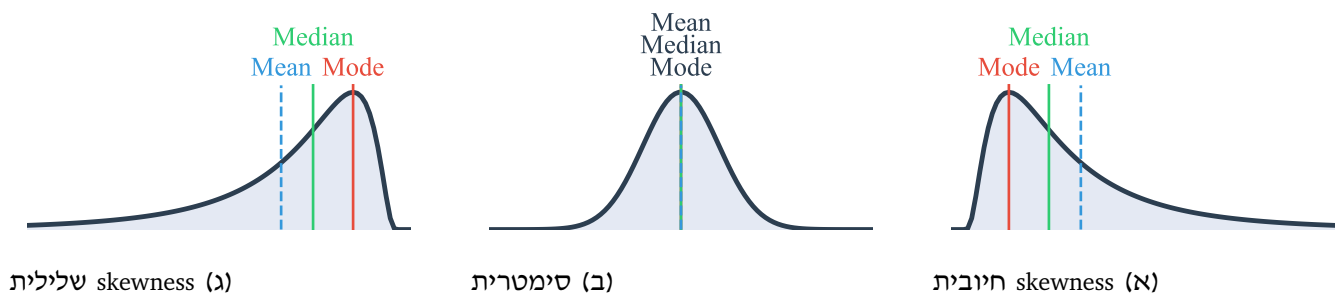
### Skewness 5.3.3

**Skewness**: Skewness (צידוד) מודדת את האסימטריה של התפלגות סביב התוחלת שלה. ה-skewness המדגמית מוגדרת ע"י

$$g_1 = \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_n} \right)^3.$$

- $g_1 = 0$ : ההתפלגות סימטרית (למשל, התפלגות נורמלית).
- $g_1 > 0$ : ההתפלגות right-skewed (בעלת צידוד ימני, skewness חיובית): הזנב הימני ארוך יותר, והממוצע בדרך כלל גדול מהחציון.
- $g_1 < 0$ : ההתפלגות left-skewed (בעלת צידוד שמאלי, skewness שלילית): הזנב השמאלי ארוך יותר, והממוצע בדרך כלל קטן מהחציון.

המחשה של הקשר בין ממוצע, חציון ו-mode עבור ערכי skewness שונים מופיעה באיור 5.3.



איור 5.3: הקשר בין ממוצע, חציון ו-mode עבור התפלגויות בעלות (א) skewness חיובית, (ב) סימטריה, ו-(ג) skewness שלילית.

### Kurtosis .5.3.4

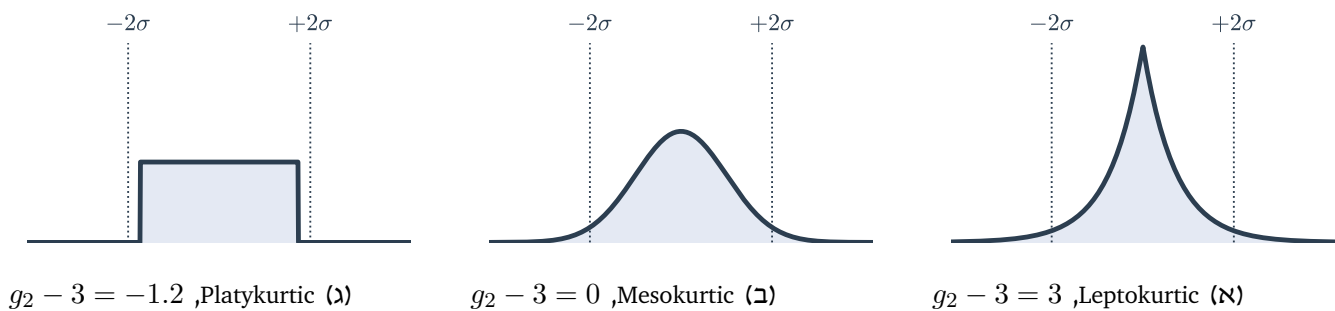
**Kurtosis:** (גבנוניות) מודדת את משקל הזנבות של התפלגות, יחסית להתפלגות נורמלית בעלת אותו פיזור. ה-kurtosis המדגמית מוגדרת ע"י

$$g_2 = \frac{1}{n} \sum_{i=1}^n \left( \frac{x_i - \bar{x}}{s_n} \right)^4,$$

חזקה אחת מעל ה-skewness  $g_1$ . התפלגות נורמלית נותנת  $g_2 = 3$ , ולכן מקובל להחסיר ערך ייחוס זה ולדווח על ה-excess kurtosis (גבנוניות עודפת)  $g_2 - 3$ :

- $g_2 - 3 = 0$ : mesokurtic (מזוקורטית), הזנבות תואמים את אלו של התפלגות נורמלית.
- $g_2 - 3 > 0$ : leptokurtic (לפטוקורטית), זנבות כבדים יותר, כך שתצפיות רחוקות מהתוחלת שכיחות יותר ממה שהתפלגות נורמלית בעלת אותו  $\sigma$  חוזה.
- $g_2 - 3 < 0$ : platykurtic (פלטוקורטית), זנבות קלים יותר, כך שהערכים הקיצוניים למעשה חסומים.

המחשה של שלושת המקרים, עבור תוחלת שווה ו- $\sigma$  שווה, מופיעה באיור 5.4.



איור 5.4: שלוש התפלגויות בעלות אותה תוחלת ואותו  $\sigma$ , הנבדלות רק במשקל הזנבות שלהן: (א) Laplace, (ב) נורמלית, (ג) אחידה. הקווים המנוקדים מסמנים  $\pm 2\sigma$ . הצפיפות ה-leptokurtic שומרת הסתברות הרחק מעבר להם, בעוד שה-platykurtic נעצרת לגמרי לפנייהם. השיא החד יותר של (א) הוא תוצר לוואי של ה- $\sigma$  המותאם; מה ש- $g_2$  משקלל הם הזנבות, מאחר וסטייה של  $2\sigma$  תורמת לסכום פי  $2^4 = 16$  יותר מסטייה של  $1\sigma$ .

כל סטייה מחולקת ב- $s_n$  לפני העלאתה בחזקה רביעית, ולכן  $g_2$  אינו נושא מידע על רוחב הנתונים; הוא מדווח רק על כמות הנתונים הרחוקה מהמרכז. זה מה שהופך אותו לראוי לציטוט לצד סטיית התקן, ולא במקומה. שתי קבוצות נתונים בעלות אותו  $\bar{x}$  ואותו  $s$  עדיין יכולות לייצר ערכים קיצוניים בקצבים שונים מאוד: עבור שתי ההתפלגויות של איור 5.4א ואיור 5.4ב, ערך מעבר ל- $\bar{x} \pm 3s$  מתקבל בהסתברות 0.014 ו-0.0027 בהתאמה, פקטור של חמש. סף הממוקם ב- $3s$  בהנחה של התפלגות נורמלית נחצה אפוא פי חמש מהמתוכנן כאשר הנתונים leptokurtic.

החזקה הרביעית הופכת את  $g_2$  לפונקציה של מעט התצפיות הקיצוניות ביותר במדגם. Outlier (ערך חריג) בודד מזיז אותו משמעותית, ועל קומץ נקודות הוא מדווח בעיקר אילו ערכים נדגמו במקרה. כדאי לצטט אותו מתוך מאות תצפיות, ו-kurtosis המדווחת ללא גודל המדגם שמאחוריה אומרת מעט.

ברירת המחדל במימושי קוד משתנה גם כאן:

- Python מחזירה את הצורה ה-excess, למשל `scipy.stats.kurtosis` עם ברירת המחדל `fisher=True`.
- Matlab מחזירה את  $g_2$  הגולמי, למשל `kurtosis`.

## 5.4 Monte-Carlo Simulation

**מטרה:** להעריך גודל סטטיסטי שאין לו ביטוי סגור נוח, ע"י מיצוע על פני מימושים אקראיים סינתטיים רבים.

ה-repeated experiments של סעיף 5.1 הם מקרה פרטי של שיטה נומרית כללית, הידועה בשם Monte-Carlo simulation (סימולציית מונטה קרלו). במקום לגזור את התנהגות הסטטיסטי באופן אנליטי, הניסוי האקראי משוחזר  $k$  פעמים במחשב, ו- $k$  הערכים המתקבלים מסוכמים באותם כלים תיאוריים בדיוק של פרק זה ושל הפרק הבא.

באופן פורמלי, יהי  $g$  סטטיסטי כלשהו המחושב ממדגם של  $n$  תצפיות, למשל  $\bar{x}$ , או  $s_{\text{biased}}^2$ . ה-MSE. התוחלת שלו מקורבת ע"י הממוצע על פני  $k$  חזרות בלתי תלויות,

$$(5.20) \quad \mathbb{E}[g] \approx \bar{g}_k = \frac{1}{k} \sum_{j=1}^k g_j,$$

כאשר חוק המספרים הגדולים מבטיח  $\bar{g}_k \rightarrow \mathbb{E}[g]$  כאשר  $k \rightarrow \infty$ . הפיזור של  $g$  מוערך מתוך אותם  $k$  ערכים עצמם,

$$(5.21) \quad \text{Var}[g] \approx s_g^2 = \frac{1}{k} \sum_{j=1}^k (g_j - \bar{g}_k)^2, \quad s_g = \sqrt{s_g^2},$$

ו- $\theta - \bar{g}_k$  מעריך את ה-bias ביחס לערך אמיתי ידוע  $\theta$  (סעיף 5.1.3).

הערך של  $s_g$  נקבע ע"י גודל המדגם  $n$  וע"י ההתפלגות של  $X$ : הוא מדווח כמה הסטטיסטי משתנה מניסוי לניסוי, ושום מספר חזרות אינו מקטין אותו. כדי להפוך את הסטטיסטי עצמו לפחות משתנה נדרש  $n$  גדול יותר. מספר החזרות  $k$  שולט בדבר אחר, והוא מידת הדיוק שבה  $\bar{g}_k$  ו- $s_g$  מעריכים את הגדלים שהם מייצגים.

החוזק של השיטה הוא שמשוואה (5.20) אינה דורשת שום פיתוח: היא מתקיימת עבור כל סטטיסטי, כל התפלגות של  $X$  וכל גודל מדגם  $n$ , כולל כל המקרים שבהם ביטוי אנליטי אינו זמין, או תקף רק עבור  $n$  גדול.

### מתכון Monte-Carlo

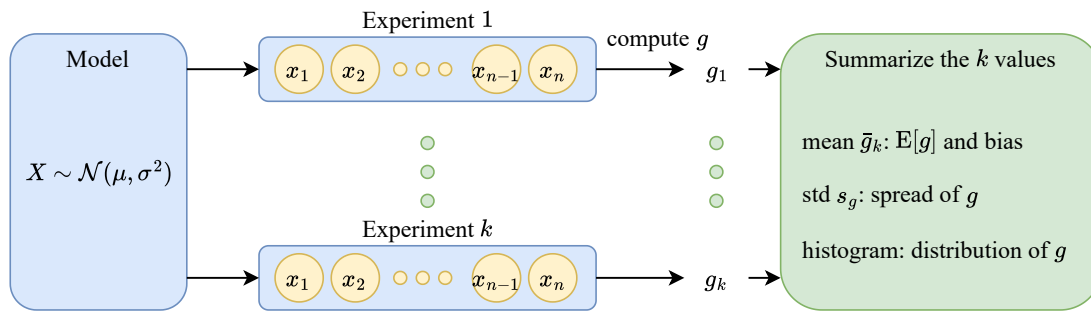
1. קבעו את מודל יצירת הנתונים ואת ערכי הפרמטרים האמיתיים, למשל  $X \sim \mathcal{N}(\mu, \sigma^2)$  עם  $\mu$  ו- $\sigma^2$ . שנבחרו.
2. הגרילו מימוש אחד של  $n$  תצפיות  $x_1, \dots, x_n$ .
3. חשבו מתוך מימוש זה את הסטטיסטי המבוקש  $g_j$ .
4. חזרו על פריטים 2 ו-3 עבור  $j = 1, \dots, k$ , ושמרו את כל  $k$  הערכים.
5. סכמו את הערכים השמורים: הממוצע שלהם (5.20) מעריך את  $\mathbb{E}[g]$  ואת ה-bias (סעיף 5.1.3), סטיית התקן שלהם (5.21)  $s_g$  מעריכה את הפיזור של  $g$ , וההיסטוגרמה, ה-boxplot (דיאגרמת קופסה) או ה-violin plot (דיאגרמת כינור) שלהם חושפים את ההתפלגות המלאה של  $g$ .

המתכון כולו מסוכם באיור 5.5: המודל מזין  $k$  ניסויים בלתי תלויים של  $n$  תצפיות כל אחד, כל ניסוי מצטמצם לערך בודד  $g_j$ , ורק אוסף  $k$  הערכים עובר פירוש.

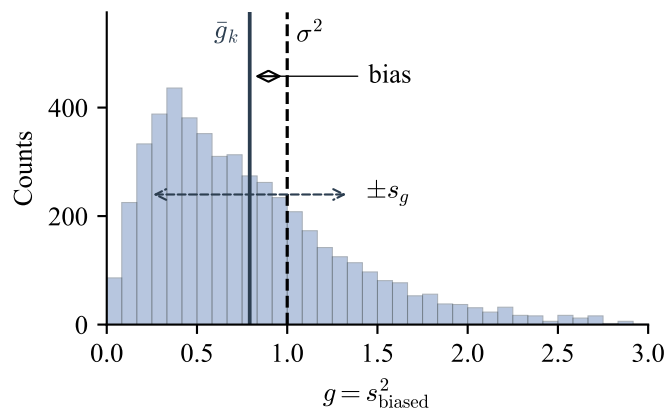
טבלה 5.2 היא בדיוק הפלט של מתכון זה, כשהוא מיושם על ה-sample mean, על שני מערכי השונות ועל ה-MSE שלהם, עם  $n \in \{5, 20, 100\}$  ו- $k = 5000$ . כל כניסה בטבלה מתמצתת  $k$  ערכים שמורים למספר בודד; איור 5.6 מציג את הערכים עצמם שמאחורי כניסה אחת.

### 5.4.1 מספר הניסויים

התוצאה המסומלצת היא עצמה גודל אקראי, ולכן היא נושאת שגיאה משלה, ושגיאה זו נשלטת ע"י  $k$  בלבד.



איור 5.5: מתכון Monte-Carlo. כל אחד מ- $k$  הניסויים מגריל  $n$  תצפיות מאותו מודל, ומצטמצם לערך אחד של הסטטיסטי  $g$ . הערכים השמורים מסוכמים לאחר מכן ע"י הכלים התיאוריים של פרק זה ושל הפרק הבא.



איור 5.6:  $k = 5000$  הערכים השמורים שמאחורי הכניסה  $\bar{s}_{\text{biased}}^2$  בשורה  $n = 5$  של טבלה 5.2. הסטטיסטי הוא  $g = s_{\text{biased}}^2$ , בסימולציה של  $n = 5$  תצפיות של  $X \sim \mathcal{N}(0, 1)$ . הממוצע שלהם,  $\bar{g}_k = 0.79$ , נופל מה- $\sigma^2 = 1$  האמיתי ב-bias של  $-\sigma^2/n = -0.2$ , וסטיית התקן שלהם  $s_g$  (5.21) מודדת כמה  $g$  זז מניסוי לניסוי. ההיסטוגרמה מוסיפה את מה ששני מספרים אלו אינם יכולים להראות: להתפלגות של  $g$  יש skewness חזקה.

**ממוצע מסומלץ** סטיית התקן של ממוצע מסומלץ היא

$$(5.22) \quad \sqrt{\text{Var}[\bar{g}_k]} = \frac{\sqrt{\text{Var}[g]}}{\sqrt{k}},$$

כאשר  $\text{Var}[g]$  הוא הפיזור (5.21) של ה- $g_j$  הבודדים: הדיוק של ממוצע מסומלץ הוא הפיזור של הסטטיסטי חלקי  $\sqrt{k}$ .

**סטיית תקן ושונות מסומלצות** באופן דומה, השגיאה היחסית של סטיית תקן מסומלצת  $s_g$  היא

$$(5.23) \quad \frac{\sqrt{\text{Var}[s_g]}}{s_g} \approx \frac{1}{\sqrt{2k}},$$

בעוד שהשגיאה היחסית של שונות מסומלצת  $s_g^2$  גדולה פי שניים,

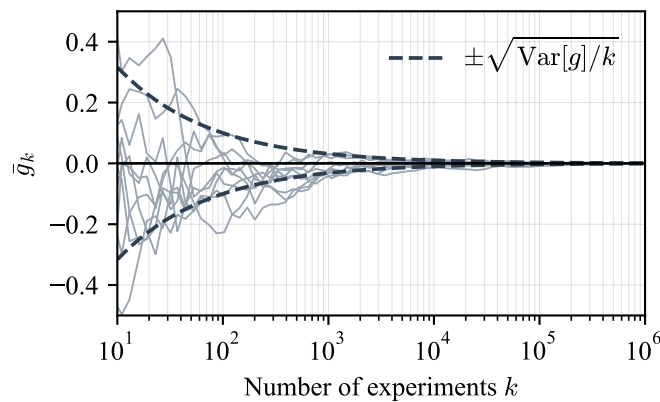
$$(5.24) \quad \frac{\sqrt{\text{Var}[s_g^2]}}{s_g^2} \approx \sqrt{\frac{2}{k}} = \frac{2}{\sqrt{2k}}.$$

הפקטור שתיים נובע מההעלאה בריבוע: שונות היא ריבוע של סטיית תקן, והעלאה בריבוע מכפילה את התנודה היחסית. לכן אי הוודאות של שונות מסומלצת גדולה תמיד פי שניים מזו של סטיית התקן המסומלצת המתקבלת מאותם  $k$  ערכים בדיוק, ואין לצטט את שתייהן באותו דיוק.

שלושתן דועכות כמו  $1/\sqrt{k}$ , ולכן הקטנת שגיאת הסימולציה בחצי עולה פי ארבעה ניסויים. ערכים מייצגים מופיעים בטבלה 5.3, וממוחשים באיור 5.7 עבור הממוצע המסומלץ ובאיור 5.8 עבור סטיית התקן המסומלצת.

טבלה 5.3: שגיאת Monte-Carlo יחסית של ממוצע מסומלץ (5.22), של סטיית תקן מסומלצת (5.23) ושל שונות מסומלצת (5.24), כפונקציה של מספר הניסויים  $k$ .

| $k$    | ממוצע, $1/\sqrt{k}$ | סטיית תקן, $1/\sqrt{2k}$ | שונות, $\sqrt{2/k}$ |
|--------|---------------------|--------------------------|---------------------|
| $10^2$ | 10%                 | 7.1%                     | 14%                 |
| $10^3$ | 3.2%                | 2.2%                     | 4.5%                |
| $10^4$ | 1.0%                | 0.7%                     | 1.4%                |
| $10^5$ | 0.32%               | 0.22%                    | 0.45%               |
| $10^6$ | 0.10%               | 0.07%                    | 0.14%               |

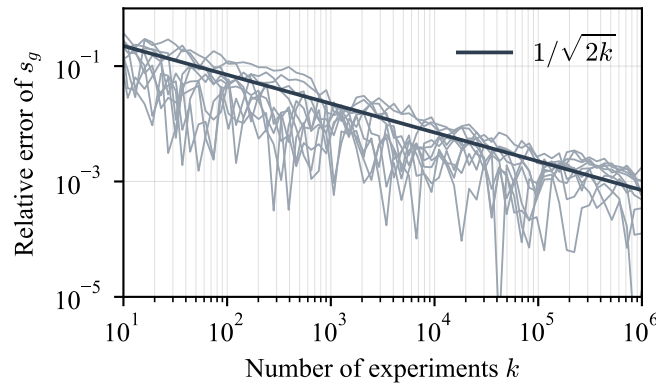


איור 5.7: ממוצע מסומלץ  $\bar{g}_k$  כפונקציה של מספר הניסויים  $k$ . אפור: עשר סימולציות בלתי תלויות. מקווקו: המעטפת  $\pm \sqrt{\text{Var}[g]/k}$  של (5.22). העקומות נשארות בתוך המעטפת, ומתהדקות סביב הערך האמיתי ככל ש- $k$  גדל.

המסקנה היא שערך מסומלץ לעולם אינו מדויק, ומספר הספרות הנכונות שהוא נושא נקבע ע"י  $k$ . בטבלה 5.2, למשל,  $k = 5000$  נותן שגיאת שונות של כ-2% (5.24), ולכן בשורה  $n = 100$  מופיע  $\bar{s}_{\text{unbiased}}^2 = 1.01$  ולא הערך המדויק 1.00.

**דוגמה 5.5 (קביעת גודלו של ניסוי Monte-Carlo):** גודל  $g$  מוערך באמצעות מתכון ה-Monte-Carlo, החוזר על עצמו  $k$  פעמים. כמה חזרות נדרשות כדי ש-

(א) סטיית התקן של הממוצע המסומלץ  $\bar{g}_k$  תהיה לכל היותר 1% מהפיזור  $\sqrt{\text{Var}[g]}$  של התוצאות הבודדות;



איור 5.8: השגיאה היחסית של סטיית התקן  $s_g$ , המוערכת מתוך  $k$  ערכי Monte-Carlo, כפונקציה של  $k$ . אפור: עשר סימולציות בלתי תלויות. שחור: החיזוי  $1/\sqrt{2k}$  של (5.23). הסימולציות הבודדות מתפזרות סביב החיזוי, ורק קצב הדעיכה מובטח.

(ב) סטיית התקן המסומלצת  $s_g$  תישא שגיאה יחסית של לכל היותר 1%;

(ג) השונות המסומלצת  $s_g^2$  תישא שגיאה יחסית של לכל היותר 1%?

פתרון: כל אחת מהמשוואות (5.22), (5.23) ו-(5.24) נפתרת ישירות עבור  $k$ :

$$\frac{\sqrt{\text{Var}[\bar{g}_k]}}{\sqrt{\text{Var}[g]}} = \frac{1}{\sqrt{k}} \leq 0.01 \implies k \geq 10^4 \quad (\text{א})$$

$$\frac{1}{\sqrt{2k}} \leq 0.01 \implies k \geq 5 \cdot 10^3 \quad (\text{ב})$$

$$\sqrt{\frac{2}{k}} \leq 0.01 \implies k \geq 2 \cdot 10^4 \quad (\text{ג})$$

השורה  $k = 10^4$  בטבלה 5.3 מתאימה לכך: הממוצע נמצא בדיוק על 1%, סטיית התקן כבר מתחת לסף, ב-0.7%, והשונות עדיין מעליו, ב-1.4%.

סטיית תקן מסומלצת היא הזולה מבין השלוש, במחצית מהריצות של הממוצע; שונות מסומלצת היא היקרה ביותר, פי ארבעה מהריצות של סטיית התקן. הפקטור ארבע בין סעיף (ב) לסעיף (ג) נובע מההעלאה בריבוע: היא מכפילה את התנודה היחסית, והכפלת שגיאה הדועכת כמו  $1/\sqrt{k}$  עולה פי ארבעה חזרות. מאותה סיבה, כל ספרה עשרונית נוספת של דיוק עולה פי מאה בזמן ריצה, ולכן תוצאות Monte-Carlo מצוטטות בשתיים או שלוש ספרות משמעותיות, ולא יותר.

**פיתוח (\*)** מהיכן מגיעות שלוש הנוסחאות? כל אחת עונה על אותה שאלה: אם כל הסימולציה חוזרת עם seed (זרע) חדש, כמה זי המספר המדויק?

נתחיל בממוצע. הערכים השמורים  $g_1, \dots, g_k$  בלתי תלויים, ולכן השונויות שלהם מתחברות, וחלוקת הסכום שלהם ב- $k$  מחלקת את השונות ב- $k^2$ ,

$$\text{Var}[\bar{g}_k] = \frac{1}{k^2} \sum_{j=1}^k \text{Var}[g_j] = \frac{\text{Var}[g]}{k}.$$

השורש של ביטוי זה הוא (5.22). זהו כלל ה- $\sigma^2/n$  המוכר של ה-sample mean (סעיף 5.2.1), המיושם על  $k$  ניסויים במקום על  $n$  תצפיות. נעשה שימוש באי תלות בלבד, ולכן קצב זה תקף עבור כל התפלגות של  $g$ . זהו גם הקצב

היחיד הניתן כערך מוחלט ולא כאחוז: אחוז היה מחלק ב- $\mathbb{E}[g]$ , שהוא אפס בכל פעם שהסימולציה בודקת bias מול אפס.

השוונות דורשת תוצאה ידועה אחת. עבור ערכים גאומיים, סכום הסטיות הריבועיות מתפלג chi-square (סעיף 2.2.4), שהשוונות שלו היא פעמיים מספר דרגות החופש שלו. עבור  $k$  ערכים זה נותן

$$\text{Var}[s_g^2] \approx \frac{2\sigma_g^4}{k},$$

וחלוקה ב- $s_g^2 \approx \sigma_g^2$  הופכת זאת לאחוז (5.24).

סטיית התקן אינה דורשת הנחה חדשה, אלא רק שורש. שינוי יחסי קטן של השוונות מצטמצם לחצי מעצמו לאחר הוצאת השורש,

$$s_g = \sigma_g \sqrt{1 + \varepsilon} \approx \sigma_g \left(1 + \frac{\varepsilon}{2}\right),$$

כך ששוונות הידועה בדיוק של  $\pm 10\%$  נותנת סטיית תקן הידועה בדיוק של  $\pm 5\%$ . חלוקה של  $\sqrt{2/k}$  בשניים נותנת  $1/\sqrt{2k}$ , וזוהי (5.23).

שימו לב שרק הצעד האמצעי הניח התפלגות גאוסית. כאשר ל- $g$  זנבות כבדים יותר, ערכים גדולים נדירים מנפחים את הסטיות הריבועיות, ואז גם השוונות וגם סטיית התקן פחות מדויקות ממה שהנוסחאות מבטיחות. הממוצע אינו מושפע. נרמול השוונות ב- $k$  או ב- $k-1$  אינו יוצר הבדל נראה ברמות דיוק אלו.

## 5.4.2. שחזוריות

מחשב מגריל מספרים pseudo-random (פסאודו-אקראיים), המיוצרים ע"י אלגוריתם דטרמיניסטי מתוך ערך התחלתי הנקרא seed. קיבוע ה-seed הופך את כל הסימולציה לניתנת לשחזור מדויק, וזוהי דרישה מכל תוצאה מדווחת, ולא אפשרות.

- Python: `numpy.random.default_rng(seed)`
- Matlab: `rng(seed)`

- מימוש בודד אינו מוכיח דבר. כל מסקנה חייבת להיגזר מתוך  $k$  הערכים שנאספו.
- דווחו את  $k$  ושמרו את ה-seed לצד כל מספר מסומלץ, כך שניתן יהיה לשפוט את הדיוק שלו (5.23).
- לעולם אין לבחור את ה-seed שמייצר תוצאה "טובה יותר", כלומר "פגיעה מקרית". ה-seed נקבע לפני הניסוי, ולא אחריו.

## 5.5. שאלות

5.1 (סטטיסטיקה תיאורית של מדגם קטן) חיישן נקרא שמונה פעמים, ומתקבלים הערכים

2, 4, 4, 4, 5, 5, 7, 9.

(א) יש לחשב את ה-sample mean  $\bar{x}$  ואת סכום הסטיות הריבועיות  $\sum_i (x_i - \bar{x})^2$ .

(ב) יש לחשב את השוונות ה-biased ואת סטיית התקן שלה, ואת השוונות ה-unbiased ואת סטיית התקן שלה.

(ג) יש למצוא את החציון, את ה-mode ואת הטווח.

(ד) פי כמה שתי השונות נבדלות זו מזו? עבור איזה  $n$  הבחירה ביניהן משנה את הערך המדויק בלא יותר מ-1%?

(ה) יש להעריך את שגיאת התקן של הממוצע, ולתת את רווח הסמך של 95% עבור  $\bar{x}$ . יש להשוות את רווח הרווח לטווח שנמצא בסעיף (ג). נתון  $t_{7,0.975} = 2.365$ .

5.2 (סימולציה: שני מערכי השונות) יש לאמת באמצעות סימולציה את ה-bias  $-\sigma^2/n$  של סעיף 5.1.3, ואת המחיר של הסרתו.

יש להגדיל  $k = 10^4$  ניסויים בלתי תלויים, שכל אחד מהם הוא  $n = 10$  דגימות מ- $\mathcal{N}(0, 1)$  (randn). בכל ניסוי יש לחשב את שני מערכי השונות, ולאחר מכן לדווח, על פני  $k$  הניסויים:

(א) את הממוצע של כל מעריך, בהשוואה ל- $\sigma^2$  ול- $\frac{n-1}{n}\sigma^2$ ;

(ב) את ה-MSE של כל מעריך סביב  $\sigma^2$  האמיתי, בהשוואה ל-

$$\overline{\text{MSE}}_{\text{unbiased}} = \frac{2}{n-1}\sigma^4, \quad \overline{\text{MSE}}_{\text{biased}} = \frac{2n-1}{n^2}\sigma^4;$$

(ג) באיזה משני המערכים עדיף להשתמש, ומדוע התשובה אינה המעריך ה-unbiased.

יש לקבוע את ה-seed כך שההרצה תהיה ניתנת לשחזור, ולהציג את התוצאות כמו בטבלה 5.2.

## 6. גרפים סטטיסטיים

**מטרה:** להציג את המיקום, הפיזור והצורה של קבוצת נתונים, ולקרוא את מידת האמינות של המוצג.

### תוכן הפרק

|    |                                     |       |
|----|-------------------------------------|-------|
| 68 | Boxplot                             | 6.1   |
| 68 | הגדרות מקדימות                      | 6.1.1 |
| 69 | הגרף                                | 6.1.2 |
| 71 | היסטוגרמה                           | 6.2   |
| 71 | ערכים בדידים                        | 6.2.1 |
| 72 | ערכים רציפים                        | 6.2.2 |
| 73 | קירוב ל-PDF של משתנה אקראי רציף (*) | 6.2.3 |
| 76 | השונויות של מספר התצפיות ב-bin      | 6.2.4 |
| 79 | Violin Plot                         | 6.3   |
| 79 | שאלות                               | 6.4   |

שלושת הגרפים של פרק זה הם גם הסיכומים הנזכרים בצעד 5 של מתכון ה-Monte-Carlo (מונטה קרלו) בסעיף 5.4. שם הם מיושמים על  $k$  הערכים השמורים של סטטיסטי מסומלץ ולא על מדידות גולמיות, אך דבר מהמובא להלן אינו תלוי במקור המספרים.

### 6.1 Boxplot

**מטרה:** הצגה תמציתית של המיקום, הפיזור וה-outliers (ערכים חריגים) הפוטנציאליים של ההתפלגות.

#### 6.1.1 הגדרות מקדימות

**Quartiles:** בהינתן סטטיסטי הסדר  $x_{(1)} \leq \dots \leq x_{(n)}$ , שלושת ה-quartiles (רבעונים) מחלקים את הנתונים הממוינים לארבעה חלקים שווים:

- $Q_1$  (ה-quartile הראשון, האחוזון ה-25): החציון של המחצית התחתונה של הנתונים.
- $Q_2$  (ה-quartile השני, האחוזון ה-50): החציון של כל הנתונים.
- $Q_3$  (ה-quartile השלישי, האחוזון ה-75): החציון של המחצית העליונה של הנתונים.

באופן שקול, כ-25% מהתצפיות נמצאות מתחת ל- $Q_1$ , כ-50% מתחת ל- $Q_2$ , וכ-75% מתחת ל- $Q_3$ .

**Interquartile range:** ה-interquartile range (טווח בין-רבעוני, IQR) מודד את הפיזור של 50% הנתונים המרכזיים,

$$(6.1) \quad \text{IQR} = Q_3 - Q_1.$$

**Outlier:** הוא כל תצפית הנמצאת מחוץ לתחום

$$\left[ Q_1 - 1.5 \text{ IQR}, Q_3 + 1.5 \text{ IQR} \right].$$

נקודות כאלו נחשבות רחוקות באופן חריג מגוף הנתונים, ועשויות להצביע על שגיאות מדידה, על אירועים נדירים, או על התנהגות של זנבות כבדים.

המקדם 1.5 הוצע באופן אמפירי ע"י הסטטיסטיקאי John Tukey בשנות ה-70.

## 6.1.2. הגרף

Boxplot (דיאגרמת קופסה, box-and-whisker plot) מסכם קבוצת נתונים באמצעות חמישה סטטיסטים:

1. הערך המינימלי שאינו outlier.

2. ה-quartile הראשון ( $Q_1$ ).

3. החציון ( $Q_2$ ).

4. ה-quartile השלישי ( $Q_3$ ).

5. הערך המקסימלי שאינו outlier.

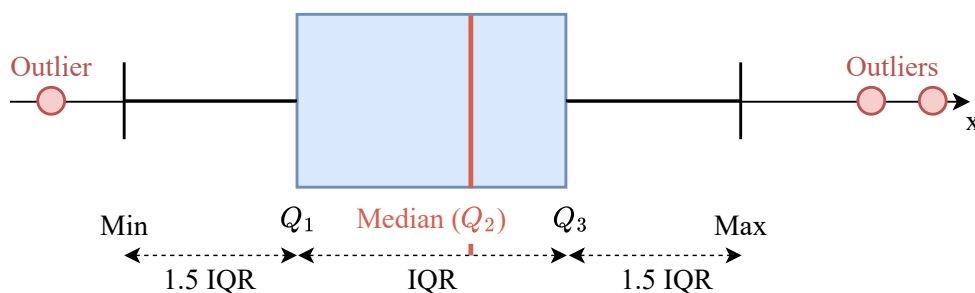
ה-boxplot משורטט באופן הבא:

- תיבה (box) משתרעת מ- $Q_1$  עד  $Q_3$ , עם קו בחציון  $Q_2$ .
- Whiskers (שפמים) נמשכים מהתיבה עד לנקודות הנתונים הקיצוניות ביותר הנמצאות בתוך

$$\left[ Q_1 - 1.5 \text{ IQR}, Q_3 + 1.5 \text{ IQR} \right].$$

- כל תצפית מחוץ לתחום ה-whiskers מסומנת בנפרד כ-outlier פוטנציאלי.

המחשה של boxplot ורכיביו מופיעה באיור 6.1.



איור 6.1: המבנה של boxplot: התיבה משתרעת על פני ה-IQR מ- $Q_1$  עד  $Q_3$ , החציון  $Q_2$  מסומן בתוכה, ה-whiskers נמשכים עד לתצפיות הקיצוניות ביותר שאינן outliers, וה-outliers הבודדים מוצגים מעבר להם.

## פירוש

- רוחב התיבה (IQR) משקף את הפיזור: תיבה רחבה מעידה על פיזור גבוה.
- מיקום קו החציון בתוך התיבה חושף skewness (צידוד): אם החציון קרוב יותר ל- $Q_1$ , ההתפלגות right-skewed (בעלת צידוד ימני); אם הוא קרוב יותר ל- $Q_3$ , היא left-skewed (בעלת צידוד שמאלי).

- אורכי ה-whiskers מצביעים על הטווח של התצפיות האופייניות.
- נקודות בודדות מעבר ל-whiskers מסמנות outliers פוטנציאליים, שייתכן שמצדיקים בדיקה נוספת.

בניגוד להיסטוגרמות, boxplots אינם מציגים את צורת ההתפלגות (למשל, multimodality, רב-שיאיות). הם שימושיים במיוחד להשוואה בין התפלגויות של קבוצות שונות, זו לצד זו.

**דוגמה 6.1 (Boxplot בחישוב ידני):** נתונות אחת עשרה מדידות, כבר ממוינות,

3, 7, 8, 12, 13, 14, 18, 21, 22, 27, 52.

(א) יש למצוא את  $Q_3, Q_2, Q_1$  ואת ה-IQR.

(ב) יש למצוא את גבולות ה-outliers ולרשום את ה-outliers, אם יש.

(ג) היכן מסתיימים ה-whiskers?

(ד) יש לשרטט את ה-boxplot, ולומר במשפט אחד מה הצורה שלו מספרת על הנתונים.

פתרון: Quartiles. מאחר ו- $n = 11$  אי-זוגי,  $Q_2$  הוא סטטיסטי הסדר השישי, וכל חצי נלקח בלעדיו, כך שנשארים חמישה ערכים בכל צד.

$$Q_2 = x_{(6)} = 14$$

$$Q_1 = \text{median}\{3, 7, 8, 12, 13\} = 8$$

$$Q_3 = \text{median}\{18, 21, 22, 27, 52\} = 22$$

$$\text{IQR} = Q_3 - Q_1 = 22 - 8 = 14$$

תוכנות שונות מבצעות אינטרפולציה שונה של ה-quartiles, ולכן פונקציית שרטוט עשויה להחזיר ערכים שונים במעט; הקריאה של הגרף אינה משתנה.

גבולות ו-outliers.

$$1.5 \text{ IQR} = 21$$

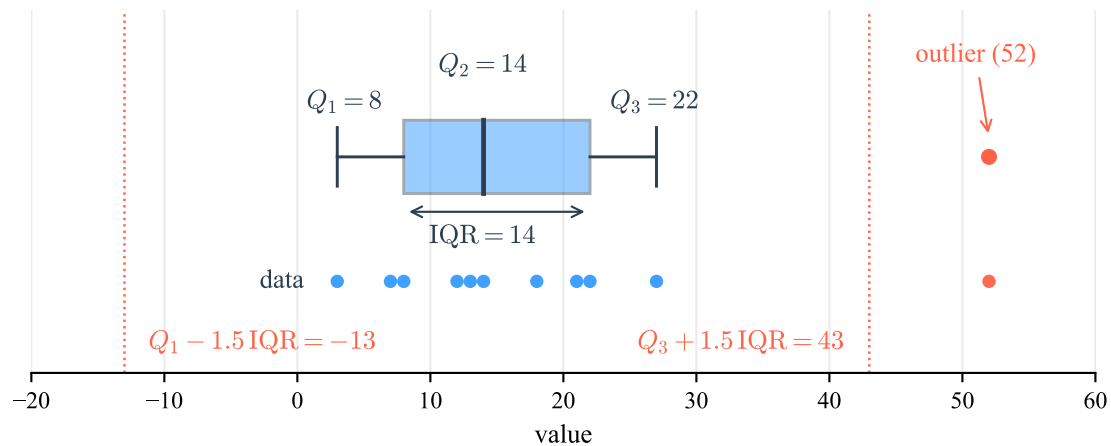
$$[Q_1 - 1.5 \text{ IQR}, Q_3 + 1.5 \text{ IQR}] = [8 - 21, 22 + 21] = [-13, 43]$$

תצפית אחת, 52, נופלת מחוץ לתחום ומסומנת כ-outlier. אף תצפית אינה נופלת מתחת ל-13-, מאחר והגבול הזה רחוק מתחום הנתונים.

Whiskers. ה-whiskers מגיעים עד התצפיות הקיצוניות ביותר שאינן outliers, ולכן הם מסתיימים ב-3 וב-27, לא ב-13- וב-43, וגם לא ב-52. חמשת הסטטיסטים המשורטטים הם לכן

3, 8, 14, 22, 27,

ו-52 מסומן כנקודה נפרדת (איור 6.2).



איור 6.2: ה-boxplot של אחת עשרה המדידות. הציר יורד עד הגבול התחתון ב-13-, הרחק מחוץ לנתונים, בעוד שהגבול העליון ב-43 חוצה אותם: האסימטריה הזו היא הסיבה שבדיוק תצפית אחת מסומנת. ה-whiskers נעצרים ב-3 וב-27, הערכים הקיצוניים ביותר בתוך הגבולות, ולא בגבולות עצמם.

קריאת הצורה.  $Q_2 - Q_1 = 6$  לעומת  $Q_3 - Q_2 = 8$ , וה-whisker העליון ארוך מהתחתון, ולכן הנתונים נוטים מעט ימינה. הנקודה הרחוקה היחידה ב-52 היא בדיוק מה שה-boxplot בנוי לחשוף, וזה הערך שכדאי לבדוק לפני שמחשבים כל דבר אחר מהמדגם.

## 6.2 היסטוגרמה

**מטרה:** הצגה גרפית של תוצאות הניסוי.

### 6.2.1 ערכים בדידים

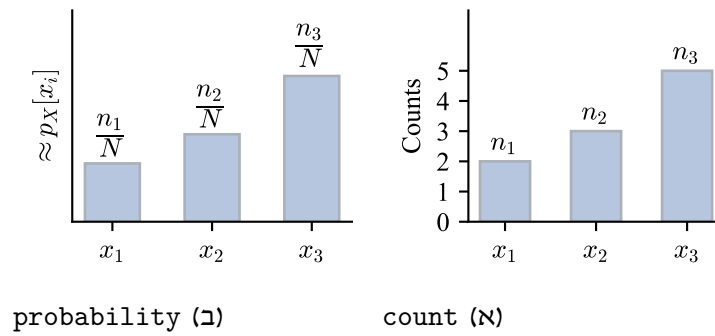
**סוג count**

**מטרה:** הצגה של מספר תוצאות ניסוי מכל סוג.

נתון ניסוי:

- בעל  $k$  תוצאות אפשריות  $x_1, x_2, \dots, x_k$ , כאשר ערך של  $k$  הוא יחסית קטן.
  - סה"כ  $N$  תוצאות ניסוי
  - בניסוי התקבלו:
    - $n_1$  תוצאות מסוג  $x_1$ ,
    - $n_2$  תוצאות מסוג  $x_2$ ,
    - ... וכך הלאה.
- כאשר  $\sum_i n_i = N$ .

העמודה הגבוהה ביותר בהיסטוגרמה מסוג count מזהה את ה-mode (השכיח) של המדגם. דרך גרפית להציג את תוצאות הניסוי היא בדוגמאת איור 6.3 (א).



איור 6.3: דוגמא להיסטוגרמה מסוג: (א) count, (ב) probability.

### סוג probability

**מטרה:** הצגה של שכיחות יחסית של תוצאות ניסוי, שהוא קירוב ל-PDF.

**קירוב ל-PDF:** הסתברות של תוצאה מסויימת היא יחסית למספר תוצאות הניסוי עם אותה תוצאה חלקי סה"כ הניסויים,

$$(6.2) \quad p_X[x_i] \approx \frac{n_i}{N}, \quad i = 1, \dots, k.$$

כמובן, תמיד עדיף  $N \rightarrow \infty$ .

דרך גרפית להציג את תוצאות הניסוי היא בדוגמאת איור 6.3(ב).

### 6.2.2. ערכים רציפים

עבור  $k$  גדול (מאות ויותר, או ערכים רציפים) ישנם 2 בעיות מרכזיות:

- קושי להציג תוצאות בצורה נוחה
- מיעוט תוצאות מסוגים מסויימים בגלל התסברות יחסית נמוכה שלהם.

הדרך להציג תוצאות הניסוי היא באופן הבא:

1. נסמן ערכים מקסימליים ומינימליים (הגדולים והמזעריים) ע"י  $x_{\max}, x_{\min}$ .

2. נחלק את המקטע  $[x_{\min}, x_{\max}]$  ל- $k$  bins (מקטעים) ברוחב אחיד של  $\Delta x$ .

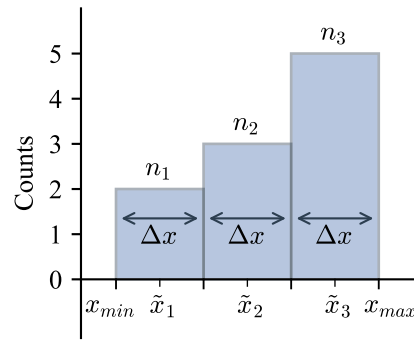
3. נסמן כל bin ע"י נקודת האמצע שלו,

$$(6.3) \quad \tilde{x}_i = x_{\min} + \left(i - \frac{1}{2}\right) \Delta x, \quad i = 1, \dots, k,$$

4. נסמן ב- $n_1$  את מספר התוצאות שהתקבלו ב-bin הראשון, ב- $n_2$  את מספר התוצאות ב-bin השני, וכך הלאה.

5. נשתמש בזוגות  $(\tilde{x}_i, n_i)$  עבור היסטוגרמות מסוג count או probability.

דוגמא להיסטוגרמה מסוג count מופיע באיור 6.4.



איור 6.4: היסטוגרמה מסוג count עבור מספר גדול של תוצאות ניסוי.

### 6.2.3 קירוב ל-PDF של משתנה אקראי רציף (\*)

**מטרה:** קירוב ל-PDF של משתנה רציף.

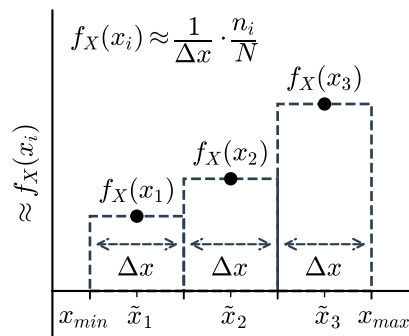
בנוסף לשיטה של קבלת ערכי  $n_i$  ע"י חלוקה ל-bins, ישנה דרך נוספת, שלישית, שהיא קירוב ל-PDF של מתנה רציף.

**קירוב ל-PDF של משתנה אקראי רציף ע"י היסטוגרמה:**

$$(6.4) \quad f_X(x_i) \approx \frac{n_i}{N} \cdot \frac{1}{\Delta x}, \quad i = 1, \dots, k$$

יש לשים לב לגורם נירמול של  $1/\Delta x$  המבדיל ממקרה של משתנה בדיד במשוואה (6.2).

דוגמא גרפית מופיע באיור 6.5.



איור 6.5: היסטוגרמה מסוג pdf עבור מספר גדול של תוצאות ניסוי.

### תאוריה והוכחה

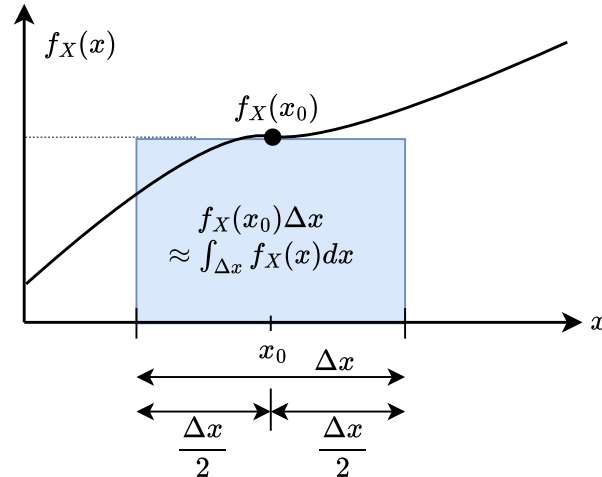
מבוסס על העיקרון של

$$(6.5) \quad \Pr(a < X \leq b) = \int_a^b f_X(x) dx,$$

ניתן לחשב בקירוב את ההסתברות ליפול בתוך bin ברוחב  $\Delta x$  שמרכזו ב- $x_0$ ,

$$(6.6) \quad \Pr\left(x_0 - \frac{\Delta x}{2} < X \leq x_0 + \frac{\Delta x}{2}\right) = \int_{\Delta x} f_X(x) dx \approx f_X(x_0) \Delta x.$$

המחשה של עיקרון זה מופיעה באיור 6.6. לאחר העברת אגפים,



איור 6.6: איור נלווה להצגת עיקרון היסטוגרמה: השטח המוצלל  $f_X(x_0) \Delta x$  מקרב את ההסתברות ליפול בתוך ה-bin.

$$(6.7) \quad f_X(x_0) \approx \Pr\left(x_0 - \frac{\Delta x}{2} < X \leq x_0 + \frac{\Delta x}{2}\right) \cdot \frac{1}{\Delta x}.$$

עבור bin  $i$  עם נקודה מרכזית  $\tilde{x}_i$ , ההסתברות להיות בתוך ה-bin היא בקירוב השכיחות היחסית (דומה מאוד למשוואה (6.2)),

$$(6.8) \quad \Pr\left(\tilde{x}_i - \frac{\Delta x}{2} < X \leq \tilde{x}_i + \frac{\Delta x}{2}\right) \approx \frac{n_i}{N}, \quad i = 1, \dots, k.$$

ההצבה של המשוואה (6.8) בתוך משוואה (6.7) נותנת את נוסחת הקירוב (6.4).

**דוגמה 6.2 (חישוב היסטוגרמת PDF):** נתונות תוצאות של ניסוי אקראי,

$$[16, 98, 96, 49, 81, 14, 43, 92, 80, 96].$$

מעוניינים להציג את ה-PDF של התוצאות בצורה גרפית ע"י היסטוגרמה בעלת  $k = 3$  bins. פתרון:

$$\begin{aligned} N &= 10 \\ x_{\min} &= 14, \quad x_{\max} = 98 \\ x_{\max} = x_{\min} + k \Delta x &= 98 \Rightarrow \Delta x = \frac{x_{\max} - x_{\min}}{k} = \frac{84}{3} = 28 \end{aligned}$$

מספר תוצאות בכל bin:

$$n_1 = 2, \quad \leftarrow \{14, 16\} \in [x_{\min}, x_{\min} + \Delta x] = [14, 42]$$

$$\begin{aligned} n_2 = 2, & \quad \leftarrow \{43, 49\} \in (x_{\min} + \Delta x, x_{\min} + 2\Delta x] & = (42, 70] \\ n_3 = 6, & \quad \leftarrow \{80, 81, 92, 96, 96, 98\} \in (x_{\max} - \Delta x, x_{\max}] & = (70, 98] \end{aligned}$$

כאשר  $n_3 = 10 - n_1 - n_2$ . מרכזים של כל bin:

$$\begin{aligned} \tilde{x}_1 &= x_{\min} + \frac{\Delta x}{2} = 28 \\ \tilde{x}_2 &= x_{\min} + \Delta x + \frac{\Delta x}{2} = 56 \\ \tilde{x}_3 &= x_{\max} - \frac{\Delta x}{2} = 84 \end{aligned}$$

לסיכום,

$$f_X(\tilde{x}_i) \approx \frac{n_i}{N \cdot \Delta x} = n_i \cdot \frac{1}{10 \cdot 28}.$$

## סיכום

דרכים להצגה של התוצאות הניסוי:

- גרף של ערכי  $n_i$  (בשפות תכנות בד"כ מסומן ב-count).
- גרף של ערכי  $n_i/N$  (בשפות תכנות בד"כ מסומן ב-probability).
- גרף של ערכי  $\frac{n_i}{N \Delta x}$  (בשפות תכנות בד"כ מסומן ב-pdf או density).

**דוגמה 6.3 (מדוע היסטוגרמת pdf מחלקת ברוחב ה-bin):** להיסטוגרמה של  $N$  תצפיות יש  $k$  bins ברוחב שווה  $\Delta x$ , עם מספרי תצפיות  $n_1, \dots, n_k$ .

(א) יש לחשב את השטח הכולל מתחת לעמודות עבור כל אחד משלושת הסוגים, count, probability ו-pdf. איזה מהם מנורמל כפי ש-PDF חייב להיות?

(ב) כעת ה-bins מאוחדים בזוגות, כך שהרוחב הופך ל- $2\Delta x$ . מה קורה לגובה העמודה בכל אחד משלושת הסוגים?

פתרון: (א) לכל עמודה רוחב  $\Delta x$ , ולכן השטח הוא הגובה כפול  $\Delta x$ , בסכום על כל ה-bins,

$$\begin{aligned} \text{count: } \sum_{i=1}^k n_i \Delta x &= N \Delta x \\ \text{probability: } \sum_{i=1}^k \frac{n_i}{N} \Delta x &= \Delta x \\ \text{pdf: } \sum_{i=1}^k \frac{n_i}{N \Delta x} \Delta x &= \frac{1}{N} \sum_{i=1}^k n_i = 1 \end{aligned}$$

רק השלישי שווה לאחד, והוא שווה לאחד עבור כל  $\Delta x$  וכל  $N$ . שני האחרים נושאים את יחידות הניסוי: השטח של count גדל עם גודל המדגם, והשטח של probability גדל עם רוחב ה-bin. החלוקה ב- $\Delta x$  ב-(6.4) אינה אפוא שינוי קנה מידה קוסמטי, אלא הצעד שהופך ספירה לצפיפות.

(ב) איחוד ה-bins  $i$  ו- $i+1$  נותן מספר תצפיות חדש  $n' = n_i + n_{i+1}$  ברוחב  $2\Delta x$ ,

$$\begin{aligned} \text{count: } n' &= n_i + n_{i+1} & \text{roughly doubles} \\ \text{probability: } \frac{n'}{N} &= \frac{n_i}{N} + \frac{n_{i+1}}{N} & \text{roughly doubles} \\ \text{pdf: } \frac{n'}{N \cdot 2\Delta x} &= \frac{1}{2} \left( \frac{n_i}{N \Delta x} + \frac{n_{i+1}}{N \Delta x} \right) & \text{the average of the two} \end{aligned}$$

כלומר, בשני הסוגים הראשונים הגובה בערך מוכפל, ובסוג pdf הוא הממוצע של שני הגבהים. שתי העמודות הראשונות זזות עם הספירה; השלישית נשארת באותו קנה מידה, והופכת לקריאה גסה יותר של אותה עקומה. זו הגזירה (6.7) מהצד השני: הצפיפות היא מה ששורד את הבחירה של  $\Delta x$ , ולכן רק את סוג pdf אפשר לשרטט על גבי  $f_X(x)$  תאורטי ללא שינוי קנה מידה.

## 6.2.4. השונות של מספר התצפיות ב-bin

**מטרה:** לכמת כמה זזה עמודה של היסטוגרמה כאשר הניסוי חוזר על עצמו, וכמה תצפיות נדרשות להיסטוגרמה קריאה.

כל מספר תצפיות  $n_i$  מתתי-הסעיפים הקודמים הוא עצמו גודל אקראי. חזרה על אותו ניסוי מציירת מחדש את כל התמונה, באותו אופן שבו ניסוי חוזר צייר מחדש את ה-sample mean (הממוצע המדגמי) בסעיף 5.1.

כל אחת מ- $N$  התצפיות נופלת ב-bin  $i$ , בהסתברות  $p_i$ , או שאינה נופלת בו, באופן בלתי תלוי בכל שאר התצפיות. לכן  $n_i$  סופר את ההצלחות של  $N$  ניסויי כן/לא בלתי תלויים, כלומר הוא משתנה אקראי בינומי,

$$(6.9) \quad n_i \sim \text{Bin}(N, p_i), \quad \mathbb{E}[n_i] = Np_i, \quad \text{Var}[n_i] = Np_i(1 - p_i),$$

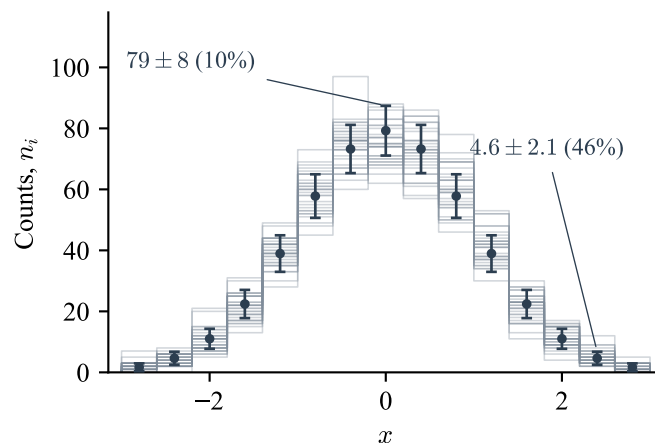
כאשר  $p_i$  היא ההסתברות האמיתית של ה-bin, בדיוק הגודל שהיחס  $n_i/N$  של (6.2) או של (6.8) נועד להעריך.

### סוג count

עמודה ממוקמת בגובה הצפוי שלה  $Np_i$ , עד כדי תנודה של

$$(6.10) \quad \sigma_{n_i} = \sqrt{Np_i(1 - p_i)},$$

שהיא סטיית התקן של הערכה, בדיוק כמו שגיאת התקן של סעיף 5.2.1. איור 6.7 מציג את שניהם: שלושים חזרות של ניסוי אחד, ואת הרצועה (6.10) שבה הן אמורות להישאר.



איור 6.7: שלושים חזרות של אותו ניסוי  $N = 500$  תצפיות של  $X \sim \mathcal{N}(0, 1)$ , (bins  $k = 15$ ), מונחות זו על זו באפור. הסמנים הם הגבהים הצפויים  $Np_i$  של (6.9), עם קווי השגיאה  $\pm \sigma_{n_i}$  שלהם (6.10). העמודות המרכזיות זזות בכעשירית מגובהן מחזרה לחזרה, ועמודות הזנבות בכמחצית.

מאחר ו- $p_i$  הוא הקטן ביותר בזנבות, עמודות הזנבות הן תמיד החלק הפחות אמין של היסטוגרמה. באיור 6.7 העמודה המרכזית היא  $79 \pm 8$  (10%), בעוד שהעמודה ב- $x \approx 2.4$  היא  $4.6 \pm 2.1$  (46%), והיא כמעט מוכפלת או נעלמת מחזרה לחזרה. בליטה בודדת שם אינה עדות ל-mode שני של התפלגות multimodal (רב-שיאית), ואותה זהירות חלה על כל bin המכיל קומץ תצפיות.

בפועל  $p_i$  אינו ידוע, ומחליפים אותו ביחס הנצפה  $n_i/N$ , באותו אופן שבו  $\sigma$  הוחלף ב- $s_x$  ב-(5.16),

$$(6.11) \quad \hat{\sigma}_{n_i} = \sqrt{n_i \left(1 - \frac{n_i}{N}\right)} \approx \sqrt{n_i},$$

כאשר הצעד האחרון מתקיים כאשר bin בודד מכיל חלק קטן מהנתונים, כלומר כאשר  $k$  גדול.

#### כלל השורש

לעמודה בגובה  $n_i$  יש קו שגיאה של  $\sqrt{n_i}$ , ולכן הדיוק היחסי שלה הוא

$$(6.12) \quad \frac{\sigma_{n_i}}{n_i} \approx \frac{1}{\sqrt{n_i}}.$$

הוא נקבע ע"י מספר התצפיות של אותה עמודה בלבד, ולא ע"י הסך  $N$ , ע"י רוחב ה-bin, או ע"י סוג ההיסטוגרמה המוצג.

ערכים מייצגים מופיעים בטבלה 6.1. עמודה הבנויה על 9 תצפיות ידועה בדיוק של שליש מגובהה, ושלוש ספרות של עמודה לעולם אינן זמינות: הן היו דורשות  $10^6$  תצפיות בתוך אותו bin בודד.

טבלה 6.1: דיוק יחסי (6.12) של עמודת היסטוגרמה, כפונקציה של מספר התצפיות שלה  $n_i$ .

| $n_i$  | שגיאה יחסית, $1/\sqrt{n_i}$ |
|--------|-----------------------------|
| 4      | 50%                         |
| 25     | 20%                         |
| 100    | 10%                         |
| 400    | 5%                          |
| $10^4$ | 1%                          |

**דוגמה 6.4 (אמינות העמודות של היסטוגרמה):** נמשיך עם  $N = 10$  התוצאות של הדוגמה הקודמת, ש- $k = 3$  ה-bins שלהן מכילים  $n_1 = 2$ ,  $n_2 = 2$  ו- $n_3 = 6$  תצפיות. עד כמה אמינות שלוש העמודות?

פתרון: כאשר  $p_i$  מוחלף ביחס הנצפה  $n_i/N$ , משוואה (6.10) נותנת

$$\begin{aligned} \sigma_{n_1} = \sigma_{n_2} &= \sqrt{10 \cdot 0.2 \cdot 0.8} = 1.26 \quad (63\%) \\ \sigma_{n_3} &= \sqrt{10 \cdot 0.6 \cdot 0.4} = 1.55 \quad (26\%) \end{aligned}$$

קריאת העמודות. העמודה הגבוהה ביותר היא  $6 \pm 1.5$ , ושתי האחרות הן  $2 \pm 1.3$  כל אחת. לא ניתן להבחין בין שני ה-bins הראשונים, ולכן לסדר ביניהם אין משמעות: חזרה על הניסוי עשויה בקלות להחליף ביניהם, ובאותה קלות לרוקן אחד מהם.

תכנון הניסוי. היפוך של (6.12) עבור שגיאה יחסית של 10% ב-bin השלישי דורש  $n_3 = 1/0.1^2 = 100$  תצפיות בתוך אותו bin, כלומר  $N = n_3/p_3 \approx 100/0.6 \approx 167$ . ספרה אמינה אחת בעמודה הגבוהה ביותר עולה פי 17 מהנתונים הנוכחיים.

**דוגמה 6.5 (עד כמה אפשר לסמוך על עמודה):** היסטוגרמה נבנתה מ- $N = 400$  תצפיות. Bin אחד מכיל  $n_i = 16$  מהן.

(א) בעזרת (6.10), יש לחשב את סטיית התקן של מספר התצפיות ב-bin, כאשר  $p_i$  נלקח מהיחס הנצפה.

(ב) יש להשוות לקירוב (6.11), ולומר מתי הקירוב הזה בטוח.

(ג) מהי השגיאה היחסית (6.12) של העמודה?

(ד) כמה תצפיות  $N$  נדרשות כדי להוריד את השגיאה היחסית ל-5%, כאשר רוחב ה-bin וההתפלגות אינם משתנים?

פתרון: (א) מספר התצפיות ב-bin מתפלג בינומית (6.9), כאשר  $p_i$  מוערך ע"י  $n_i/N = 16/400 = 0.04$ ,

$$\sigma_{n_i} = \sqrt{Np_i(1-p_i)} = \sqrt{400 \cdot 0.04 \cdot 0.96} = \sqrt{15.36} \approx 3.92.$$

(ב) הקירוב (6.11) משמיט את הגורם  $(1-p_i)$ ,

$$\sigma_{n_i} \approx \sqrt{n_i} = \sqrt{16} = 4,$$

בטווח של 2% מהערך המדויק. הגורם המושמט הוא  $\sqrt{1-p_i}$ , ולכן הקירוב בטוח כאשר bin מכיל חלק קטן מהנתונים, נניח  $p_i \lesssim 0.1$ . זה המקרה הרגיל: היסטוגרמה בעלת רזולוציה כלשהי מפזרת את התצפיות על פני bins רבים.

(ג)

$$\frac{\sigma_{n_i}}{n_i} \approx \frac{1}{\sqrt{n_i}} = \frac{1}{4} = 25\%.$$

עמודה בגובה 16 פירושה למעשה  $16 \pm 4$ . שתי עמודות שכנות שגובהן שונה ברבע אינן ניתנות להבחנה.

(ד) השגיאה היחסית תלויה במספר התצפיות ב-bin בלבד,

$$\frac{1}{\sqrt{n_i}} \leq 0.05 \quad \Rightarrow \quad n_i \geq 400.$$

כאשר רוחב ה-bin וההתפלגות אינם משתנים,  $p_i$  נשאר 0.04, ולכן

$$N = \frac{n_i}{p_i} = \frac{400}{0.04} = 10^4,$$

פי עשרים וחמישה מהמדגם המקורי עבור דיוק טוב פי חמישה, המחיר הרגיל של  $1/\sqrt{N}$ .

## בחירת רוחב ה-bin

הדיוק נקנה במספר התצפיות בכל bin, ומספר התצפיות בכל bin נשלט ע"י  $\Delta x$ : bin רחב פי שניים אוסף בערך פי שניים תצפיות. לכן הקטנת  $\Delta x$  בחצי, כלומר הכפלת  $k$ , מקטינה כל  $n_i$  בחצי ומכפילה את השגיאה היחסית (6.12) ב- $\sqrt{2}$ . עבור  $N$  קבוע, שתי התכונות הרצויות פועלות זו נגד זו:

- Bins רחבים ( $k$  קטן): עמודות אמינות, אך צורת ההתפלגות מוצגת ברזולוציה גסה.
- Bins צרים ( $k$  גדול): רזולוציה עדינה, אך עמודות רועשות כל כך שחלק מהצורה שהן מציגות הוא תוצר של הספירה.

יעד מעשי הוא לשמור את ה-bins שעליהם מתבסס הפירוש מעל כמה עשרות תצפיות,  $n_i \gtrsim 25$  בטבלה 6.1, ולא לקרוא שום צורה מתוך ה-bins שמתחת לכך.

## סיכום

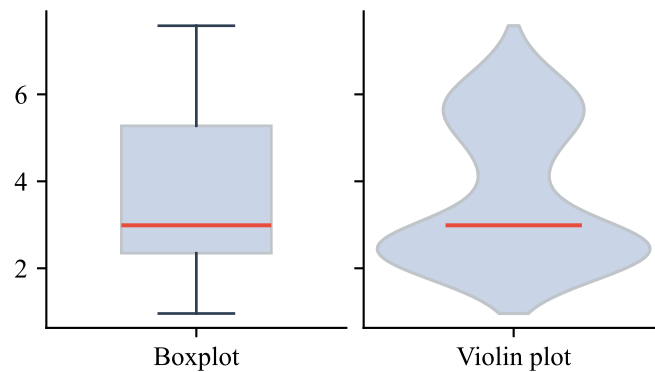
שלושת סוגי התצוגה של הסיכום הקודם הם אותו מספר תצפיות מוכפל בקבוע, ולכן הם נושאים אותה שגיאה ביחידות שונות:

- סוג count:  $n_i$ , עם סטיית תקן  $\sigma_{n_i}$ .
- סוג probability:  $\frac{n_i}{N}$ , עם סטיית תקן  $\frac{\sigma_{n_i}}{N} = \sqrt{\frac{p_i(1-p_i)}{N}}$ .
- סוג pdf:  $\frac{n_i}{N \Delta x}$ , עם סטיית תקן  $\frac{\sigma_{n_i}}{N \Delta x}$ .

חלוקת הגובה ב- $N$  או ב- $N \Delta x$  מחלקת את התנודה שלו באותו קבוע, ולכן השגיאה היחסית נשארת  $1/\sqrt{n_i}$  בשלושתם. הנרמול משנה את יחידות הגרף, ולעולם לא את האמינות של עמודה.

## 6.3 Violin Plot

**מטרה:** להציג גם את הסטטיסטיים המסכמים (boxplot) וגם את צורת התפלגות הנתונים (היסטוגרמה). Violin plot (דיאגרמת כינור) מרחיב את ה-boxplot ע"י הוספת היסטוגרמה (מוחלקת)<sup>1</sup> של הנתונים מכל צד. בדרך כלל משורטט בתוכו גם קו חציון.



איור 6.8: השוואה בין boxplot לבין violin plot עבור אותם נתונים bimodal (דו-שיאיים): ה-boxplot מסתיר את המבנה של שתי הקבוצות, בעוד שה-violin plot חושף אותו באמצעות צורתו.

איור 6.8 משווה בין שתי ההצגות עבור אותה קבוצת נתונים bimodal: ה-boxplot אינו מרמז כלל על שתי קבוצות, בעוד שה-violin plot מראה בבירור שתי בליטות.

## 6.4 שאלות

<sup>1</sup>החלקה זו נקראת kernel density estimation (שערוך צפיפות גרעין, KDE). הדיון בשיטה זו חורג ממסגרת פרק זה.

6.1 (חישוב הסתברות) נתונה פונקציית PDF של משתנה אקראי  $X$ , מהצורה

$$f_X(x) = \frac{1}{2}e^{2x}; \quad -\infty < x < \ln(a)$$

יש למצוא:

(א) ערך של המקדם  $a$

(ב)  $F_X(x)$

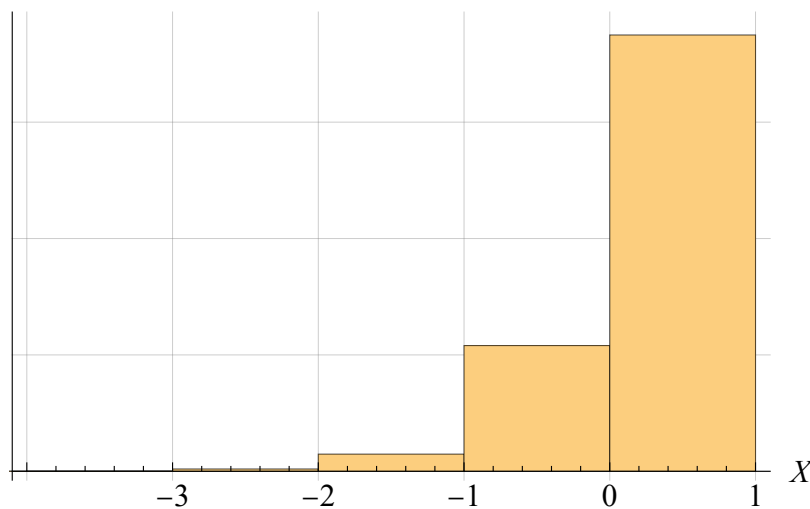
(ג)  $\mathbb{E}[X]$

(ד)  $\mathbb{E}[X^2]$

(ה)  $\Pr(X < 0)$

(ו) ערך חציון

(ז) נתונה היסטוגרמה (איור 6.9) עבור ערכים אקראיים המתפלגים לפי  $X$ . ציר  $x$  מחושב עם מרווחים של  $\Delta x = 1$ , כמתואר באיור 6.9.



איור 6.9: היסטוגרמה של משתנה  $X$ .

עבור ההסטוגרמה, יש לחשב את הערכים הבאים:

(1) מהו מספר הערכים התאורטי במרווח בין 0 ל-1 בתוך ההיסטוגרמה, מתוך סה"כ 100,000 הערכים (סוג count)?

(2) מהי הסתברות לערכים במרווח בין 0 ל-1, בתוך ההיסטוגרמה (סוג probability)?

(3) מהו הערך התאורטי של  $\hat{f}_X(0.5)$  עבור ההיסטוגרמה (סוג pdf)?

תזכורת:

$$\lim_{x \rightarrow -\infty} x^n e^x \rightarrow 0, \quad e^{\ln(a)} = a, \quad \ln(2) \approx 0.69314718$$

6.2 היסטוגרמה בוצע ניסוי אקראי שבו נדגמו  $N = 500$  ערכים של משתנה אקראי רציף  $X$ . לצורך ניתוח התוצאות, טווח הערכים חולק ל-bins בעלי רוחב אחיד של  $\Delta x = 4$ . נתמקד ב-bin ספציפי המוגדר על ידי התחום  $[10, 14]$ . ב-bin זה נספרו  $n_i = 80$  תוצאות מתוך הניסוי.

(א) חשב את גובה העמודה ( $y$ -axis) עבור ה-bin הנתון [10, 14] בשלושת סוגי ההיסטוגרמות הבאים:

(1) היסטוגרמה מסוג count.

(2) היסטוגרמה מסוג probability (שכיחות יחסית).

(3) היסטוגרמה מסוג pdf (צפיפות).

(ב) עבור היסטוגרמה מסוג pdf, מהו השטח של העמודה ב-bin ה"ל"? מהי המשמעות ההסתברותית של שטח זה?

(ג) נתון כי ב-bin אחר, [20, 24], גובה העמודה בהיסטוגרמה מסוג pdf הוא 0.03. כמה תוצאות ( $n_j$ ) נפלו ב-bin זה?

6.3 (סימולציה: סריקת רוחב ה-bin) יש לבחון באמצעות סימולציה כיצד רוחב ה-bin מחליף רזולוציה באמינות. יש להגדיל  $N = 1000$  תצפיות מ- $\mathcal{N}(0, 1)$  (randn) ולשרטט היסטוגרמות מסוג pdf של אותם נתונים עם  $k = 5$ , 20 ו-100 bins, ובכל אחת לשרטט על גבי ההיסטוגרמה את  $f_X(x)$  האמיתי.

(א) עבור כל  $k$ , יש לדווח את מספר התצפיות הטיפוסי  $n_i$  ב-bin שבמרכז ההתפלגות, ואת השגיאה היחסית  $1/\sqrt{n_i}$ .

(ב) איזו מהשלוש משחזרת הכי טוב את הצפיפות האמיתית, ואילו שתיים נכשלות, מסיבות הפוכות?

(ג) עבור איזה  $k$  כל ה-bins המרכזיים נשארים מעל כלל האצבע  $n_i \gtrsim 25$ ?

יש להשוות את הפיזור הנצפה של העמודות לאיור 6.7 ולטבלה 6.1.

## 7. ריבועים פחותים לינאריים חד-משתניים

מטרת הפרק היא להגדיר ולדון בשיטת הריבועים הפחותים (LS, least squares) הלינארית החד-משתנית עבור מודל לינארי.

שיטת ה-LS נקראת גם רגרסיה לינארית (linear regression). LS חד-משתני נקרא גם קו מגמה לינארי (linear trend-line).

### תוכן הפרק

|    |                                                |
|----|------------------------------------------------|
| 82 | 7.1 LS לינארי חד-משתני                         |
| 82 | 7.1.1 הגדרות                                   |
| 84 | 7.1.2 ממוצע ושונות                             |
| 84 | 7.1.3 מודל לינארי                              |
| 85 | 7.2 שונות משותפת ומקדם קורלציה                 |
| 86 | 7.2.1 שונות משותפת מדגמית                      |
| 86 | 7.2.2 מקדם קורלציה                             |
| 87 | 7.2.3 Normal Equations באמצעות גדלים סטטיסטיים |
| 88 | 7.3 Metrics ( $MSE, R^2$ )                     |
| 88 | 7.3.1 ( $R^2$ ) Coefficient of Determination   |
| 90 | 7.3.2 ה-MSE של LS לינארי                       |
| 92 | 7.4 רגרסיה לינארית מנורמלת (z-score) (*)       |
| 96 | 7.5 Spearman Rank Correlation                  |
| 96 | 7.6 קורלציה אינה סיבתיות                       |
| 97 | 7.7 שאלות                                      |

## 7.1 LS לינארי חד-משתני

### 7.1.1 הגדרות

**מערך נתונים:** ניסוי אקראי מייצר מערך נתונים (dataset)<sup>1</sup> של  $M$  זוגות תצפיות  $\{x_k, y_k\}_{k=1}^M$ .

**מודל:** המודל המונח בבסיס מערך הנתונים הוא

$$(7.1) \quad y = h(x) + \epsilon,$$

כאשר  $h(x)$  היא פונקציה דטרמיניסטית ו- $\epsilon$  הוא רעש בעל תוחלת אפס.

<sup>1</sup>אין לבלבל עם מסד נתונים (database), שהוא דבר אחר.

המטרה היא להציע ולהפעיל מודל,

$$(7.2) \quad \hat{y}_k = f(x_k)$$

שנותן את התוצאות המתאימות ביותר ביחס ל- $y_k$ .

האתגר: קיימות אפשרויות רבות עבור  $f(\cdot)$ .

**שגיאה:** השגיאה של המודל היא

$$(7.3) \quad e_k = y_k - \hat{y}_k$$

## Loss

**מטרה:** למצוא את הפרמטרים האופטימליים של המודל.

**Loss function:** (פונקציית הפסד), או cost function (פונקציית מחיר),  $\mathcal{L}(y_k, \hat{y}_k)$  היא מדד מרחק כלשהו בין  $y_k$  ל- $\hat{y}_k$ . הפרמטרים האופטימליים של המודל המופעל נמצאים ע"י מזעור ה-loss function.

דוגמאות נפוצות ל-loss functions:

- סכום ריבועי השגיאות (SSE, sum of squared errors),

$$(7.4) \quad \text{SSE} = \sum_{k=1}^M e_k^2,$$

- השגיאה הריבועית הממוצעת (MSE), או שונות השגיאה (ה-biased),

$$(7.5) \quad \text{MSE} = \frac{1}{M} \sum_{k=1}^M e_k^2 = \frac{1}{M} \text{SSE} = s_e^2$$

- ושורש השגיאה הריבועית הממוצעת (RMSE, root-mean-square error),

$$(7.6) \quad \text{RMSE} = \sqrt{\text{MSE}}.$$

מאחר וכפל בקבוע והפעלת טרנספורמציה מונוטונית אינם משנים את מיקום המינימום, לכל ה-loss functions האלו אותם פרמטרים אופטימליים:

$$(7.7) \quad \begin{aligned} w_0, w_1 &= \arg \min_{w_0, w_1} \text{SSE} \\ &= \arg \min_{w_0, w_1} \text{MSE} \\ &= \arg \min_{w_0, w_1} \text{RMSE}. \end{aligned}$$

**Metric:** Metric (מדד ביצועים) הוא מדד כמותי המשמש להערכת טיב פעולתו של המודל.

לדוגמה, את ה-RMSE קל הרבה יותר לפרש מאשר את ה-MSE או את ה-SSE, מאחר ותחום הערכים ויחידות המדידה שלו זהים לאלו של מערך הנתונים המקורי.

- קל לבלבל בין metrics ל-loss functions, אך הם משרתים מטרות שונות:
- ה-loss function משמשת את האלגוריתם במהלך האימון, כדי ללמוד את הפרמטרים האופטימליים של המודל.
  - ה-metric משמש את בני האדם להערכת ביצועי המודל.

## סיכום

1. בחירת מודל  $f(\cdot)$ .
2. מציאת הפרמטרים של המודל ע"י מזעור ה-loss function,  $\mathcal{L}$ .
3. שימוש ב-metrics להערכת ביצועי המודל.

## 7.1.2. ממוצע ושונות

**מטרה:** לתת פירוש לממוצע ולשונות בהקשר של LS. מקרה פרטי של המודל הלינארי הוא מודל מהצורה

$$(7.8) \quad \hat{y} = w_0$$

ה-loss function מסוג MSE (7.5) המתאימה היא

$$(7.9) \quad \mathcal{L}(w_0) = \frac{1}{M} \sum_{k=1}^M (y_k - w_0)^2,$$

והמינימום שלה מתקבל בממוצע של ערכי  $y_k$ ,

$$(7.10) \quad w_0 = \frac{1}{M} \sum_{k=1}^M y_k = \bar{y}$$

הוכחה. כדי למצוא את המינימום,

$$(7.11) \quad \frac{d}{dw_0} \mathcal{L}(w_0) = 2 \frac{1}{M} \sum_{k=1}^M (y_k - w_0)(-1) = 0$$

ה-MSE המתאים של המודל (הצבת (7.10) ב-(7.9)) הוא

$$(7.12) \quad \text{MSE} = \frac{1}{M} \sum_{k=1}^M (y_k - \bar{y})^2 = s_y^2$$

שהוא ה-sample variance (שונות מדגמית) של ערכי  $y_k$ .

לסיכום, החזאי הטוב ביותר הוא ה-sample mean (ממוצע מדגמי),  $\hat{y} = \bar{y}$ , עם  $\text{MSE} = s_y^2$ .

## 7.1.3. מודל לינארי

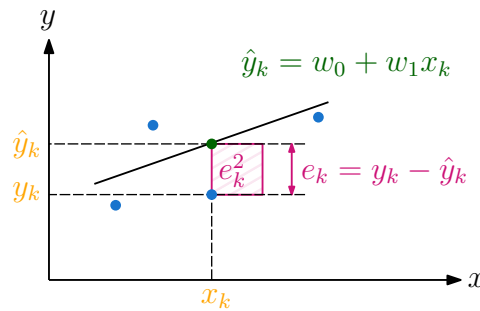
דוגמה למודל לינארי היא

$$(7.13) \quad \hat{y}_k = f(x_k; w_0, w_1) = w_0 + w_1 x_k,$$

כאשר  $w_0$  ו- $w_1$  הם הפרמטרים של המודל.

השגיאה של המודל הלינארי היא

$$(7.14) \quad e_k = y_k - w_0 - w_1 x_k.$$



איור 7.1: המחשה של גרסיה לינארית. המטרה היא מזעור ה-SSE, שהוא השטח הכולל  $\sum_k e_k^2$  של הריבועים.

מזעור ה-loss function מסוג SSE (משוואה (7.4) ואיור 7.1),

$$(7.15) \quad \mathcal{L}(w_0, w_1) = \sum_{k=1}^M (y_k - w_0 - w_1 x_k)^2,$$

מתבצע ע"י השוואת הנגזרות החלקיות של  $\mathcal{L}$  לאפס:

$$(7.16) \quad \begin{cases} \frac{\partial}{\partial w_0} \mathcal{L}(w_0, w_1) = 0 \\ \frac{\partial}{\partial w_1} \mathcal{L}(w_0, w_1) = 0 \end{cases}$$

מה שנותן

$$(7.17) \quad \begin{cases} 2 \sum_{k=1}^M (y_k - w_0 - w_1 x_k) \cdot (-1) = 0 \\ 2 \sum_{k=1}^M (y_k - w_0 - w_1 x_k) \cdot (-x_k) = 0. \end{cases}$$

לבסוף, לאחר מעט אלגברה:

$$(7.18) \quad \begin{cases} w_0 M + w_1 \sum_{k=1}^M x_k = \sum_{k=1}^M y_k, \\ w_0 \sum_{k=1}^M x_k + w_1 \sum_{k=1}^M x_k^2 = \sum_{k=1}^M x_k y_k. \end{cases}$$

מערכת משוואות זו נקראת normal equations (המשוואות הנורמליות).

## 7.2. שונות משותפת ומקדם קורלציה

**מטרה:** לרשום מחדש את ה-normal equations (7.18) באמצעות גדלים סטטיסטיים.

<sup>2</sup>ביטוי שונות זה נקרא biased (מוטה), והוא משמש לאורך כל הפרק (סעיף 5.1).

## 7.2.1. שונות משותפת מדגמית

**מטרה:** לכמת את הקשר הלינארי בין  $x_k$  ל- $y_k$ .

השונות המשותפת המדגמית (biased-ה, sample covariance) בין  $x_1, \dots, x_M$  לבין  $y_1, \dots, y_M$  נתונה ע"י

$$(7.19) \quad s_{xy} = \frac{1}{M} \sum_{k=1}^M (x_k - \bar{x})(y_k - \bar{y})$$

הסימן והגודל של השונות המשותפת המדגמית  $s_{xy}$  מצביעים על הכיוון ועל העוצמה של הקשר הלינארי בין  $x$  ל- $y$ :

- $s_{xy} > 0$ : כאשר  $x$  גדל,  $y$  נוטה לגדול (רוב המכפלות  $(x_i - \bar{x})(y_i - \bar{y})$  חיוביות).
- $s_{xy} < 0$ : כאשר  $x$  גדל,  $y$  נוטה לקטון.
- $s_{xy} \approx 0$ : אין קשר לינארי.

## 7.2.2. מקדם קורלציה

**מטרה:** שונות משותפת מדגמית מנורמלת.

מקדם הקורלציה המדגמי של Pearson (sample Pearson correlation coefficient) בין  $x$  ל- $y$  הוא השונות המשותפת המנורמלת (חסרת היחידות):

$$(7.20) \quad r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{\sum_{k=1}^M (x_k - \bar{x})(y_k - \bar{y})}{\sqrt{\sum_{k=1}^M (x_k - \bar{x})^2} \sqrt{\sum_{k=1}^M (y_k - \bar{y})^2}}$$

### פירוש

- תחום ערכים:  $-1 \leq r_{xy} \leq +1$ .
- $r_{xy} > 0$ : קשר לינארי חיובי (כאשר  $x$  גדל,  $y$  נוטה לגדול).
- $r_{xy} < 0$ : קשר לינארי שלילי (כאשר  $x$  גדל,  $y$  נוטה לקטון).
- $r_{xy} = 0$ : אין קשר לינארי (ייתכן עדיין קשר לא לינארי).
- $|r_{xy}| = 1$ : קשר לינארי מושלם ( $y_k = a + bx_k$  לכל  $k$ ).

שימו לב שאין הבדל בין ההגדרה ה-biased להגדרה ה-unbiased עבור  $r_{xy}$ , מאחר והמקדמים המתאימים מצטמצמים.

**דוגמה 7.1:** באיור 7.2 נתונות תוצאות של 4 ניסויים דו-ממדיים שונים. יש למיין מהגבוה לנמוך את הסדר של הניסויים ע"פ ערך מוחלט של המקדם הקורלציה,  $|r_{xy}|$ , ללא חישוב. יש להסביר/לנמק את תשובתך.

פתרון: באופן כללי,

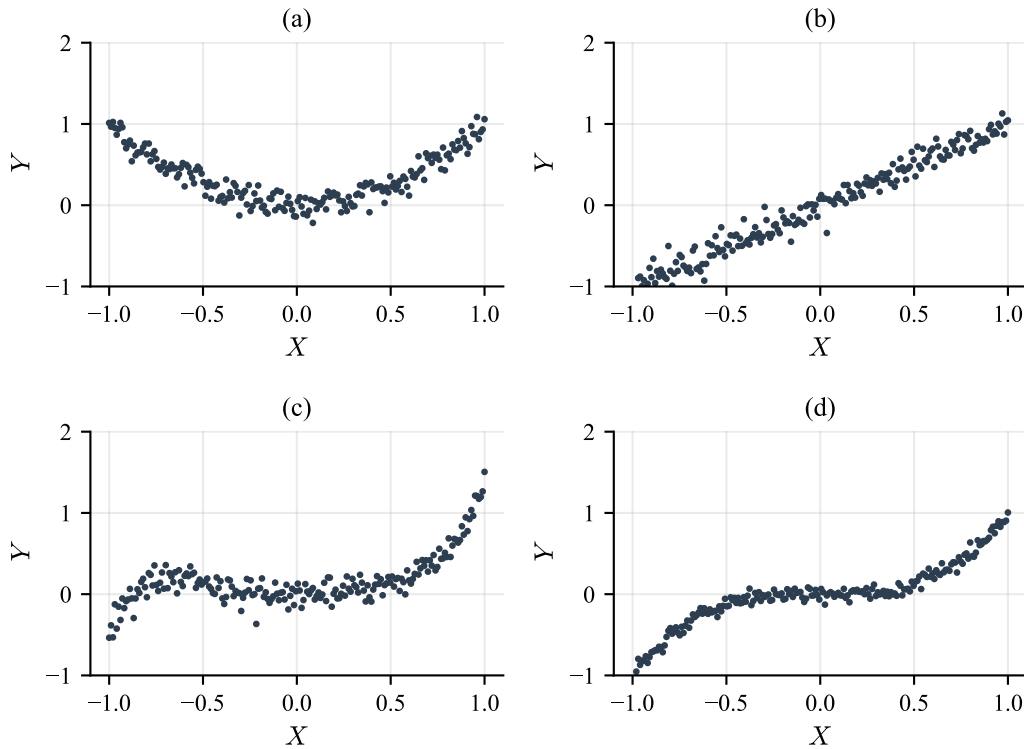
- עבור קו "ישר" יותר  $|r_{xy}|$  יותר קרוב ל-1.
- עבור קו ישר, עבור פיזור "קטן" יותר  $|r_{xy}|$  יותר קרוב ל-1.

מכאן הסדר (b), (d), (c), (a). מקדמי קורלציה לא ניתן לחשב ישירות מתוך השרטוט. הערכים הם:

- (b)  $r_{xy} = 0.985$ , קו ישר כמעט מושלם;

- (d)  $r_{xy} = 0.908$ , מונוטוני אך מעוקל בבירור;
- (c)  $r_{xy} = 0.615$ , מונוטוני רק בחלק מהתחום;
- (a)  $r_{xy} = -0.019$ .

הגרף (a) הוא החשוב מכולם. הנקודות שלו נמצאות על פרבולה, כך ש- $y$  הוא פונקציה מדויקת של  $x$  והשניים קשורים זה בזה ככל שניתן, ובכל זאת  $r_{xy}$  כמעט אפס. הפרבולה סימטרית סביב  $x = 0$ : היא יורדת עבור  $x$  שלילי ועולה עבור  $x$  חיובי, ושני החצאים תורמים מכפלות  $(x_k - \bar{x})(y_k - \bar{y})$  בעלות סימנים מנוגדים המבטלות זו את זו. מקדם הקורלציה מודד את החלק הליניארי של הקשר, ותו לא.



איור 7.2: יש לרשום את הסדר של הניסויים ע"פ ערך מוחלט של המקדם הקורלציה, מהגבוה לנמוך בתהאם לדוגמה 7.1.

### 7.2.3 Normal Equations באמצעות גדלים סטטיסטיים

נוח מבחינה נומרית לרשום מחדש את ה-normal equations (7.18) באמצעות מומנטים מדגמיים,

$$\begin{aligned}
 w_1 &= \frac{s_{xy}}{s_x^2} = r_{xy} \frac{s_y}{s_x} \\
 &= \frac{\sum_k (x_k - \bar{x})(y_k - \bar{y})}{\sum_k (x_k - \bar{x})^2}, \\
 (7.21) \quad w_0 &= \bar{y} - w_1 \bar{x} \\
 &= \bar{y} - \frac{s_{xy}}{s_x^2} \bar{x},
 \end{aligned}$$

ובצורת חיזוי,

$$(7.22) \quad \hat{y} = \bar{y} + w_1 (x - \bar{x}).$$

שימו לב שביטויים אלו אינם תלויים בשאלה אם עבור  $s_{xy}$  ו- $s_x$  משתמשים בהגדרות ה-biased או ה-unbiased, מאחר וגורמי הנרמול המתאימים מצטמצמים.

### הערות.

- פתרון תקף דורש  $s_x \neq 0$ , כלומר פיזור בערכי  $x_i \neq \text{const}$ .
- אם שני המשתנים ממורכזים ( $\bar{x} = \bar{y} = 0$ ), משוואה (7.21) מצטמצמת ל-

$$(7.23) \quad \begin{aligned} w_0 &= 0, \\ w_1 &= \frac{\sum_k x_k y_k}{\sum_k x_k^2} \end{aligned}$$

## 7.3 Metrics ( $MSE, R^2$ )

ביצועי המודל מכומתים ע"י metric ביצועים, שאינו בהכרח זהה ל-loss function.

### 7.3.1 Coefficient of Determination ( $R^2$ )

כדי להדגיש את ההבדל בין loss ל-metrics, מובאת להלן דוגמה ל-metric של רגרסיה לינארית. ה-coefficient of determination (מקדם הקביעה), המסומן  $R^2$  (R-squared), מבוסס על הקשר

$$(7.24) \quad \underbrace{\sum_{k=1}^M (y_k - \bar{y})^2}_{\text{SST}} = \underbrace{\sum_{k=1}^M (\hat{y}_k - \bar{y})^2}_{\text{SSR}} + \underbrace{\sum_{k=1}^M e_k^2}_{\text{SSE}},$$

כאשר:

- SST: סכום הריבועים הכולל (total sum of squares).
- SSR: סכום הריבועים המוסבר ע"י הרגרסיה (sum of squares due to regression).
- SSE: סכום ריבועי השגיאות (residual sum of squares, סכום ריבועי השאריות).

ה-metric  $R^2$  מוגדר כ-

$$(7.25) \quad R^2 = \frac{\text{SSR}}{\text{SST}} = 1 - \frac{\text{SSE}}{\text{SST}},$$

והוא מדד חסר יחידות ל-goodness-of-fit (טיב ההתאמה), בעל תחום אינטואיטיבי  $[0, 1]$ :

- $R^2 = 1$ : התאמה מושלמת.
- $R^2 = 0$ : המודל אינו טוב יותר מחיזוי ע"י הממוצע,  $\hat{y}_i = \bar{y}$ .
- $R^2 < 0$ : המודל גרוע יותר מהממוצע (אפשרי עבור מודלים של למידת מכונה על נתוני בדיקה).

הוכחה. כדי להוכיח את (7.24), נשים לב ש-

$$y_k - \bar{y} = (\hat{y}_k - \bar{y}) + (y_k - \hat{y}_k) = (\hat{y}_k - \bar{y}) + e_k.$$

לכן,

$$\begin{aligned}\sum_{k=1}^M (y_k - \bar{y})^2 &= \sum_{k=1}^M \left[ (\hat{y}_k - \bar{y}) + e_k \right]^2 \\ &= \sum_{k=1}^M (\hat{y}_k - \bar{y})^2 + \sum_{k=1}^M e_k^2 + 2 \sum_{k=1}^M (\hat{y}_k - \bar{y}) e_k.\end{aligned}$$

נותר להראות שהאיבר המעורב מתאפס:

$$\sum_{k=1}^M (\hat{y}_k - \bar{y}) e_k = \sum_{k=1}^M \hat{y}_k e_k - \bar{y} \sum_{k=1}^M e_k = 0 - 0 = 0,$$

מאחר וב-LS מתקיים  $\sum_{k=1}^M e_k = 0$  וגם  $\sum_{k=1}^M \hat{y}_k e_k = 0$ .

$R^2$  של **LS לינארי חד-משתני** במקרה של רגרסיה לינארית חד-משתנית,  $R^2$  הוא החלק של השונות המדגמית של  $y$  המוסבר ע"י ההתאמה הלינארית,

$$(7.26) \quad R^2 = r_{xy}^2.$$

הוכחה. נתחיל מ- $\hat{y}_k = w_0 + w_1 x_k$ :

$$\begin{aligned}(7.27) \quad \text{SSR} &= \sum_{k=1}^M (\hat{y}_k - \bar{y})^2 \\ &= \sum_{k=1}^M (w_0 + w_1 x_k - \bar{y})^2 \\ &= \sum_{k=1}^M (\bar{y} - w_1 \bar{x} + w_1 x_k - \bar{y})^2 \\ &= w_1^2 \sum_{k=1}^M (x_k - \bar{x})^2 \\ &= w_1^2 M s_x^2 = M \left( \frac{s_{xy}}{s_x^2} \right)^2 s_x^2 = M \frac{s_{xy}^2}{s_x^2}, \\ \text{SST} &= \sum_{k=1}^M (y_k - \bar{y})^2 = M s_y^2\end{aligned}$$

$$\text{לכן } \text{SSR}/\text{SST} = s_{xy}^2 / (s_x^2 s_y^2) = r_{xy}^2.$$

- הזהות היא חד-משתנית. ההוכחה נשענת על  $\hat{y}_k - \bar{y} = w_1(x_k - \bar{x})$ , כלומר על עמודה ממורכזת אחת. כאשר יש יותר מ-1 feature אחד, החיזוי כבר אינו פרופורציוני לעמודה אחת, ו- $R^2$  אינו הריבוע של אף מקדם קורלציה בודד.
- $R^2$  הוא ריבוע, ולכן הוא מאבד את הסימן: קו עולה מושלם וקו יורד מושלם מקבלים שניהם  $R^2 = 1$ . כיוון הקשר נישא ע"י  $r_{xy}$ , או באופן שקול ע"י הסימן של  $w_1$ , שע"פ (7.21) הוא הסימן של  $s_{xy}$ . דיווח של  $R^2$  בלבד אומר כמה מהשונות מוסבר, ולעולם לא באיזה כיוון.

## 7.3.2. ה-MSE של LS לינארי

ה-MSE הכללי (ה-biased<sup>3</sup>) נתון ע"י

$$(7.28) \quad \text{MSE} = s_y^2 (1 - r_{xy}^2) = s_e^2$$

הוכחה. ע"פ (7.24):

$$(7.29) \quad \begin{aligned} \text{MSE} &= \frac{1}{M} \text{SSE} = \frac{1}{M} (\text{SST} - \text{SSR}) \\ &= \frac{\text{SST}}{M} \left( 1 - \frac{\text{SSR}}{\text{SST}} \right) \\ &= s_y^2 (1 - R^2) \\ &= s_y^2 (1 - r_{xy}^2) \end{aligned}$$

דוגמה 7.2 (התאמה חד-משתנית מלאה בחישוב ידני): נתון מערך נתונים של  $M = 5$  תצפיות,

| 5 | 4 | 3 | 2 | 1 | $k$   |
|---|---|---|---|---|-------|
| 5 | 4 | 3 | 2 | 1 | $x_k$ |
| 9 | 6 | 5 | 3 | 2 | $y_k$ |

(א) יש לחשב את  $\bar{x}$ ,  $\bar{y}$ ,  $s_x^2$ ,  $s_y^2$  ואת השונות המשותפת  $s_{xy}$ .(ב) יש למצוא את  $w_1$  ואת  $w_0$  ע"פ (7.21).

(ג) יש לרשום את החיזויים ואת השגיאות, ולחשב ישירות את SSE, MSE ו-RMSE.

(ד) יש לחשב את  $r_{xy}$  ואת  $R^2$ , לבדוק ש- $R^2 = r_{xy}^2$ , ולבדוק את MSE מול  $s_y^2(1 - r_{xy}^2)$ .(ה) מהו ה-MSE של המודל הקבוע הטוב ביותר (7.10),  $\hat{y} = \bar{y}$ , ופי כמה השיפוע משפר אותו?

(ו) ממזערים את SSE, אך מדווחים את RMSE. מדוע?

פתרון: ממוצעים, שונות, שונות משותפת. שני הממוצעים שלמים, וזה שומר על כל ההמשך מדויק.

$$\begin{aligned} \bar{x} &= \frac{1+2+3+4+5}{5} = 3, & \bar{y} &= \frac{2+3+5+6+9}{5} = \frac{25}{5} = 5 \\ x_k - \bar{x} &= -2, -1, 0, 1, 2, & y_k - \bar{y} &= -3, -2, 0, 1, 4 \end{aligned}$$

$$\begin{aligned} s_x^2 &= \frac{1}{M} \sum_k (x_k - \bar{x})^2 = \frac{4+1+0+1+4}{5} = \frac{10}{5} = 2 \\ s_y^2 &= \frac{1}{M} \sum_k (y_k - \bar{y})^2 = \frac{9+4+0+1+16}{5} = \frac{30}{5} = 6 \\ s_{xy} &= \frac{1}{M} \sum_k (x_k - \bar{x})(y_k - \bar{y}) = \frac{6+2+0+1+8}{5} = \frac{17}{5} = 3.4 \end{aligned}$$

המקדמים.

$$w_1 = \frac{s_{xy}}{s_x^2} = \frac{3.4}{2} = 1.7$$

<sup>3</sup>המעריך ה-unbiased (חסר ההטיה) של  $s_e^2$  דורש גורם  $1/(M-2)$  במקום  $1/M$ .

$$w_0 = \bar{y} - w_1 \bar{x} = 5 - 1.7 \cdot 3 = -0.1$$

חיזויים ושגיאות.

| 5   | 4    | 3   | 2    | 1   | $k$         |
|-----|------|-----|------|-----|-------------|
| 8.4 | 6.7  | 5.0 | 3.3  | 1.6 | $\hat{y}_k$ |
| 0.6 | -0.7 | 0   | -0.3 | 0.4 | $e_k$       |

סכום השגיאות הוא אפס,  $0.4 - 0.3 + 0 - 0.7 + 0.6 = 0$ , וזו הבדיקה האריתמטית של  $w_0$ .

$$\text{SSE} = 0.16 + 0.09 + 0 + 0.49 + 0.36 = 1.10$$

$$\text{MSE} = \frac{\text{SSE}}{M} = \frac{1.10}{5} = 0.22$$

$$\text{RMSE} = \sqrt{0.22} \approx 0.469$$

מקדם קורלציה ו- $R^2$ .

$$r_{xy} = \frac{s_{xy}}{s_x s_y} = \frac{3.4}{\sqrt{2}\sqrt{6}} = \frac{3.4}{\sqrt{12}} \approx 0.9815$$

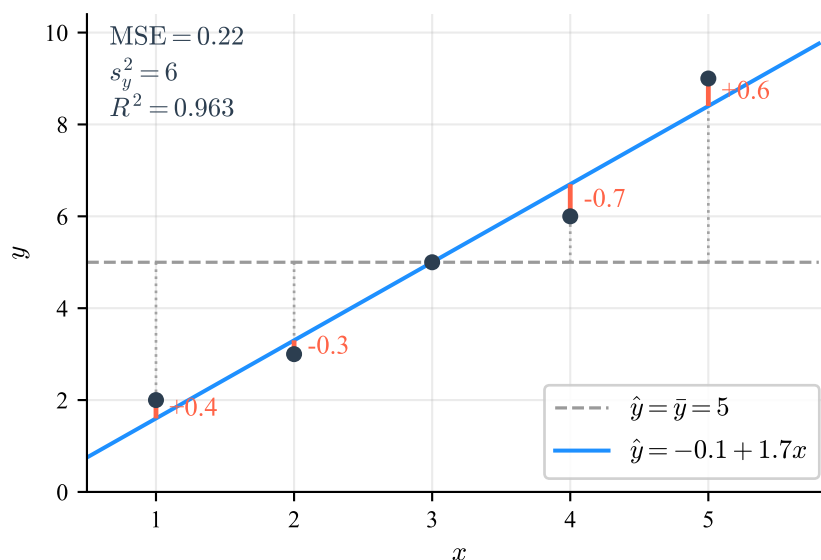
$$R^2 = 1 - \frac{\text{SSE}}{\text{SST}} = 1 - \frac{1.10}{30} \approx 0.9633$$

$$r_{xy}^2 = \frac{s_{xy}^2}{s_x^2 s_y^2} = \frac{3.4^2}{2 \cdot 6} = \frac{11.56}{12} = 0.9633 = R^2,$$

כפי שדורשת הזהות  $R^2 = r_{xy}^2$ . גם הבדיקה של השגיאה נסגרת,

$$s_y^2 (1 - r_{xy}^2) = 6 (1 - 0.9633) = 0.22 = \text{MSE},$$

בהתאמה לערך שהתקבל מהשגיאות אחת-אחת.



איור 7.3: ההתאמה של דוגמה 7.2 וקו הבסיס שעליה לנצח. הקטעים הרציפים הם השגיאות של הקו, והמנוקדים הם השגיאות של המודל הקבוע הטוב ביותר  $\hat{y} = \bar{y} = 5$ . סכום כל אחת מהקבוצות הוא אפס; הפקטור 27 בין  $\text{MSE} = 0.22$  לבין  $s_y^2 = 6$  הוא ההבדל באורכיהן, ו- $R^2 = 0.963$  הוא אותה השוואה כשבר.

קו הבסיס הקבוע. המודל הטוב ביותר עם מקדם יחיד הוא  $\hat{y} = \bar{y}$ , והשגיאה שלו היא השונות של המטרה עצמה,

$$\text{MSE}_{\text{const}} = s_y^2 = 6 \quad \text{vs.} \quad \text{MSE} = 0.22,$$

פקטור של כ-27. זה בדיוק מה ש- $R^2$  מודד: המודל הקבוע מקבל  $R^2 = 0$  מעצם הגדרתו, והשיפוע לוקח 96.3% מהשונות הזמינה.

Loss מיל metric. SSE הוא ה-loss: אותו ממזערות המשוואות הנורמליות (7.18). אבל הוא ביחידות של  $y^2$  וגדל עם  $M$ , ולכן אי אפשר להשוות באמצעותו שני מערכי נתונים בגדלים שונים.  $\text{RMSE} = \sqrt{\text{MSE}}$  מחלק בגודל המדגם ומוציא שורש, וכך חוזר ליחידות של  $y$  עצמו: כאן 0.469, החטאה טיפוסית של כחצי יחידה מול ערכי מטרה בין 2 ל-9. מזעור של SSE, MSE או RMSE נותן את אותו  $w$  ע"פ (7.7); הבחירה ביניהם היא בחירה של אופן הדיווח, לא של ההתאמה.

## 7.4. גרסיה לינארית מנורמלת (z-score) (\*)

מטרת הסעיף היא לתת פירוש אינטואיטיבי למקדם הקורלציה.

**שינוי קנה מידה** כדי שהשיפוע יהיה ניתן לפירוש ישיר, נעביר את שני המשתנים לתוחלת אפס ולשונות יחידה, כלומר ל-z-score (ציון תקן):

$$(7.30) \quad z_i = \frac{x_i - \bar{x}}{s_x}, \quad t_i = \frac{y_i - \bar{y}}{s_y}, \quad i = 1, \dots, M$$

התכונות של  $z_i$  ו- $t_i$  הן

$$\begin{aligned} \bar{z} &= \frac{1}{M} \sum z_i = 0, \\ \bar{t} &= \frac{1}{M} \sum t_i = 0, \\ s_z &= s_t = 1. \end{aligned}$$

הוכחה. נשתמש בהגדרות השונות ה-biased<sup>4</sup>:

$$(7.31) \quad s_x^2 = \frac{1}{M} \sum_{k=1}^M (x_k - \bar{x})^2, \quad s_y^2 = \frac{1}{M} \sum_{k=1}^M (y_k - \bar{y})^2.$$

עבור הממוצעים,

$$\begin{aligned} \sum_{k=1}^M z_k &= \frac{1}{s_x} \sum_{k=1}^M (x_k - \bar{x}) \\ (7.32) \quad &= \frac{1}{s_x} \left( \sum_{k=1}^M x_k - M\bar{x} \right) \\ &= \frac{1}{s_x} (M\bar{x} - M\bar{x}) = 0 \end{aligned}$$

<sup>4</sup>עבור חישוב unbiased יש להשתמש בגורם  $M - 1$ .

ולכן  $\bar{z} = 0$ , ובאופן דומה  $\bar{t} = 0$ . עבור השונויות:

$$(7.33) \quad s_z^2 = \frac{1}{M} \sum_{k=1}^M z_k^2 = \frac{1}{M} \sum_{k=1}^M \frac{(x_k - \bar{x})^2}{s_x^2} = \frac{1}{s_x^2} \left[ \frac{1}{M} \sum_{k=1}^M (x_k - \bar{x})^2 \right] = \frac{s_x^2}{s_x^2} = 1.$$

ההוכחה עבור  $s_t^2 = 1$  זהה.

### Normal Equations המודל הלינארי במרחב המנורמל הוא

$$(7.34) \quad \hat{t}_i = r_{xy} z_i,$$

כאשר

$$(7.35) \quad r_{xy} = \frac{s_{xy}}{s_x s_y}$$

הוא מקדם הקורלציה.

הוכחה. הפעלת (7.23) על

$$(7.36) \quad \hat{t}_i = w_0^* + w_1^* z_i$$

נותנת

$$(7.37) \quad w_0^* = 0,$$

$$(7.38) \quad w_1^* = \frac{s_{zt}}{s_z^2} = \frac{1}{M} \sum_{i=1}^M z_i t_i = \frac{1}{M} \sum_{i=1}^M \left( \frac{x_i - \bar{x}}{s_x} \right) \left( \frac{y_i - \bar{y}}{s_y} \right) = r_{xy}$$

### שגיאת הרגרסיה כפונקציה של $r_{xy}$

עם המודל המנורמל האופטימלי  $\hat{t} = r_{xy} z$ , השאריות הן

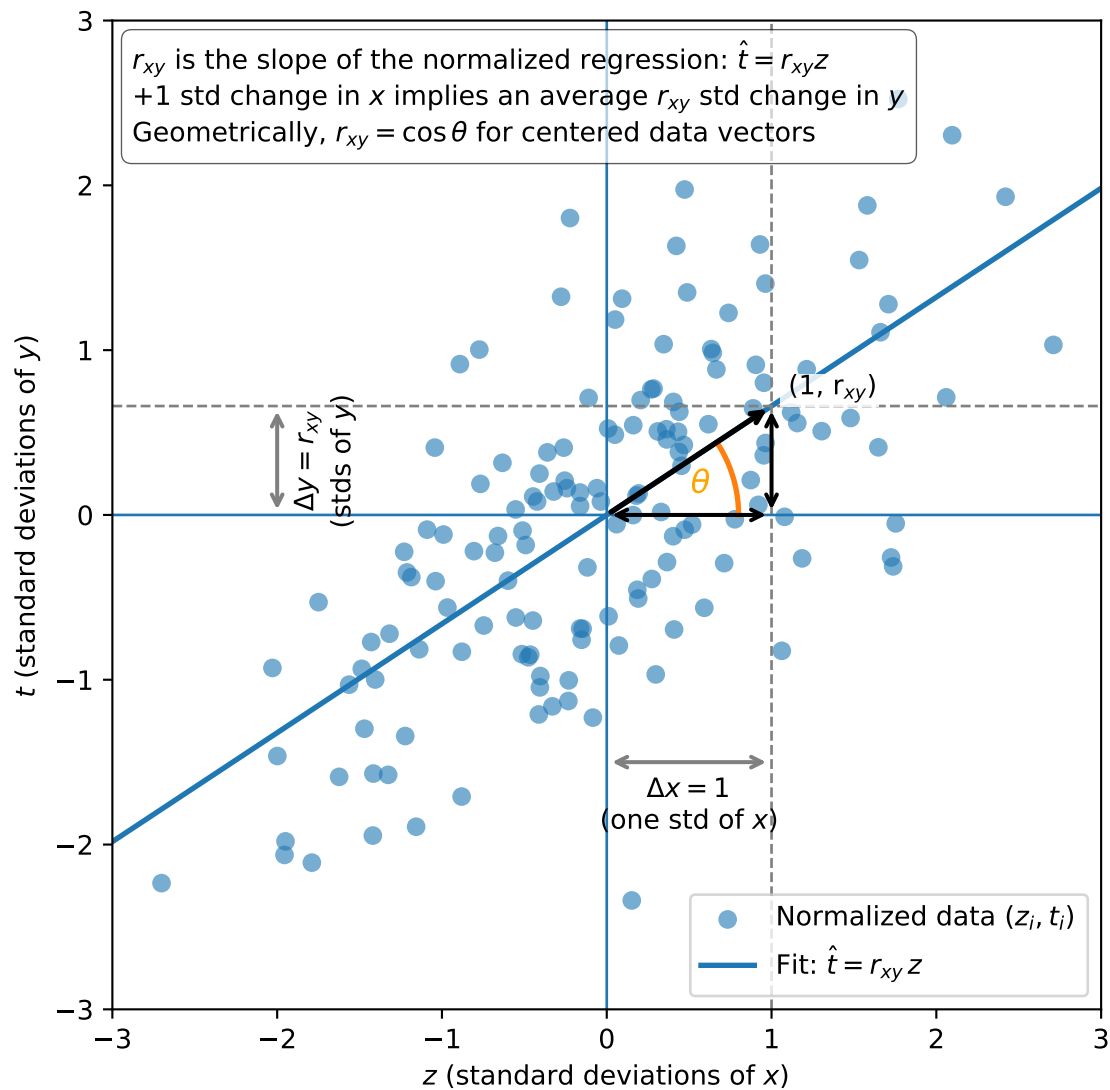
$$(7.39) \quad e_i = t_i - \hat{t}_i = t_i - r_{xy} z_i$$

עם

$$(7.40) \quad \text{MSE} = 1 - r_{xy}^2 = s_e^2.$$

הוכחה.

$$(7.41) \quad \begin{aligned} s_e^2 &= \frac{1}{M} \sum_{i=1}^M e_i^2 = \frac{1}{M} \sum_{i=1}^M (t_i - r_{xy} z_i)^2 \\ &= \frac{1}{M} \sum_{i=1}^M t_i^2 - 2r_{xy} \frac{1}{M} \sum_{i=1}^M t_i z_i + r_{xy}^2 \frac{1}{M} \sum_{i=1}^M z_i^2 \\ &= s_t^2 - 2r_{xy} (r_{xy}) + r_{xy}^2 s_z^2 \\ &= 1 - r_{xy}^2 \end{aligned}$$



איור 7.4: פירוש גאומטרי של  $r_{xy}$  בגרסיה לינארית מנורמלת.

### פירוש גאומטרי של $r_{xy}$

- יחידות הצירים הן סטיות תקן מהראשית, מאחר ו- $t_i$  ו- $z_i$  ממורכזים ומתוקננים.
- במרחב המנורמל,  $r_{xy}$  הוא שיפוע הרגרסיה. עלייה של סטיית תקן אחת ב- $x$  (משתנה ב-1) משנה את  $y$  במוצע ב- $r_{xy}$  סטיות תקן ( $\hat{t}$  משתנה ב- $r_{xy}$ ).
- נגדיר שני וקטורים,  $z$  ו- $t$ , שניהם ב- $\mathbb{R}^M$ . הזווית ביניהם היא  $\theta$ , ומתקיים  $r_{xy} = \cos \theta$  (איור 7.4). הוכחה. המכפלה הסקלרית המתאימה היא

$$(7.42) \quad \mathbf{z} \cdot \mathbf{t} = \|\mathbf{z}\| \|\mathbf{t}\| \cos(\theta)$$

ע"פ (7.38),

$$(7.43) \quad \mathbf{z} \cdot \mathbf{t} = \sum_{i=1}^M z_i t_i = M r_{xy}.$$

מאחר ו- $\|\mathbf{z}\| = \sqrt{\sum z_i^2} = \sqrt{M s_z^2} = \sqrt{M}$ , מתקיים

$$(7.44) \quad \|\mathbf{z}\| \|\mathbf{t}\| = \sqrt{M} \cdot \sqrt{M} = M$$

השלכות:

- תחום ערכים:  $-1 \leq r_{xy} \leq +1$ .
- מקרים מיוחדים:

- $r_{xy} = +1$  מעיד על קשר לינארי חיובי מושלם,  $\theta = 0^\circ$ .
- $r_{xy} = -1$  מעיד על קשר לינארי שלילי מושלם,  $\theta = 180^\circ$ .
- $r_{xy} = 0$  מעיד על היעדר קשר לינארי,  $\theta = 90^\circ$ .

**דוגמה 7.3 (רגרסיית z-score והקוסינוס):** יש לנרמל את הנתונים של דוגמה 7.2 ל-z-scores ולהתאים  $\hat{t} = w_0^* + w_1^* z$ .

(א) יש לרשום בטבלה את  $z_k$  ואת  $t_k$ .

(ב) יש לאמת מספרית ש- $w_0^* = 0$  (7.37) וש- $w_1^* = r_{xy}$  (7.38).

(ג) יש לאמת ש- $\text{MSE} = 1 - r_{xy}^2$ , ולהסביר מדוע זו אותה התאמה כמו קודם, ולא התאמה טובה יותר.

(ד) בעזרת  $r_{xy} = \cos \theta$  (איור 7.4), יש למצוא את הזווית  $\theta$  בין שני הווקטורים הממורכזים.

פתרון: ה-z-scores עם  $s_x = \sqrt{2} \approx 1.4142$  ו- $s_y = \sqrt{6} \approx 2.4495$ .

| 5      | 4      | 3 | 2       | 1       | $k$   |
|--------|--------|---|---------|---------|-------|
| 1.4142 | 0.7071 | 0 | -0.7071 | -1.4142 | $z_k$ |
| 1.6330 | 0.4082 | 0 | -0.8165 | -1.2247 | $t_k$ |

לשתי השורות ממוצע 0, ומעצם הבנייה שונות 1.

המקדמים. מאחר ו- $\bar{z} = \bar{t} = 0$ , הקבוע החופשי מתקבל מיד,

$$w_0^* = \bar{t} - w_1^* \bar{z} = 0,$$

וזו (7.37): מירכוז מבטל את הקבוע החופשי, תמיד. עבור השיפוע,  $s_z^2 = 1$ , ולכן  $w_1^*$  הוא השונות המשותפת של ה-z-scores בלבד,

$$\sum_k z_k t_k = 1.7321 + 0.5774 + 0 + 0.2887 + 2.3094 = 4.9075$$

$$w_1^* = \frac{1}{M} \sum_k z_k t_k = \frac{4.9075}{5} \approx 0.9815 = r_{xy}.$$

השגיאה. עם  $s_t^2 = 1$

$$\text{MSE} = 1 - r_{xy}^2 = 1 - 0.9633 = 0.0367.$$

זה אינו שיפור לעומת 0.22 של ההתאמה המקורית; זו אותה שגיאה, הנמדדת על ציר בקנה מידה אחר. חלוקת  $y$  ב- $s_y = \sqrt{6}$  מחלקת כל שגיאה ב- $\sqrt{6}$ , ולכן מחלקת את ה-MSE ב-6, ואכן  $0.0367 = 0.22/6$ . נרמול הוא שינוי יחידות. הוא אינו יכול לשנות את מה שהמודל לוכד, ו- $r_{xy}$ , שהוא חסר יחידות, אינו משתנה בו. מה שהנרמול כן נותן הוא סקאלה שבה ה-MSE נע בין 0 ל-1 ונקרא ישירות כ- $1 - R^2$ .

הזווית. כאשר אוספים את הערכים הממורכזים והמנורמלים לווקטורים  $z$  ו- $t$ , מקדם הקורלציה הוא הקוסינוס ביניהם,

$$r_{xy} = \cos \theta = 0.9815 \implies \theta = \arccos(0.9815) \approx 11.0^\circ.$$

שני הווקטורים כמעט מקבילים, ו-3.7% השונות שאינה מוסברת הם ריבוע הסינוס,  $\sin^2 \theta = 1 - \cos^2 \theta = 0.0367$ , שהוא ה-MSE שחושב לעיל.

## 7.5 Spearman Rank Correlation

**מטרה:** חלופה non-parametric (א-פרמטרית) לקורלציה של Pearson, עבור קשרים מונוטוניים שאינם בהכרח לינאריים.

מקדם הקורלציה של Pearson,  $r_{xy}$ , מודד קשר לינארי. כאשר הקשר בין  $x$  ל- $y$  מונוטוני אך לא לינארי, או כאשר הנתונים מכילים outliers (ערכים חריגים), ה-Spearman rank correlation (מקדם מתאם הדרגות של Spearman) מספק מדד עמיד יותר.

יהי  $R(x_i)$  ה-rank (הדרגה, כלומר המיקום בסדר הממוין) של  $x_i$  מבין  $\{x_1, \dots, x_M\}$  (עם דרגה ממוצעת עבור ties, ערכים שווים), ובאופן דומה  $R(y_i)$ . קורלציית Spearman היא קורלציית Pearson המחושבת על הדרגות:

$$(7.45) \quad r_s = r_{R(x)R(y)}$$

- תחום ערכים:  $-1 \leq r_s \leq +1$ .
- $|r_s| = 1$ : קשר מונוטוני מושלם (לא בהכרח לינארי).
- עמיד ל-outliers ולקשרים מונוטוניים לא לינאריים.
- עבור נתונים לינאריים,  $r_s \approx r_{xy}$ .

**דוגמה 7.4:** נתונות  $M = 5$  תצפיות בעלות קשר אקספוננציאלי  $y \approx e^x$ . הטבלה משווה בין מערך נתונים נקי (מונוטוני לחלוטין) לבין מערך נתונים רועש (החלפת דרגות ב-2, 3):

| $i$      | נקי ( $y_i = e^{x_i}$ ) |     |      |      |       | רועש |     |     |      |       |
|----------|-------------------------|-----|------|------|-------|------|-----|-----|------|-------|
|          | 1                       | 2   | 3    | 4    | 5     | 1    | 2   | 3   | 4    | 5     |
| $x_i$    | 1                       | 2   | 3    | 4    | 5     | 1    | 2   | 3   | 4    | 5     |
| $y_i$    | 2.7                     | 7.4 | 20.1 | 54.6 | 148.4 | 3.5  | 8.1 | 5.3 | 52.0 | 135.7 |
| $R(x_i)$ | 1                       | 2   | 3    | 4    | 5     | 1    | 2   | 3   | 4    | 5     |
| $R(y_i)$ | 1                       | 2   | 3    | 4    | 5     | 1    | 3   | 2   | 4    | 5     |

- נקי:  $r_s = 1$ ,  $r_{xy} \approx 0.89$ . Pearson מעריך בחסר את עוצמת הקשר (העקומה האקספוננציאלית אינה קו ישר), בעוד ש-Spearman מזהה נכון קשר מונוטוני מושלם.
- רועש:  $r_s = 0.90$ ,  $r_{xy} \approx 0.86$ . הרעש גורם ל- $y_2 > y_3$ , ומחליף בין הדרגות 2 ו-3. שני המקדמים קטנים, אך Spearman משקף במפורש את שבירת המונוטוניות.

## 7.6 קורלציה אינה סיבתיות

ערך  $r_{xy}$  (או  $r_s$ ) שונה מאפס מכמת association (קשר סטטיסטי) בין  $x$  ל- $y$ , אך אינו אומר דבר על כך ש- $x$  גורם ל- $y$ , כלומר על causation (סיבתיות). מספר מנגנונים שונים יכולים לייצר את אותה קורלציה:

- Direct causation (סיבתיות ישירה):  $x$  אכן משפיע על  $y$ .
- Reverse causation (סיבתיות הפוכה):  $y$  משפיע על  $x$ , ולא להפך.
- Confounding (גורם מבלבל, סיבה משותפת): משתנה שלישי  $z$  משפיע גם על  $x$  וגם על  $y$ , ויוצר קשר ביניהם אף שאף אחד מהם אינו גורם לשני.
- Selection bias (הטיית בחירה): אופן הדגימה של מערך הנתונים יוצר קורלציה שאינה קיימת באוכלוסייה עצמה.
- Coincidence (צירוף מקרים): במדגמים קטנים, או כאשר נבדקים זוגות משתנים רבים, ערכי  $|r_{xy}|$  גדולים מתקבלים במקרה בלבד (spurious correlation, קורלציה מדומה).

**דוגמה 7.5:** בעיירת חוף, מכירות הגלידה היומיות ( $x$ ) ומקרי הטביעה ( $y$ ) נמצאים בקורלציה חיובית חזקה. הגלידה אינה גורמת לטביעה; שניהם מונעים ע"י confounder (גורם מבלבל), טמפרטורת הקיץ, שמגדילה באופן בלתי תלוי גם את צריכת הגלידה וגם את מספר המתרחצים בחוף.

**דוגמה 7.6 (זיהוי המנגנון):** כל אחת מהקורלציות הבאות אמיתית וניתנת לשחזור. יש לזהות את המנגנון שמאחוריה, ולהציע בדיקה אחת שתבדיל בינה לבין קשר סיבתי.

(א) מטופלים בבית חולים שטופלו ביחידה לטיפול נמרץ מתים בשיעור גבוה יותר מאלה שטופלו במחלקה רגילה.

(ב) מספר הכבאים שנשלחים לשרפה נמצא בקורלציה חזקה עם הנזק שהשרפה גורמת.

(ג) בקרב המועמדים שהתקבלו לעבודה בחברה, הציון במבחן הכתוב נמצא בקורלציה שלילית עם הציון בראיון, אף שבקרב כלל המועמדים השניים אינם קשורים.

פתרון: (א) Reverse causation. מצב קרוב למוות גורם לאשפוז בטיפול נמרץ, ולא להפך. בדיקה: השוואה בין מטופלים בעלי חומרה שווה בעת הקבלה.

(ב) Confounding. גודל השרפה קובע גם את מספר הכבאים וגם את הנזק. בדיקה: השוואה בין שרפות בגודל דומה; בתוכן, יותר כבאים אמורים להביא לפחות נזק.

(ג) Selection bias. החברה מקבלת כל מי שהצליח באחד מהשניים, ולכן בקבוצת המתקבלים ציון חלש באחד מופיע רק לצד ציון חזק בשני. בדיקה: חישוב הקורלציה על כל המועמדים, כולל הנדחים; היא נעלמת.

אף אחד מהשלושה אינו צירוף מקרים: איסוף הנתונים מחדש מחזיר כל אחת מהקורלציות, כי המבנה שמייצר אותה עדיין קיים.

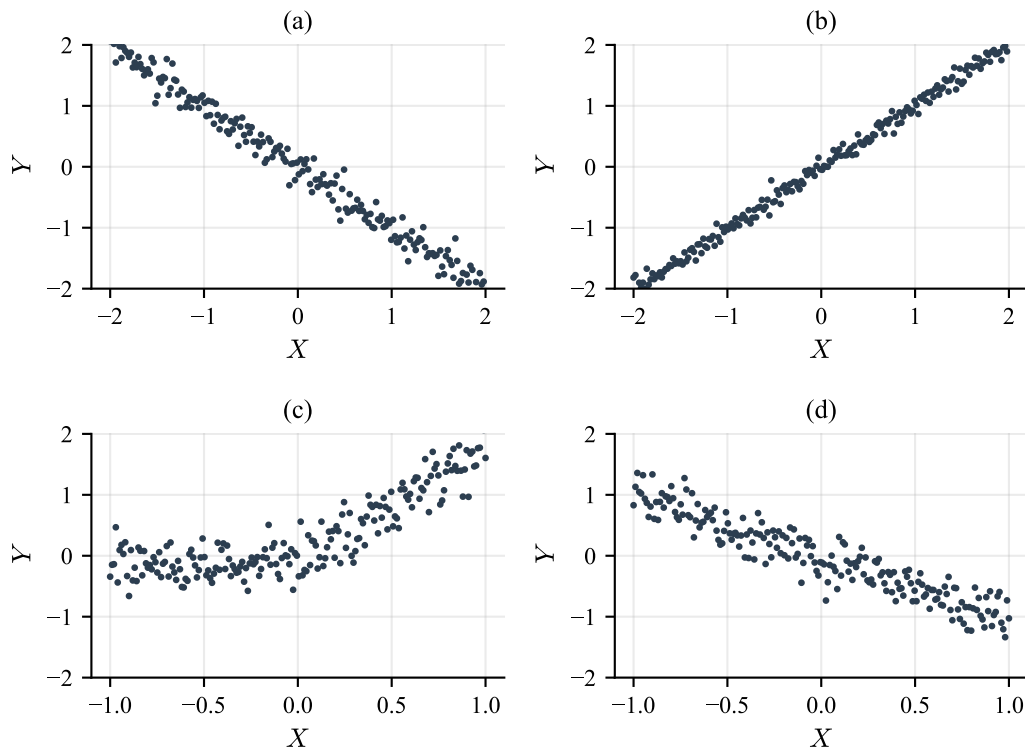
הוכחת סיבתיות דורשת ניסוי מבוקר או מסגרת ייעודית של causal inference (הסקה סיבתית). עוצמת הקורלציה לבדה, קרובה ככל שתהיה ל- $\pm 1$ , אינה עדות לקשר סיבתי.

## לקריאה נוספת

- **Spurious Correlations** מאת Tyler Vigen הוא המחשה מצוינת לעיקרון שקורלציה אינה סיבתיות.

## 7.7. שאלות

7.1 (קריאת מקדם הקורלציה) באיור 7.5 נתונות תוצאות של 4 ניסויים דו-ממדיים שונים. יש למיין מהגבוה לנמוך את הסדר של הניסויים ע"פ ערך מוחלט של המקדם הקורלציה,  $|r_{xy}|$ . יש לנמק את תשובתך.



איור 7.5: יש לרשום את הסדר של הניסויים ע"פ ערך מוחלט של המקדם הקורלציה, מהגבוה לנמוך בתהתאם לשאלה 7.1.

7.2 (Spearman מול Pearson) נלקחו שש מדידות של גודל שלפי ההשערה גדל באופן גאומטרי,

| רועש |    |   |   |   |   | נקי |    |   |   |   |   |       |
|------|----|---|---|---|---|-----|----|---|---|---|---|-------|
| 6    | 5  | 4 | 3 | 2 | 1 | 6   | 5  | 4 | 3 | 2 | 1 | $k$   |
| 6    | 5  | 4 | 3 | 2 | 1 | 6   | 5  | 4 | 3 | 2 | 1 | $x_k$ |
| 32   | 16 | 4 | 8 | 2 | 1 | 32  | 16 | 8 | 4 | 2 | 1 | $y_k$ |

עבור כל אחד ממערכי הנתונים:

(א) יש לחשב את מקדם Pearson,  $r_{xy}$ .

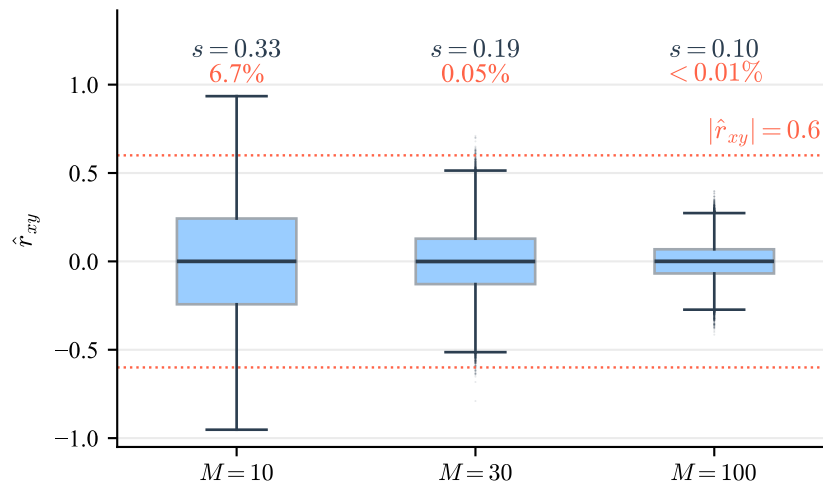
(ב) יש לחשב את מקדם Spearman,  $r_s$  (סעיף 7.5).

(ג) יש להסביר את הפער בין השניים במקרה הנקי, ומה משתנה במקרה הרועש.

סכומים שימושיים לשני מערכי הנתונים:  $\sum_k (x_k - \bar{x})^2 = 17.5$  ו-  $\sum_k (y_k - \bar{y})^2 = 703.5$ .

7.3 (סימולציה: באיזו תדירות רעש נראה מתואם) יש לבדוק, באמצעות סימולציה, כמה גדול צריך להיות מקדם קורלציה כדי שתהיה לו משמעות.

יש לייצר  $M$  זוגות  $(x_k, y_k)$  של ערכים נורמליים סטנדרטיים גלתי תלויים (randn), כך שמקדם הקורלציה האמיתי הוא אפס, ולחשב את  $\hat{r}_{xy}$ . יש לחזור על כך  $k$  פעמים, עבור  $M = 10$ , 30 ו-100. באיור 7.6 מוצגת התוצאה של סימולציה כזו עם  $k = 10^5$ .



איור 7.6:  $\hat{r}_{xy}$  של  $k = 10^5$  קבוצות של  $M$  זוגות בלתי תלויים, תיבה אחת לכל  $M$ . הערך  $s$  מעל כל תיבה הוא סטיית התקן של  $\hat{r}_{xy}$ .

(א) יש לשחזר את איור 7.6 בסימולציה משלכם עם  $k = 10^4$ . מה משותף לשלוש התיבות, ומה משתנה עם  $M$ ?

(ב) יש להשוות את ערכי סטיית התקן  $s$  שבאיור ל- $1/\sqrt{M-1}$ .

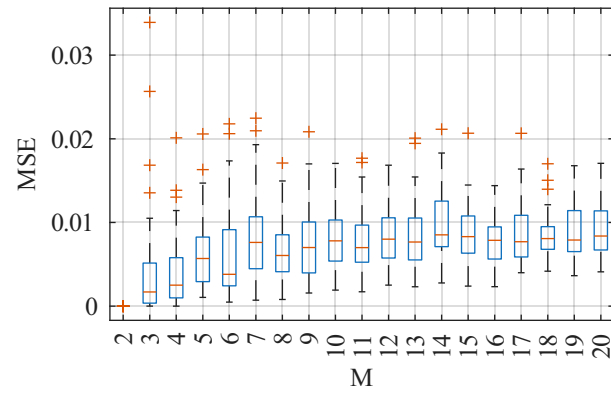
(ג) מה מוצג באדום? יש להסביר את הקווים האדומים ואת המספרים האדומים, ומה משמעות המספר האדום עבור  $M = 10$  לקורא שמוצא  $|r_{xy}| > 0.6$ , ערך שרוב הקוראים היו מכנים קורלציה חזקה, במערך נתונים של עשר נקודות.

(ד) מהו הכלל המעשי לדיווח מקדם קורלציה שעולה מכאן?

7.4 (סימולציה: שגיאת הרגרסיה כפונקציה של גודל המדגם) מטרת התרגיל היא לבחון את השפעת גודל המדגם על ריבועים פחותים לינאריים. יש לייצר נתונים ממודל לינארי ידוע,  $y_k = w_1 x_k + w_0 + \epsilon_k$ , כאשר ל- $\epsilon_k$  שונות  $\sigma^2$ , להתאים אותו ב-LS, ולבחון את ה-MSE המתקבל.

- יש לבצע רגרסיה לינארית בשיטת LS ולנתח את ה-MSE עבור גדלי מדגם  $M$  שונים.
- יש לחזור על כל הערכת MSE לפחות 30 פעמים.
- יש לסכם את התוצאות ב-boxplot (למשל, כמו באיור 7.7).
- האם סביר שה-MSE גדל כאשר  $M$  גדל (יותר נתונים)?

רמז לסעיף האחרון: יש לספור כמה מתוך  $M$  השגיאות חופשיות להשתנות לאחר ששני מקדמים הותאמו לאותם נתונים.



איור 7.7: ה-MSE כפונקציה של מספר נקודות הנתונים,  $M$ .

# א'. פתרונות לשאלות

פתרון השאלה 1.1 הדרכה: יש להציב  $Y = aX + b$  בתוך משוואה (1.17).

פתרון השאלה 1.2

$$\begin{aligned} p[k \geq 4] &= 1 - p[k < 4] \\ &= 1 - (p[k = 1] + p[k = 2] + p[k = 3]) \\ &= 1 - \left( \frac{4}{7} + \frac{2}{7} + \frac{1}{8} \right) = \frac{1}{56} \end{aligned}$$

פתרון השאלה 1.3 יש לספור את מספר התוצאות המתקבלות עבור כל אחת מתוצאות האפשרויות של ניסוי אקראי, ולנרמל ע"י סה"כ הניסויים.

```
N = 1e6; % number of random experiments
x = randi(6,[1 N]);
y = randi(6,[1 N]);
z = x + y;
empirical_cdf = sum(z == (1:12)',2)/N;
bar(empirical_cdf)
```

פתרון השאלה 1.4 תוצאות של זריקות מטבע מתפלג בינומית, כאשר מספר הצלחות  $k$  שווה למספר זריקות  $n$ ,

$$X \sim \text{Bin}(n, 1/2) \Rightarrow p = p_X[k = n] = \binom{n}{n} p^n (1-p)^{n-n} = \left(\frac{1}{2}\right)^n$$

דרך א': תוצאה של חזרה על הניסוי מתפלג בינומית, כאשר סיכוי לרצף אחדות בכל ניסוי הוא  $p$  מהסעיף הקודם. אחרי  $N$  ניסויים, סיכוי להצלחה אחת או יותר הוא המשלים של הסיכוי ל-0 הצלחות,  $p_Y[0]$ ,

$$Y \sim \text{Bin}(N, p) \Rightarrow p_2 = 1 - p_Y[0] = 1 - (1-p)^N = 1 - \left(1 - \left(\frac{1}{2}\right)^n\right)^N$$

בדרך ב': סיכוי לתוצאת ניסוי בודדת ללא רצף של  $n$  אחדות  $1 - \left(\frac{1}{2}\right)^n$ .

סיכוי ל- $N$  תוצאות ניסוי ללא רצף אחדות  $\left(1 - \left(\frac{1}{2}\right)^n\right)^N$ . כמובן, תשובה סופית זהה.

$$p_2 = 1 - \left(1 - \left(\frac{1}{2}\right)^{10}\right)^{2^{10}} = 0.632$$

### פתרון השאלה 1.5

$$X \sim \text{Bin}(n = 10^3, p = 10^{-3}), \quad p_X[0] = \binom{n}{0} p^0 (1-p)^n \approx 0.37$$

$$X \sim \text{Bin}(n = 10^3, p = 10^{-3}), \quad p_X[1] = \binom{n}{1} p^1 (1-p)^{n-1} \approx 0.37$$

$$X \sim \text{Bin}(n = 10^3, p = 10^{-3}), \quad \Pr(X \geq 3) = 1 - F_X(2) = 1 - p_X[0] - p_X[1] - p_X[2] \approx 0.08$$

### פתרון השאלה 1.6 נגדיר:

•  $p$ : ההסתברות שביט יגיע שגוי (Error).

•  $1-p$ : ההסתברות שביט יגיע תקין (Correct).

סעיף: שידור ללא כפילות (2 ביטים). בשידור של 2 ביטים, שגיאה מוגדרת כמצב שבו לפחות ביט אחד שגוי. נחשב זאת באמצעות המאורע המשלים: המקרה שבו אין שגיאה כלל (שני הביטים תקינים).

ההסתברות ששני הביטים תקינים:

$$\Pr(\text{correct message}) = (1-p) \cdot (1-p) = (1-p)^2$$

הסתברות השגיאה היא המשלים ל-1:

$$\begin{aligned} \Pr(\text{error}) &= 1 - \Pr(\text{correct message}) = 1 - (1-p)^2 \\ &= 1 - (1 - 2p + p^2) = 2p - p^2 \end{aligned}$$

סעיף: שידור תקין לחלוטין (4 ביטים). כדי שהשדר יתקבל באופן מושלם, נדרש שכל 4 הביטים (2 המקוריים + 2 החזרות) יגיעו ללא שגיאה. מכיוון שהשגיאות בביטים בלתי תלויות:

$$\Pr(\text{perfect transmission}) = (1-p) \cdot (1-p) \cdot (1-p) \cdot (1-p) = (1-p)^4$$

סעיף: שגיאה בלתי מזוהה עם כפילות (4 ביטים). נסמן את השדר כ- $M_1 M_2$ , כאשר  $M_1$  הוא הזוג הראשון ו- $M_2$  הוא החזרה. שגיאה בלתי מזוהה מתרחשת כאשר  $M_1 = M_2$  (המקלט מקבל את השדר), אך הערך שונה מהערך המקורי. זה קורה כאשר בשני הזוגות מתרחשת בדיוק אותה טעות.

נבחן את סוגי הטעויות האפשריים בבלוק של 2 ביטים:

- טעות בביט הראשון בלבד: ההסתברות היא  $p(1-p)$ .
- טעות בביט השני בלבד: ההסתברות היא  $(1-p)p$ .
- טעות בשני הביטים: ההסתברות היא  $p \cdot p = p^2$ .

כדי שהטעות לא תזוהה, עליה להתרחש גם ב- $M_1$  וגם ב- $M_2$ :

1. טעות בבית הראשון בשניהם:  $[p(1-p)] \times [p(1-p)] = p^2(1-p)^2$

2. טעות בבית השני בשניהם:  $[(1-p)p] \times [(1-p)p] = p^2(1-p)^2$

3. טעות בשני הביטים בשניהם:  $[p^2] \times [p^2] = p^4$

נסכום את ההסתברויות:

$$\Pr(\text{undetected error}) = 2 \cdot p^2(1-p)^2 + p^4 = 2p^2(1-p)^2 + p^4$$

פתרון השאלה 1.7 (א, ו)

פתרון השאלה 1.8

$$\begin{aligned} X &\sim \text{Geo}\left(p = \frac{1}{36}\right) \\ F_X(9) &= 1 - (1-p)^9 \\ \Pr(X \geq 10) &= 1 - \Pr(X \leq 9) = 1 - F_X(9) \\ &= \left(1 - \frac{1}{36}\right)^9 = \left(\frac{35}{36}\right)^9 \\ &= \frac{78815638671875}{101559956668416} \approx 0.776 \end{aligned}$$

דרך נוספת, ע"י חישוב סיכוי של 0 'הצלחות' מתוך 9,

$$\begin{aligned} X &\sim \text{Bin}\left(9, \frac{1}{36}\right) \\ p_X[0] &= \binom{9}{0} p^0 (1-p)^9 = \left(\frac{35}{36}\right)^9 \end{aligned}$$

סעיפים ב', ג'

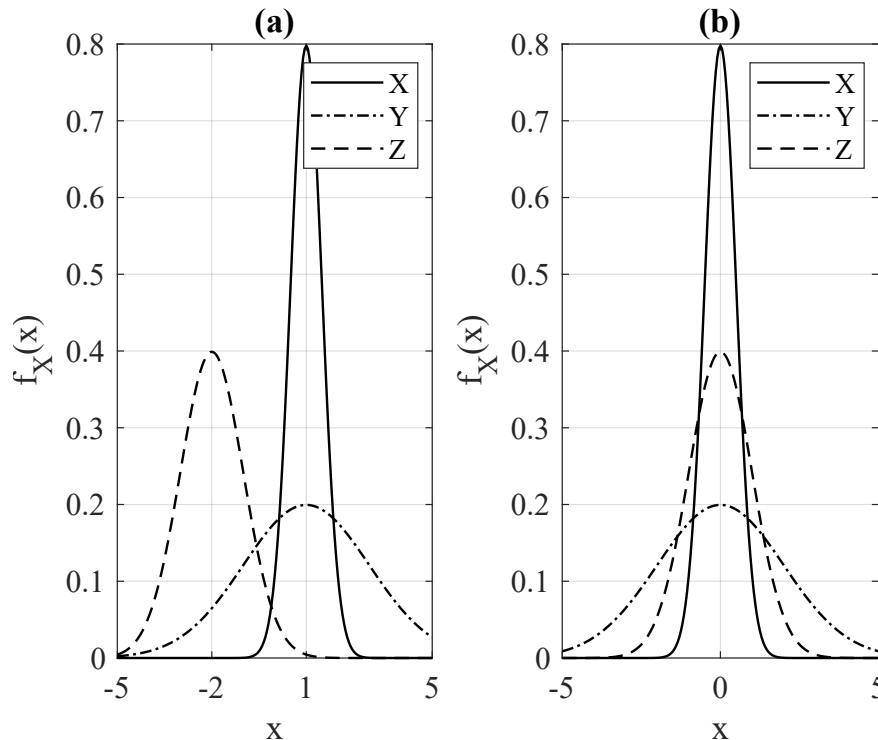
$$\begin{aligned} X &\sim \text{Bin}\left(10, \frac{1}{36}\right) \\ p_X[0] &= \binom{10}{0} \left(\frac{1}{36}\right)^0 \left(1 - \frac{1}{36}\right)^{10} \approx 0.754493 \\ \Pr(X \geq 1) &= 1 - F_X(0) = 1 - p_X[0] \\ &= 1 - (1-p)^n \\ &= 1 - \left(\frac{35}{36}\right)^{10} \approx 0.2455 \\ p_X[1] &= \binom{10}{1} p^1 (1-p)^{n-1} = 10 \left(\frac{1}{36}\right) \left(1 - \frac{1}{36}\right)^9 \\ &= \frac{10}{36} \left(\frac{35}{36}\right)^9 \approx 0.21557 \end{aligned}$$

## פתרון השאלה 2.1

$$\mu_X = \mu_Y = 1, \mu_Z = -2 \quad (\text{א})$$

$$\sigma_X < \sigma_Z < \sigma_Y \quad (\text{ב})$$

האיור להלן מציג את הפונקציות  $f_X(x)$  עבורן הוצגו בשאלה  $F_X(x) = \int_{-\infty}^x f_X(x) dx$ . באיור (א) הוצגו הפונקציות  $f_X(x)$  המקוריות, ובאיור (ב) נעשה נרמול של  $\mu_X = \mu_Y = \mu_Z = 0$ .



איור 2.6: (א) פונ' PDF, (ב) פונ' PDF אחרי הצבה של תוחלת 0.

## פתרון השאלה 2.2

(א) חישוב ע"י סימולציה

```
N = 1e6; % number of random experiments
X = normrnd(1/2,1/2,200,N); % Each random experiment is 200 numbers
Y = sum(X,1); % sum of 200 random numbers
nnz(find(Y >= 90 & Y <= 110))/N % numerical probability
```

(ב) חישוב תאורטי

$$\mathbb{E}[Y] = 200\mathbb{E}[X_i] = 100$$

$$\text{Var}[Y] = 200\text{Var}[X_i] = \frac{200}{4} = 50$$

$$\Pr(90 \leq Y) = Q\left(\frac{90 - \mathbb{E}[Y]}{\sqrt{\text{Var}[Y]}}\right) = Q\left(\frac{90 - 100}{\sqrt{50}}\right)$$

$$\Pr(110 \leq Y) = Q\left(\frac{110 - \mathbb{E}[Y]}{\sqrt{\text{Var}[Y]}}\right) = Q\left(\frac{110 - 100}{\sqrt{50}}\right)$$

$$\begin{aligned}
\Pr(90 \leq Y \leq 110) &= \Pr(90 \leq Y) - \Pr(110 \leq Y) \\
&= Q\left(\frac{-10}{\sqrt{50}}\right) - Q\left(\frac{10}{\sqrt{50}}\right) \\
&= 1 - 2Q\left(\frac{10}{\sqrt{50}}\right) \approx 0.8427
\end{aligned}$$

ולבסוף תשובה מספרית ע"י הקוד:

```
qfunc(-10/sqrt(50)) - qfunc(10/sqrt(50))
1-2*qfunc(10/sqrt(50))
```

### פתרון השאלה 2.3

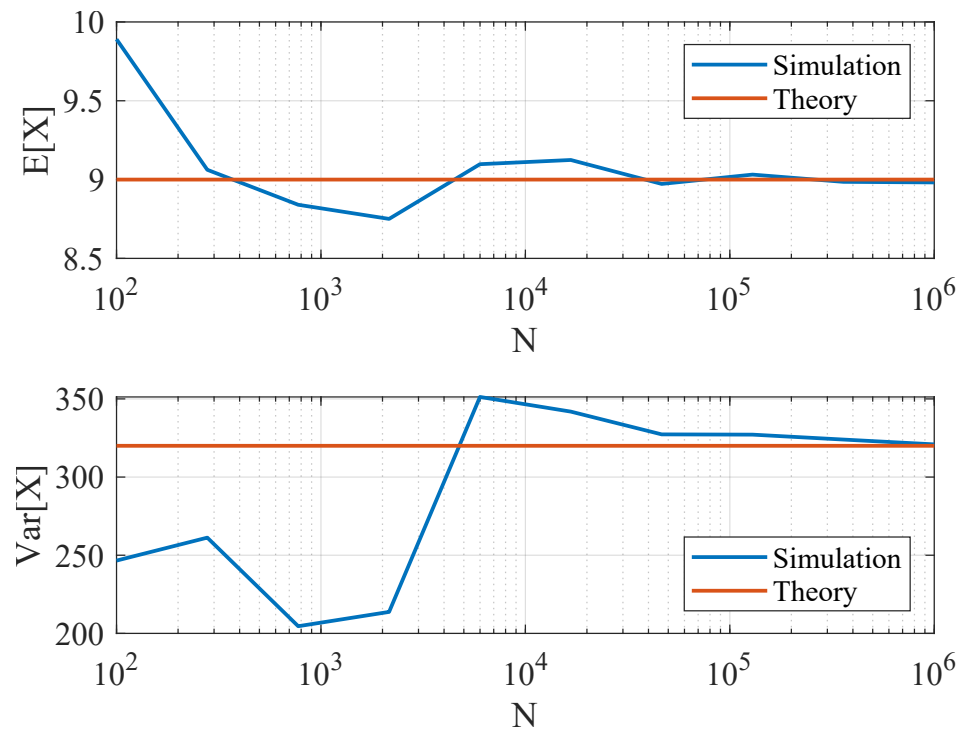
(א)

$$\begin{aligned}
\mathbb{E}[X] &= \frac{1}{\lambda} \\
\text{Var}[X] &= \frac{1}{\lambda^2} \\
\mathbb{E}[X^2] &= \text{Var}[X] + \left(\mathbb{E}[X]\right)^2 = \frac{2}{\lambda^2} \\
\mathbb{E}[X^4] &= \int_0^\infty x^4 f_X(x) dx = \int_0^\infty x^4 \lambda e^{-\lambda x} dx = \frac{24}{\lambda^4}
\end{aligned}$$

$$\begin{aligned}
\mathbb{E}[Y] &= \mathbb{E}[X^2 + 1] = \mathbb{E}[X^2] + 1 = \frac{2}{\lambda^2} + 1 \\
\text{Var}[Y] &= \text{Var}[X^2 + 1] = \text{Var}[X^2] = \mathbb{E}[X^4] - \left(\mathbb{E}[X^2]\right)^2 = \frac{24}{\lambda^4} - \frac{4}{\lambda^4} = \frac{20}{\lambda^4}
\end{aligned}$$

(ב)

$$\begin{aligned}
\lambda &= \frac{1}{2} \\
\mathbb{E}[Y] &= \frac{2}{\left(\frac{1}{2}\right)^2} + 1 = 9 \\
\text{Var}[Y] &= \frac{20}{\left(\frac{1}{2}\right)^4} = 320
\end{aligned}$$



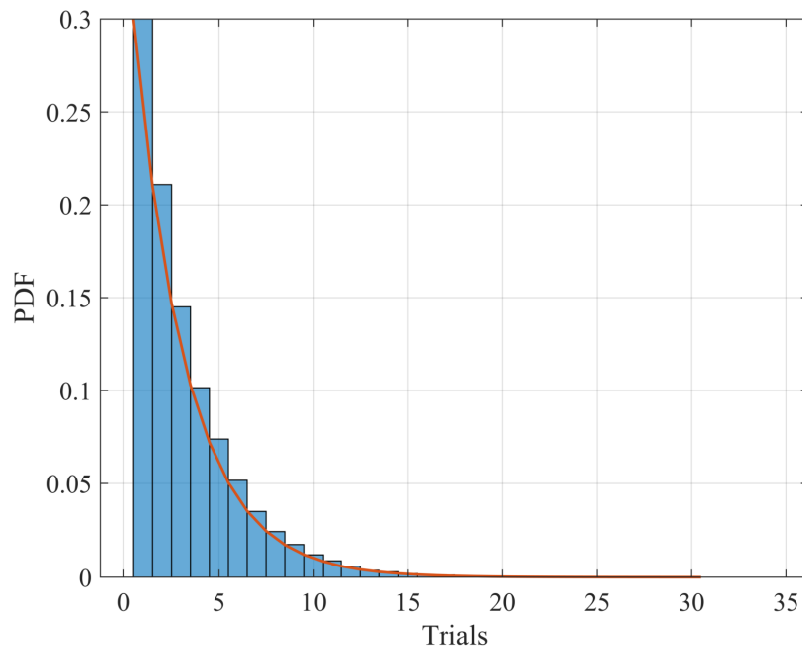
איור 2.7: דוגמה לסימולציה מתקבלת.

```
rng(42)
N = 1e6;
x = exprnd(2, [1 N]);
y = x.^2+1;

M = round(logspace(2,6,10));
for k = 1:length(M)
    E(k) = mean(y(1:M(k)));
    V(k) = var(y(1:M(k)));
end
```

(ג) התוחלת והשונות מתחילים להתקרב לערכים תאורטיים עם הגדלת ערך של  $N$ . עם זאת, התוצאה הנומרית עדיין אינה מדויקת גם עבור  $N = 10^6$ .

## פתרון השאלה 2.4



### פתרון השאלה 2.5

```
N = 1e6;           % number of random experiments
X = rand(1,N);     % random experiments
p = nnz(X.^(1/4) < 1/2)/N % numerical probability
```

פתרון השאלה 3.1 הצבה במשוואה (13.32) יחד עם תכונת שינוי קנה המידה של השונות המשותפת,

$$\text{Var}[\alpha X + \beta Y] = \alpha^2 \text{Var}[X] + 2\alpha\beta \text{Cov}[X, Y] + \beta^2 \text{Var}[Y]$$

נותנת זוג משוואות,

$$\begin{cases} \text{Var}[X] + 4\text{Cov}[X, Y] + 4\text{Var}[Y] = 40 \\ \text{Var}[X] - 4\text{Cov}[X, Y] + 4\text{Var}[Y] = 20 \\ \text{Var}[X] = \text{Var}[Y] \end{cases}$$

חיבור שתי הראשונות מבטל את השונות המשותפת, ויחד עם  $\text{Var}[X] = \text{Var}[Y]$  מתקבל  $10\text{Var}[X] = 60$ . חיסור שלהן מבודד את השונות המשותפת,

$$\begin{aligned} \text{Var}[X] &= 6 = \text{Var}[Y] \\ \text{Cov}[X, Y] &= \frac{20}{8} = 2.5 \end{aligned}$$

### פתרון השאלה 3.2 נרשום את התרגיל בצורה של טבלה,

| $x_k$ | $y_j$ | $p_{XY}[x_k, y_j]$ |
|-------|-------|--------------------|
| -1    | 0     | 1/3                |
| -1    | -1    | 1/6                |
| 0     | -1    | 1/3                |
| 1     | 1     | 1/6                |

תחילה נחשב את ההתפלגויות השוליות,

$$p_X[x] = \sum_y p_{XY}[x, y] = \begin{cases} \Pr[X = -1] = \frac{1}{2}, \\ \Pr[X = 0] = \frac{1}{3}, \\ \Pr[X = 1] = \frac{1}{6}, \end{cases}$$

$$p_Y[y] = \sum_x p_{XY}[x, y] = \begin{cases} \Pr[Y = -1] = \frac{1}{2}, \\ \Pr[Y = 0] = \frac{1}{3}, \\ \Pr[Y = 1] = \frac{1}{6} \end{cases}$$

נשים לב שההתפלגויות של  $X$  ושל  $Y$  זהות, ולכן גם התוחלות והשונויות,

$$\begin{aligned} \mathbb{E}[X] &= \mathbb{E}[Y] = (-1) \cdot \frac{1}{2} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{6} = -\frac{1}{3} \\ \text{Var}[X] &= \text{Var}[Y] = \mathbb{E}[X^2] - \left(\mathbb{E}[X]\right)^2 \\ &= \left(1 \cdot \frac{1}{2} + 0 \cdot \frac{1}{3} + 1 \cdot \frac{1}{6}\right) - \left(-\frac{1}{3}\right)^2 \\ &= \frac{2}{3} - \frac{1}{9} = \frac{5}{9} \end{aligned}$$

השונות המשותפת נובעת מהמומנט המעורב,

$$\begin{aligned} \mathbb{E}[XY] &= \sum_{x,y} x y p_{XY}[x, y] \\ &= (-1) \cdot 0 \cdot \frac{1}{3} + (-1) \cdot (-1) \cdot \frac{1}{6} \\ &\quad + 0 \cdot (-1) \cdot \frac{1}{3} + 1 \cdot 1 \cdot \frac{1}{6} = \frac{1}{3} \\ \text{Cov}[X, Y] &= \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y] \\ &= \frac{1}{3} - \left(-\frac{1}{3}\right) \left(-\frac{1}{3}\right) = \frac{2}{9} \end{aligned}$$

ולסיכום, מקדם הקורלציה,

$$\rho_{XY} = \frac{\text{Cov}[X, Y]}{\sqrt{\text{Var}[X] \text{Var}[Y]}} = \frac{2/9}{5/9} = \frac{2}{5}$$

פתרון השאלה 3.3 המכפלה מתפרקת ע"פ משוואה (3.24), והסכום ע"פ משוואות (3.22) ו-(3.34),

$$\begin{aligned} \mathbb{E}[X_i] &= \frac{1}{2}, \quad \text{Var}[X_i] = \frac{1}{12} \\ \mathbb{E}[X_i^2] &= \text{Var}[X_i] + \left(\mathbb{E}[X_i]\right)^2 = \frac{1}{12} + \frac{1}{4} = \frac{1}{3} \\ \mathbb{E}[Y] &= \prod_{i=1}^n \mathbb{E}[X_i] = \frac{1}{2^n} \\ \mathbb{E}[Z] &= \sum_{i=1}^n \mathbb{E}[X_i] = \frac{n}{2} \end{aligned}$$

$$\begin{aligned}
\text{Var}[Y] &= \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 \\
&= \prod_{i=1}^n \mathbb{E}[X_i^2] - \left( \prod_{i=1}^n \mathbb{E}[X_i] \right)^2 \\
&= \frac{1}{3^n} - \frac{1}{4^n} \\
\text{Var}[Z] &= \sum_{i=1}^n \text{Var}[X_i] = \frac{n}{12}
\end{aligned}$$

פתרון השאלה 3.4 המעריך הוא סכום, ולכן הצפיפות מתפרקת למכפלה של פונקציה של  $x$  בפונקציה של  $y$ ,

$$f_{XY}(x, y) = \left[ e^{-2x} u(x) \right] \left[ e^{-y/2} u(y) \right] = f_X(x) f_Y(y)$$

כל אחד מהסוגריים הוא צפיפות אקספוננציאלית, ולכן התנאי (3.19) מתקיים, המשתנים בלתי תלויים, ומכאן ש- $\text{Cov}[X, Y] = 0$  וגם  $\rho_{XY} = 0$ .

### פתרון השאלה 3.5

$$\begin{aligned}
\mathbb{E}[Y] &= -6 \mathbb{E}[X] + 22 = -18 + 22 = 4, \\
\text{Var}[Y] &= \text{Var}[-6X + 22] = 36 \text{Var}[X] = 72, \\
\mathbb{E}[X^2] &= \text{Var}[X] + (\mathbb{E}[X])^2 = 2 + 9 = 11, \\
\mathbb{E}[XY] &= \mathbb{E}[-6X^2 + 22X] = -6 \mathbb{E}[X^2] + 22 \mathbb{E}[X] \\
&= -66 + 66 = 0, \\
\text{Cov}[X, Y] &= \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y] = 0 - 12 = -12, \\
\rho_{XY} &= \frac{-12}{\sqrt{2 \cdot 72}} = -1
\end{aligned}$$

מאחר ו- $Y$  הוא פונקציה לינארית מדויקת של  $X$  בעלת שיפוע שלילי, מקדם הקורלציה מגיע לחסם התחתון שלו.

### פתרון השאלה 3.6 המומנטים של הרדיוס האמיתי הם

$$\begin{aligned}
\mathbb{E}[r] &= \frac{a+b}{2} = 2, \\
\text{Var}[r] &= \frac{(b-a)^2}{12} = \frac{1}{3}, \\
\mathbb{E}[r^2] &= \text{Var}[r] + (\mathbb{E}[r])^2 = \frac{13}{3} \approx 4.3333
\end{aligned}$$

כאשר  $\mathbb{E}[r^2]$  מתקבל גם ממשוואה (2.18). השגיאה תורמת  $\mathbb{E}[n] = 0$  וכן  $\mathbb{E}[n^2] = \text{Var}[n] = 0.01$ . מאחר ו- $r$  ו- $n$  בלתי תלויים,

$$\mathbb{E}[X] = \mathbb{E}[r] + \mathbb{E}[n] = 2,$$

$$\mathbb{E}[C] = 2\pi \mathbb{E}[X] = 4\pi,$$

$$\text{Var}[X] = \text{Var}[r] + \text{Var}[n] = \frac{1}{3} + 0.01 \approx 0.3433,$$

$$\text{Var}[C] = 4\pi^2 \text{Var}[X] \approx 13.5543,$$

$$\mathbb{E}[X^2] = \text{Var}[X] + (\mathbb{E}[X])^2 \approx 4.3433,$$

$$\mathbb{E}[A] = \pi \mathbb{E}[X^2] \approx 13.6450$$

אותו ערך של  $\mathbb{E}[X^2]$  מתקבל גם ע"י פיתוח  $\mathbb{E}[(r+n)^2] = \mathbb{E}[r^2] + 2\mathbb{E}[r]\mathbb{E}[n] + \mathbb{E}[n^2]$ , שבו האיבר האמצעי מתאפס. ולבסוף,

$$\mathbb{E}[Xr] = \mathbb{E}[(r+n)r] = \mathbb{E}[r^2] + \mathbb{E}[r]\mathbb{E}[n] = \frac{13}{3},$$

$$\text{Cov}[X, r] = \mathbb{E}[Xr] - \mathbb{E}[X]\mathbb{E}[r] = \frac{13}{3} - 4 = \frac{1}{3}$$

השונות המשותפת שווה ל- $\text{Var}[r]$ : המדידה עוקבת אחר הרדיוס במדויק, והרעש אינו מוסיף דבר שהרדיוס שותף לו.

פתרון השאלה 3.7 כל התוחלות מחושבות עם הצפיפות הקבועה  $f_\Theta(\theta) = 1/(2\pi)$  על  $[-\pi, \pi]$

$$\mathbb{E}[X] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(\theta) d\theta = 0,$$

$$\mathbb{E}[Y] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \sin(\theta) d\theta = 0,$$

$$\begin{aligned} \mathbb{E}[XY] &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos(\theta) \sin(\theta) d\theta \\ &= \frac{1}{4\pi} \int_{-\pi}^{\pi} \sin(2\theta) d\theta = 0, \end{aligned}$$

$$\text{Cov}[X, Y] = 0$$

ולכן הזוג חסר קורלציה. המומנטים השניים הם

$$\begin{aligned} \mathbb{E}[X^2] &= \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos^2(\theta) d\theta \\ &= \frac{1}{4\pi} \int_{-\pi}^{\pi} (1 + \cos(2\theta)) d\theta = \frac{1}{2} = \mathbb{E}[Y^2] \end{aligned}$$

עבור  $g_1(x) = x^2$  ו- $g_2(y) = y^2$ ,

$$\mathbb{E}[X^2 Y^2] = \frac{1}{2\pi} \int_{-\pi}^{\pi} \cos^2(\theta) \sin^2(\theta) d\theta = \frac{1}{8}$$

ואילו  $\mathbb{E}[X^2] \mathbb{E}[Y^2] = \frac{1}{4}$ . השניים שונים, ולכן משוואה (3.23) אינה מתקיימת והמשתנים אינם בלתי תלויים. ואכן, תמיד מתקיים  $X^2 + Y^2 = 1$ , וזוהי תלות חזקה ככל שניתן.

פתרון השאלה 3.8 סעיף (א) הוא דוגמה 3.11 עם היפוך הסימן של  $Y$ , ומתקבלת משוואה (3.32). עבור סעיף (ב), נציב  $\text{Cov}[X, Y] = \rho\sigma_1\sigma_2$  ע"פ משוואה (3.36),

$$\begin{aligned}\mathbb{E}[X - Y] &= \mu_1 - \mu_2, \\ \text{Var}[X - Y] &= \sigma_1^2 + \sigma_2^2 - 2\rho\sigma_1\sigma_2\end{aligned}$$

פתרון השאלה 3.9 כל הסתברות היא מלבן, אולי לא חסום, ולכן משוואה (3.17) חלה. עבור הראשונה נציב  $d = 1, c = -\infty, b = 1, a = -1$

$$\begin{aligned}\Pr(-1 < X < 1, Y < 1) \\ &= F_{XY}(1, 1) - F_{XY}(-1, 1) - F_{XY}(1, -\infty) + F_{XY}(-1, -\infty) \\ &= F_{XY}(1, 1) - F_{XY}(-1, 1)\end{aligned}$$

מאחר ו- $F_X(-\infty) = 0$  כופה  $F_{XY}(\cdot, -\infty) = F_{XY}(-\infty, \cdot) = 0$ . עבור השנייה נציב  $b = 1, a = -\infty, d = \infty, c = -1$

$$\begin{aligned}\Pr(X < 1, -1 < Y) \\ &= F_{XY}(1, \infty) - \cancel{F_{XY}(-\infty, \infty)} - F_{XY}(1, -1) + \cancel{F_{XY}(-\infty, -1)} \\ &= F_X(1) - F_{XY}(1, -1)\end{aligned}$$

פתרון השאלה 3.10 עבור זוג גאוסי במשותף, אי תלות שקולה לשוונות משותפת אפס ע"פ תכונה 3.4, ולכן סעיף (א) מבקש  $\text{Cov}[V, W] = 0$ . שתי התוחלות הן

$$\begin{aligned}\mathbb{E}[V] &= \mathbb{E}[X_1] \cos(\theta) + \mathbb{E}[X_2] \sin(\theta) = m [\cos(\theta) + \sin(\theta)] \\ \mathbb{E}[W] &= -\mathbb{E}[X_1] \sin(\theta) + \mathbb{E}[X_2] \cos(\theta) = m [\cos(\theta) - \sin(\theta)] \\ \mathbb{E}[V] \mathbb{E}[W] &= m^2 [\cos^2(\theta) - \sin^2(\theta)] = m^2 \cos(2\theta)\end{aligned}$$

עבור המומנט המעורב, שני האיברים הנושאים את  $\mathbb{E}[X_1^2]$  ואת  $\mathbb{E}[X_2^2]$  מתבטלים מאחר ושתי השונויות שוות, והאיברים הנותרים מצטרפים ל- $\cos(2\theta)$ ,

$$\begin{aligned}\mathbb{E}[VW] &= -\mathbb{E}[X_1^2] \sin(\theta) \cos(\theta) + \mathbb{E}[X_2^2] \sin(\theta) \cos(\theta) \\ &\quad + \mathbb{E}[X_1 X_2] (\cos^2(\theta) - \sin^2(\theta)) \\ &= \mathbb{E}[X_1 X_2] \cos(2\theta)\end{aligned}$$

ולכן

$$\text{Cov}[V, W] = (b - m^2) \cos(2\theta)$$

עבור  $b \neq m^2$  הביטוי מתאפס כאשר  $\cos(2\theta) = 0$ , כלומר

$$\theta = \frac{\pi}{2}n - \frac{\pi}{4}, \quad n \in \mathbb{Z},$$

ובתוך  $[-\pi, \pi]$  פירוש הדבר  $\theta \in \left\{-\frac{3\pi}{4}, -\frac{\pi}{4}, \frac{\pi}{4}, \frac{3\pi}{4}\right\}$ .

עבור סעיף (ב),  $V$  ו- $W$  הם קומבינציות לינאריות של זוג גאוסי במשותף, ולכן ע"פ הגדרה 3.4 הם עצמם גאויים במשותף. ההתפלגות שלהם היא אפוא משוואה (3.41) עם חמשת הפרמטרים

$$\begin{aligned}\mathbb{E}[V] &= m(\cos \theta + \sin \theta), \\ \mathbb{E}[W] &= m(\cos \theta - \sin \theta), \\ \text{Var}[V] &= \mathbb{E}[V^2] - (\mathbb{E}[V])^2, \\ \text{Var}[W] &= \mathbb{E}[W^2] - (\mathbb{E}[W])^2, \\ \rho_{VW} &= \frac{(b - m^2) \cos(2\theta)}{\sqrt{\text{Var}[V] \text{Var}[W]}}\end{aligned}$$

כאשר  $\mathbb{E}[V^2]$  ו- $\mathbb{E}[W^2]$  מפותחים מתוך ההגדרות של  $V$  ו- $W$  באותו אופן כמו  $\mathbb{E}[VW]$  לעיל.

פתרון השאלה 3.11 הסכום שווה לאחד כאשר בדיוק אחד משלושת המשתנים שווה לאחד. שלושת המקרים זרים, וכל אחד מהם הוא מכפלה של שלוש הסתברויות בלתי תלויות,

$$\begin{aligned}\Pr(X_1 + X_2 + X_3 = 1) &= \Pr[X_1 = 1] \Pr[X_2 = 0] \Pr[X_3 = 0] \\ &\quad + \Pr[X_1 = 0] \Pr[X_2 = 1] \Pr[X_3 = 0] \\ &\quad + \Pr[X_1 = 0] \Pr[X_2 = 0] \Pr[X_3 = 1] \\ &= p_1(1 - p_2)(1 - p_3) + (1 - p_1)p_2(1 - p_3) \\ &\quad + (1 - p_1)(1 - p_2)p_3\end{aligned}$$

הסכום שווה לשתיים כאשר בדיוק אחד מהם שווה לאפס,

$$\begin{aligned}\Pr(X_1 + X_2 + X_3 = 2) &= p_1p_2(1 - p_3) + p_1(1 - p_2)p_3 \\ &\quad + (1 - p_1)p_2p_3\end{aligned}$$

סימולציה מגרילה שלושה רצפי Bernoulli וסופרת את הסכומים,

```
N = 1e6;
p = [0.2 0.4 0.7];
X = rand(3,N) < p(:);
S = sum(X,1);
p1 = nnz(S == 1)/N
p2 = nnz(S == 2)/N
```

### פתרון השאלה 3.12

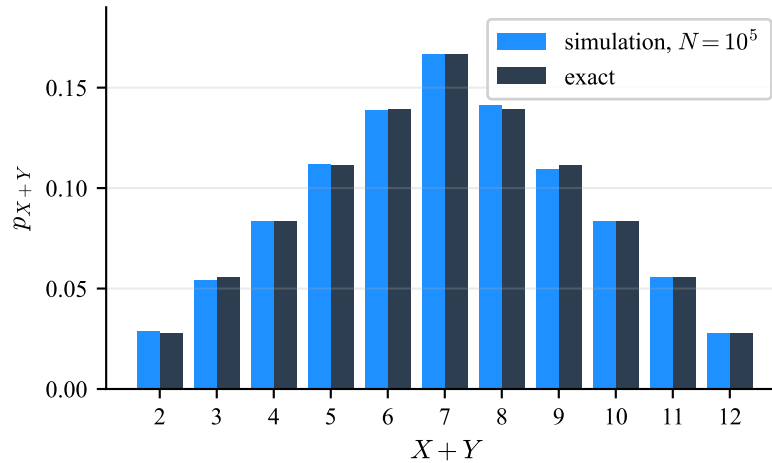
```
N = 1e6;           % number of random experiments
X = rand(1,N);     % random experiments
Y = rand(1,N);     %
p = nnz(X.^(1/3) > Y/2)/N % numerical probability
```

פתרון השאלה 3.13 נגריל שני רצפים בלתי תלויים של פאות, נחבר אותם, נספור כמה פעמים מתקבל כל סכום ונחלק במספר הניסויים (ראה איור להלן).

```

N = 1e5;
X = randi(6,1,N);
Y = randi(6,1,N);
S = X + Y;
p = histcounts(S, 1.5:1:12.5)/N

```



איור 3.2: התפלגות אמפירית של סכום זוג קוביות הוגנות, בהשוואה להתפלגות המדויקת.

פתרון השאלה 4.1 עם מקדם יחיד, הגזירה דורשת נגזרת אחת בלבד,

$$\begin{aligned}
 \text{MSE}(a) &= \mathbb{E}[(Y - aX)^2] = \mathbb{E}[Y^2] - 2a\mathbb{E}[XY] + a^2\mathbb{E}[X^2] \\
 \frac{d}{da}\text{MSE}(a) &= -2\mathbb{E}[XY] + 2a\mathbb{E}[X^2] = 0 \\
 a_{opt} &= \frac{\mathbb{E}[XY]}{\mathbb{E}[X^2]}
 \end{aligned}$$

יש לשים לב שהמומנטים כאן אינם ממורכזים, מאחר ולחזאי אין קבוע חופשי שיבלע את התוחלות. כתוצאה מכך השגיאה הממוצעת אינה מתאפסת,

$$\begin{aligned}
 \mathbb{E}[Y - \hat{Y}] &= \mathbb{E}[Y] - \frac{\mathbb{E}[XY]}{\mathbb{E}[X^2]}\mathbb{E}[X] \\
 \text{MSE}_{min} &= \mathbb{E}[Y^2] - \frac{(\mathbb{E}[XY])^2}{\mathbb{E}[X^2]}
 \end{aligned}$$

ויש להשוות זאת לתכונה 4.2.

פתרון השאלה 4.2 הגזירה זהה למקרה הקבוע ולמקרה הלינארי, עם  $X^2$  במקום  $X$ ,

$$\begin{aligned}
 \text{MSE}(a) &= \mathbb{E}[(Y - aX^2)^2] \\
 &= \mathbb{E}[Y^2] - 2a\mathbb{E}[X^2Y] + a^2\mathbb{E}[X^4]
 \end{aligned}$$

$$\frac{\partial}{\partial a} \text{MSE}(a) = -2\mathbb{E}[X^2Y] + 2a\mathbb{E}[X^4] = 0$$

$$a_{opt} = \frac{\mathbb{E}[X^2Y]}{\mathbb{E}[X^4]}$$

$$\text{MSE}_{min} = \mathbb{E}[Y^2] - \frac{(\mathbb{E}[X^2Y])^2}{\mathbb{E}[X^4]}$$

כל המומנטים הדרושים להלן נובעים מהמומנטים של המשתנה האחד,

$$\mathbb{E}[W_i^n] = \int_0^1 x^n dx = \frac{1}{n+1}, \quad i = 1, 2$$

ומאי התלות של  $W_1$  ו- $W_2$ ,

$$\begin{aligned} \mathbb{E}[Y^2] &= \mathbb{E}[W_1^4] = \frac{1}{5} = 0.2 \\ \mathbb{E}[X^2Y] &= \mathbb{E}\left[\left(\frac{9}{10}W_1 + \frac{1}{10}W_2\right)^2 W_1^2\right] \\ &= \left(\frac{9}{10}\right)^2 \mathbb{E}[W_1^4] + 2 \cdot \frac{9}{10} \cdot \frac{1}{10} \mathbb{E}[W_1^3] \mathbb{E}[W_2] \\ &\quad + \left(\frac{1}{10}\right)^2 \mathbb{E}[W_1^2] \mathbb{E}[W_2^2] \\ &= \left(\frac{9}{10}\right)^2 \frac{1}{5} + \frac{18}{100} \cdot \frac{1}{4} \cdot \frac{1}{2} + \frac{1}{100} \cdot \frac{1}{3} \cdot \frac{1}{3} \\ &= \frac{3341}{18000} \approx 0.18561 \\ \mathbb{E}[X^4] &= \mathbb{E}\left[\left(\frac{9}{10}W_1 + \frac{1}{10}W_2\right)^4\right] \\ &= \left(\frac{9}{10}\right)^4 \frac{1}{5} + 4 \left(\frac{9}{10}\right)^3 \frac{1}{10} \cdot \frac{1}{4} \cdot \frac{1}{2} \\ &\quad + 6 \left(\frac{9}{10}\right)^2 \left(\frac{1}{10}\right)^2 \frac{1}{3} \cdot \frac{1}{3} \\ &\quad + 4 \cdot \frac{9}{10} \left(\frac{1}{10}\right)^3 \frac{1}{2} \cdot \frac{1}{4} + \left(\frac{1}{10}\right)^4 \frac{1}{5} \\ &= \frac{8677}{50000} \approx 0.17354 \\ \text{MSE}_{min} &= \frac{83111}{56226960} \approx 1.478 \times 10^{-3} \end{aligned}$$

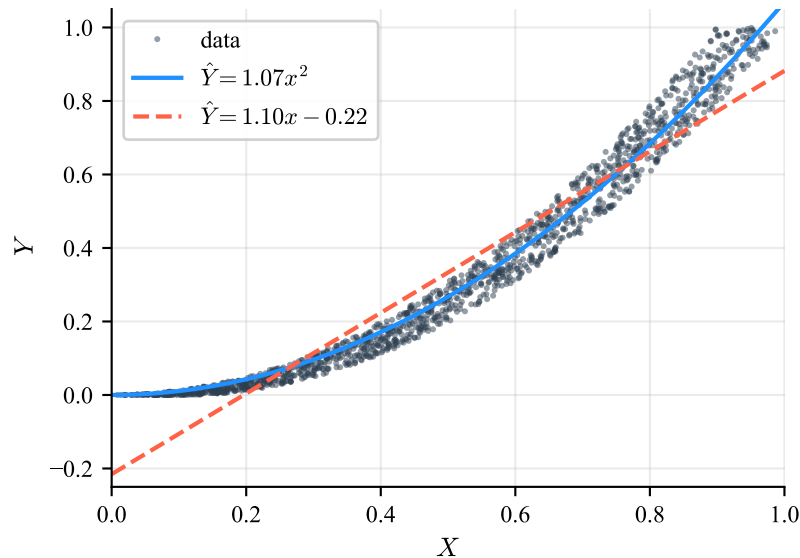
עבור החזאי הלינארי, אותם מומנטים נותנים

$$\begin{aligned} \text{Var}[X] &= \left(\frac{9}{10}\right)^2 \frac{1}{12} + \left(\frac{1}{10}\right)^2 \frac{1}{12} = \frac{41}{600} \approx 0.06833 \\ \text{Var}[Y] &= \mathbb{E}[W_1^4] - \left(\mathbb{E}[W_1^2]\right)^2 = \frac{1}{5} - \frac{1}{9} = \frac{4}{45} \approx 0.08889 \\ \text{Cov}[X, Y] &= \frac{9}{10} \mathbb{E}[W_1^3] + \frac{1}{10} \mathbb{E}[W_1^2] \mathbb{E}[W_2] \end{aligned}$$

$$- \mathbb{E}[W_1^2] \left( \frac{9}{10} \mathbb{E}[W_1] + \frac{1}{10} \mathbb{E}[W_2] \right) = \frac{3}{40}$$

$$\text{MSE}_{\min} = \text{Var}[Y] (1 - \rho_{XY}^2) = \frac{97}{14760} \approx 6.572 \times 10^{-3}$$

החזאי הריבועי מדויק פי ארבעה בקירוב. זה צפוי:  $Y$  הוא פונקציה ריבועית של  $W_1$ , ו- $X$  הוא כמעט  $W_1$ , ולכן הקשר האמיתי בין  $X$  ל- $Y$  קרוב לריבועי (ראה איור להלן). קו ישר יכול רק לקרב אותו, ומשוואה (4.15) מודדת בדיוק את הפער הזה.



איור 4.2: החזאי הריבועי האופטימלי והחזאי הלינארי האופטימלי עבור זוג שהקשר האמיתי ביניהם ריבועי.

פתרון השאלה 4.3 רישום הנתונים בצורת טבלת PDF משותף,

| $x_k$ | $y_j$ | $p_{XY}[x_k, y_j]$ |
|-------|-------|--------------------|
| 0     | 0     | 1/4                |
| 1     | 1     | 1/4                |
| 2     | 2     | 1/4                |
| 2     | 3     | 1/4                |

הערך  $x = 2$  מופיע פעמיים, ולכן ההסתברות השולית שלו כפולה,

$$p_X[x] = \begin{cases} 1/4 & x = 0 \\ 1/4 & x = 1 \\ 1/2 & x = 2 \end{cases} \quad p_Y[y] = \begin{cases} 1/4 & y = 0 \\ 1/4 & y = 1 \\ 1/4 & y = 2 \\ 1/4 & y = 3 \end{cases}$$

$$\mathbb{E}[X] = \frac{1}{4} \cdot 0 + \frac{1}{4} \cdot 1 + \frac{1}{2} \cdot 2 = \frac{5}{4}$$

$$\mathbb{E}[Y] = \frac{1}{4} (0 + 1 + 2 + 3) = \frac{3}{2}$$

$$\begin{aligned}
\mathbb{E}[X^2] &= \frac{1}{4} \cdot 1 + \frac{1}{2} \cdot 4 = \frac{9}{4} \\
\mathbb{E}[Y^2] &= \frac{1}{4} (0 + 1 + 4 + 9) = \frac{14}{4} \\
\text{Var}[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{9}{4} - \frac{25}{16} = \frac{11}{16} \\
\text{Var}[Y] &= \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = \frac{14}{4} - \frac{9}{4} = \frac{5}{4} \\
\mathbb{E}[XY] &= \frac{1}{4} (0 \cdot 0 + 1 \cdot 1 + 2 \cdot 2 + 2 \cdot 3) = \frac{11}{4} \\
\text{Cov}[X, Y] &= \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y] = \frac{11}{4} - \frac{15}{8} = \frac{7}{8} \\
\rho_{XY} &= \frac{7/8}{\sqrt{\frac{11}{16} \cdot \frac{5}{4}}} \approx 0.944 \\
\hat{Y} &= \mathbb{E}[Y] + \frac{\text{Cov}[X, Y]}{\text{Var}[X]} (x - \mathbb{E}[X]) = \frac{14}{11}x - \frac{1}{11}
\end{aligned}$$

פתרון השאלה 4.4 ערכי  $X$  סימטריים סביב אפס, ולכן  $\mathbb{E}[X] = 0$  והשונויות המשותפת מצטמצמת למומנט המעורב,

$$\begin{aligned}
\mathbb{E}[X] &= 0 \\
\text{Var}[X] &= \frac{1}{5} (4 + 1 + 0 + 1 + 4) = 2 \\
\mathbb{E}[Y] &= \frac{1}{5} (0 + 0 + 3 + 3 + 4) = 2 \\
\text{Var}[Y] &= \frac{1}{5} (4 + 4 + 1 + 1 + 4) = \frac{14}{5} = 2.8 \\
\mathbb{E}[XY] &= \frac{1}{5} (0 + 0 + 0 + 3 + 8) = \frac{11}{5} = 2.2 \\
\text{Cov}[X, Y] &= \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y] = \frac{11}{5}
\end{aligned}$$

הצבה במשוואות (4.14) ו-(4.15),

$$\begin{aligned}
\hat{Y} &= 2 + \frac{11}{5} \cdot \frac{1}{2}x = 2 + \frac{11}{10} \cdot 1 = 3.1 \\
\rho_{XY} &= \frac{11/5}{\sqrt{2.8 \cdot 2}} \approx 0.9297 \\
\text{MSE}_{\min} &= \text{Var}[Y] (1 - \rho_{XY}^2) \approx 0.38
\end{aligned}$$

אותו חישוב בקוד:

```

X = [-2 -1 0 1 2];
Y = [0 0 3 3 4];
Ex = mean(X); Ey = mean(Y);
VarX = mean((X-Ex).^2); VarY = mean((Y-Ey).^2);
CovXY = mean(X.*Y) - Ex*Ey;
rho = CovXY/sqrt(VarX*VarY)
VarY*(1 - rho^2)

```

פתרון השאלה 4.5 סכימה לפי שורות ולפי עמודות בטבלה נותנת את ההתפלגויות השוליות,

$$p_X[x] = \{0.2, 0.3, 0.3, 0.2\}, \quad x = 1, \dots, 4$$

$$p_Y[y] = \{0.3, 0.4, 0.3\}, \quad y = 1, 2, 3$$

ומהן

$$\begin{aligned}\mathbb{E}[X] &= 1 \cdot 0.2 + 2 \cdot 0.3 + 3 \cdot 0.3 + 4 \cdot 0.2 = 2.5 \\ \mathbb{E}[X^2] &= 1 \cdot 0.2 + 4 \cdot 0.3 + 9 \cdot 0.3 + 16 \cdot 0.2 = 7.3 \\ \text{Var}[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = 1.05 \\ \mathbb{E}[Y] &= 1 \cdot 0.3 + 2 \cdot 0.4 + 3 \cdot 0.3 = 2 \\ \mathbb{E}[Y^2] &= 1 \cdot 0.3 + 4 \cdot 0.4 + 9 \cdot 0.3 = 4.6 \\ \text{Var}[Y] &= \mathbb{E}[Y^2] - (\mathbb{E}[Y])^2 = 0.6\end{aligned}$$

המומנט המעורב עובר על שש התאים השונים מאפס בלבד,

$$\begin{aligned}\mathbb{E}[XY] &= 1 \cdot 3 \cdot 0.2 + 2 \cdot 2 \cdot 0.2 + 2 \cdot 3 \cdot 0.1 \\ &\quad + 3 \cdot 1 \cdot 0.1 + 3 \cdot 2 \cdot 0.2 + 4 \cdot 1 \cdot 0.2 = 4.3 \\ \text{Cov}[X, Y] &= \mathbb{E}[XY] - \mathbb{E}[X] \mathbb{E}[Y] = 4.3 - 5 = -0.7 \\ \rho_{XY} &= \frac{-0.7}{\sqrt{1.05 \cdot 0.6}} = -\frac{\sqrt{7}}{3} \approx -0.882 \\ \hat{y} &= \mathbb{E}[Y] + \frac{\text{Cov}[X, Y]}{\text{Var}[X]} (x - \mathbb{E}[X]) \\ &= 2 - \frac{0.7}{1.05} (3 - 2.5) \approx 1.667\end{aligned}$$

הקורלציה שלילית חזקה, ובהתאם הערך החזוי של  $y$  עבור  $x = 3$  נופל מתחת ל- $\mathbb{E}[Y]$ .

פתרון השאלה 4.6 ההתפלגויות השוליות הן סכומי שורות ועמודות של ה-PDF המשותף,

$$\begin{aligned}p_X[1] &= p_{XY}[1, 0] + p_{XY}[1, 1] = p_1 = 0.6 \\ p_Y[1] &= p_{XY}[0, 1] + p_{XY}[1, 1] = p_2 = 0.8\end{aligned}$$

ע"פ משוואות (4.6) ו-(4.7), הקבוע הטוב ביותר הוא התוחלת ושגיאתו היא השונות,

$$\begin{aligned}b_x &= \mathbb{E}[X] = p_1 = 0.6 \\ \text{MSE}_{\min} &= \text{Var}[X] = p_1(1 - p_1) = 0.24\end{aligned}$$

אותן שתי תוצאות עבור  $Z$ , תוך שימוש בלינאריות התוחלת ובמשוואה (3.34) עבור השונות של סכום משתנים בלתי תלויים,

$$\begin{aligned}b_z &= \mathbb{E}[Z] = a\mathbb{E}[X] + c\mathbb{E}[Y] = ap_1 + cp_2 \\ \text{MSE}_{\min} &= \text{Var}[Z] = a^2p_1(1 - p_1) + c^2p_2(1 - p_2) \\ &= 0.24a^2 + 0.16c^2\end{aligned}$$

עבור הסעיף האחרון, נפתח את השונות המשותפת של  $X$  עם  $Z$ . האיבר  $\mathbb{E}[XY]$  מתפרק מאחר ו- $X$  ו- $Y$  בלתי תלויים, והוא מצטמצם במדויק,

$$\begin{aligned}\mathbb{E}[XZ] &= a\mathbb{E}[X^2] + c\mathbb{E}[X]\mathbb{E}[Y] = ap_1 + cp_1p_2 \\ \text{Cov}[X, Z] &= \mathbb{E}[XZ] - \mathbb{E}[X]\mathbb{E}[Z] \\ &= ap_1 + cp_1p_2 - p_1(ap_1 + cp_2) \\ &= a(p_1 - p_1^2) = a\text{Var}[X]\end{aligned}$$

כאשר  $\mathbb{E}[X^2] = p_1$  מאחר ו- $X$  מקבל רק את הערכים 0 ו-1. מאחר ומתקיים  $\text{Var}[X] > 0$  עבור  $0 < p_1 < 1$ , השונות המשותפת מתאפסת רק עבור  $a = 0$ . לא קיים אפוא זוג קבועים שונים מאפס: התרומה של  $Y$  ל- $Z$  אינה רלוונטית כאן, ורק הסרת  $X$  מתוך  $Z$  יכולה לנקות את המתאם ביניהם.

פתרון השאלה 4.7 מההגדרה של מקדם הקורלציה, משוואה (3.36), עם  $\sigma_X = \sqrt{4} = 2$  ו- $\sigma_Y = \sqrt{1} = 1$ ,

$$\text{Cov}[X, Y] = \rho_{XY}\sigma_X\sigma_Y = \frac{1}{2} \cdot 2 \cdot 1 = 1$$

מכיוון שהתוחלת היא אפס, המומנט השני של הסכום הוא השונות שלו,

$$\begin{aligned}\mathbb{E}[(X + Y)^2] &= \text{Var}[X + Y] \\ &= \text{Var}[X] + \text{Var}[Y] + 2\text{Cov}[X, Y] \\ &= 4 + 1 + 2 \cdot 1 = 7\end{aligned}$$

מאחר והתוחלות אפס, הקבוע החופשי של משוואה (4.13) מתאפס והחזאי מצטמצם למקדם היחיד של משוואה (4.13), כאשר תפקידי  $X$  ו- $Y$  מוחלפים,

$$a = \frac{\text{Cov}[X, Y]}{\text{Var}[Y]} = \frac{1}{1} = 1 \quad \Rightarrow \quad \hat{X} = Y$$

את השגיאה ניתן לחשב ישירות, או ממשוואה (4.15),

$$\begin{aligned}\mathbb{E}[(X - Y)^2] &= \text{Var}[X - Y] = \text{Var}[X] + \text{Var}[Y] - 2\text{Cov}[X, Y] \\ &= 4 + 1 - 2 \cdot 1 = 3 \\ &= \text{Var}[X] \left(1 - \rho_{XY}^2\right) = 4 \left(1 - \frac{1}{2^2}\right) = 3\end{aligned}$$

ולבסוף,  $X + Y$  הוא קומבינציה לינארית של זוג גאוסי במשותף, ולכן גאוסי, בעל תוחלת אפס ושונות 7,

$$Z = X + Y \sim \mathcal{N}(0, 7), \quad \Pr(Z > 1) = Q\left(\frac{1}{\sqrt{7}}\right) \approx 0.354$$

פתרון השאלה 4.8 השונות המשותפת של הזוג הראשון נובעת ממשוואה (3.36),

$$\text{Cov}[X_1, X_2] = \rho_{X_1X_2}\sigma_1\sigma_2 = \frac{1}{\sqrt{2}} \cdot 1 \cdot \sqrt{2} = 1$$

ולכן, ע"פ משוואה (4.14),

$$\hat{X}_2 = \mathbb{E}[X_2] + \frac{\text{Cov}[X_1, X_2]}{\text{Var}[X_1]} (x_1 - \mathbb{E}[X_1]) = 1 + x_1$$

עבור הזוג השני, פיתוח כל גודל ע"פ משוואות (3.22) ו-(13.32),

$$\mathbb{E}[Y_3] = 2, \quad \mathbb{E}[Y_4] = 2, \quad \text{Var}[Y_3] = 13, \quad \text{Var}[Y_4] = 29, \quad \text{Cov}[Y_3, Y_4] = 19$$

נותן

$$\hat{Y}_4 = \mathbb{E}[Y_4] + \frac{\text{Cov}[Y_3, Y_4]}{\text{Var}[Y_3]} (y_3 - \mathbb{E}[Y_3]) = 2 + \frac{19}{13} (y_3 - 2)$$

$$\rho_{Y_3 Y_4} = 19 / \sqrt{13 \cdot 29} \approx 0.979$$

עבור הסעיף האחרון, נשווה את השונות הנתונות למה שמשוואה (13.32) מנבאת. אילו היו שני המשתנים בלתי תלויים, האיבר המעורב היה מתאפס והשונות הייתה יוצאת קטנה יותר,

$$\text{Var}[Y_2] = \text{Var}[X_1] + 4\text{Var}[X_3] - 4\text{Cov}[X_1, X_3] = 3 \implies \text{Cov}[X_1, X_3] = \frac{1}{2} \neq 0,$$

$$\text{Var}[Y_1] = \text{Var}[X_2] + \text{Var}[X_3] + 2\text{Cov}[X_2, X_3] = 5 \implies \text{Cov}[X_2, X_3] = 1 \neq 0$$

שונות משותפת שונה מאפס שוללת אי תלות, ע"פ תכונה 3.3.1, ולכן אף אחד מהזוגות אינו בלתי תלוי.

לבסוף, כל צירוף לינארי של  $Y_3$  ו- $Y_4$  הוא צירוף לינארי של  $X_1$  ו- $X_2$ , שהם גאוסיים במשותף, ולכן הוא גאוס. לפי הגדרה 3.4,  $Y_3$  ו- $Y_4$  גאוסיים במשותף. ההתפלגות המשותפת שלהם היא ה-PDF הגאוסית הדו-ממדית (3.41), עם חמשת הגדלים שכבר חושבו:

$$\mu_{Y_3} = \mu_{Y_4} = 2, \quad \sigma_{Y_3}^2 = 13, \quad \sigma_{Y_4}^2 = 29, \quad \rho_{Y_3 Y_4} \approx 0.979.$$

#### פתרון השאלה 4.9 ההיקף והשטח הם

$$C = 2X + 2Y, \quad A = XY$$

תוך שימוש ב- $\mathbb{E}[X] = \frac{1}{2}$ , ב- $\mathbb{E}[X^2] = \frac{1}{3}$  ובאי התלות של  $X$  ו- $Y$ ,

$$\mathbb{E}[C] = 2(\mathbb{E}[X] + \mathbb{E}[Y]) = 2, \quad \mathbb{E}[A] = \mathbb{E}[X] \mathbb{E}[Y] = \frac{1}{4},$$

$$\begin{aligned} \mathbb{E}[CA] &= 2(\mathbb{E}[X^2] \mathbb{E}[Y] + \mathbb{E}[X] \mathbb{E}[Y^2]) = \frac{2}{3}, \\ \text{Cov}[C, A] &= \mathbb{E}[CA] - \mathbb{E}[C] \mathbb{E}[A] = \frac{2}{3} - \frac{1}{2} = \frac{1}{6} \end{aligned}$$

כאן  $A$  הוא המשתנה הנצפה, ולכן הוא תופס את מקומו של  $X$  במשוואה (4.14),

$$\hat{C} = \mathbb{E}[C] + \frac{\text{Cov}[C, A]}{\text{Var}[A]} (A - \mathbb{E}[A])$$

כאשר

$$\text{Var}[A] = \mathbb{E}[X^2] \mathbb{E}[Y^2] - (\mathbb{E}[X] \mathbb{E}[Y])^2 = \frac{1}{9} - \frac{1}{16} = \frac{7}{144},$$

ולכן

$$\hat{C} = 2 + \frac{1/6}{7/144} \left( A - \frac{1}{4} \right) = \frac{24}{7} A + \frac{8}{7}$$

עבור השגיאה, השונות של ההיקף ומקדם הקורלציה הם

$$\begin{aligned} \text{Var}[C] &= 4\text{Var}[X] + 4\text{Var}[Y] = \frac{2}{3}, \\ \rho_{CA} &= \frac{\text{Cov}[C, A]}{\sqrt{\text{Var}[C] \text{Var}[A]}} = \frac{\sqrt{42}}{7} \end{aligned}$$

ולכן

$$\text{MSE}_{\min} = \text{Var}[C] \left( 1 - \rho_{CA}^2 \right) = \frac{2}{3} \left( 1 - \frac{42}{49} \right) = \frac{2}{21} \approx 0.0952$$

### פתרון השאלה 5.1

(א)

$$\begin{aligned} \bar{x} &= \frac{1}{8} (2 + 4 + 4 + 4 + 5 + 5 + 7 + 9) = \frac{40}{8} = 5 \\ x_i - \bar{x} &= -3, -1, -1, -1, 0, 0, 2, 4 \\ \sum_i (x_i - \bar{x})^2 &= 9 + 1 + 1 + 1 + 0 + 0 + 4 + 16 = 32 \end{aligned}$$

סכום הסטיות הוא אפס, וזוהי הבדיקה החשבונית של  $\bar{x}$ .

(ב)

$$\begin{aligned} s_{\text{biased}}^2 &= \frac{32}{n} = \frac{32}{8} = 4, & s_{\text{biased}} &= 2 \\ s_{\text{unbiased}}^2 &= \frac{32}{n-1} = \frac{32}{7} \approx 4.571, & s_{\text{unbiased}} &\approx 2.138 \end{aligned}$$

(ג) הנתונים כבר ממוינים. עבור  $n = 8$  זוגי, החציון הוא הממוצע של שני סטטיסטי הסדר המרכזיים,  $x_{(4)} = 4$  ו- $x_{(5)} = 5$ , כלומר  $(4 + 5)/2 = 4.5$ . ה-mode הוא 4 (שלוש הופעות), והטווח הוא  $9 - 2 = 7$ .

(ד) שני המעריכים נבדלים בגורם קבוע התלוי ב- $n$  בלבד,

$$\frac{s_{\text{biased}}^2}{s_{\text{unbiased}}^2} = \frac{n-1}{n} = \frac{7}{8} = 0.875,$$

פער של 12.5%, גדול מכדי להתעלם ממנו. הפער היחסי הוא  $1/n$ , ה-bias היחסי של דוגמה 5.2, ולכן הבחירה מפסיקה להשפיע עבור  $n \geq 100$ .

(ה) שגיאת התקן (5.16) משתמשת ב- $s$  ה-unbiased של סעיף (ב), מאחר וה-quantiles של  $t$  מחושבים עבור נרמול זה,

$$\begin{aligned} \widehat{\text{SE}}(\bar{x}) &= \frac{s_{\text{unbiased}}}{\sqrt{n}} = \frac{2.138}{\sqrt{8}} = 0.756 \\ t_{7, 0.975} \cdot \widehat{\text{SE}}(\bar{x}) &= 2.365 \cdot 0.756 = 1.788 \\ \bar{x} \pm 1.788 &= [3.21, 6.79] \end{aligned}$$

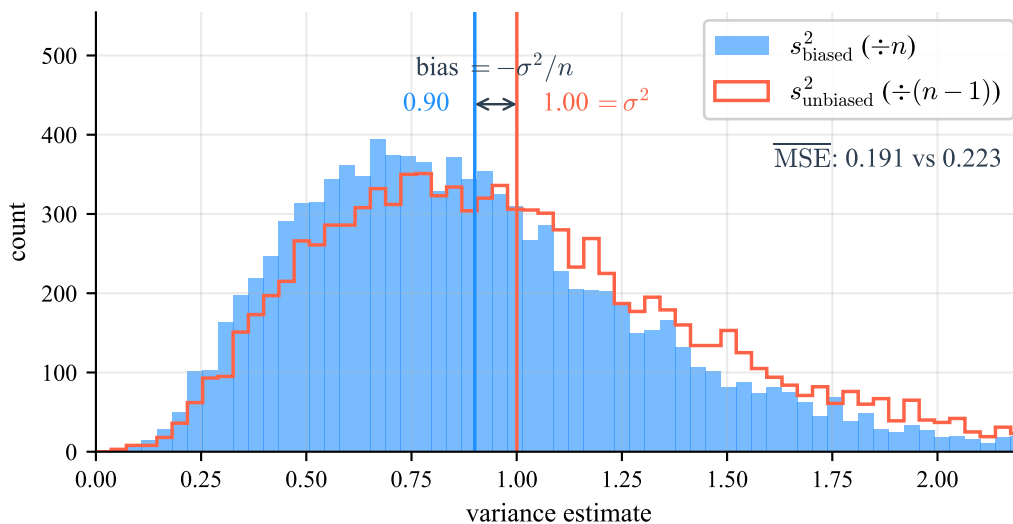
לפי (5.18), עם  $n - 1 = 7$  דרגות חופש.

רוחב הרווח הוא 3.58, לעומת טווח של 7: אי הוודאות לגבי מיקום המרכז מכסה מחצית מהפיזור של הנתונים עצמם. שמונה קריאות ממקמות את החייון ב- $5 \pm 1.8$ , והבחירה בין שני נרמולי השונות, ששווה 12.5% בסעיף (ד), אינה יכולה להתחרות בכך. ההחלטה כמה דגימות לאסוף היא השאלה העיקרית; ההחלטה כיצד לנרמל אותן היא פרט.

פתרון השאלה 5.2 עבור  $\sigma^2 = 1$  ו- $n = 10$  ארבעת ערכי המטרה הם

$$\begin{aligned} \mathbb{E}[s_{\text{unbiased}}^2] &= 1, & \overline{\text{MSE}}_{\text{unbiased}} &= \frac{2}{9} \approx 0.2222 \\ \mathbb{E}[s_{\text{biased}}^2] &= \frac{9}{10} = 0.9, & \overline{\text{MSE}}_{\text{biased}} &= \frac{19}{100} = 0.19 \end{aligned}$$

עבור  $k = 10^4$  הממוצעים המסומלצים צריכים ליפול בטווח של כ-1% מערכים אלו, לפי (5.22); ערכי ה-MSE המסומלצים הם עצמם ממוצעים, ומתייצבים מעט לאט יותר.



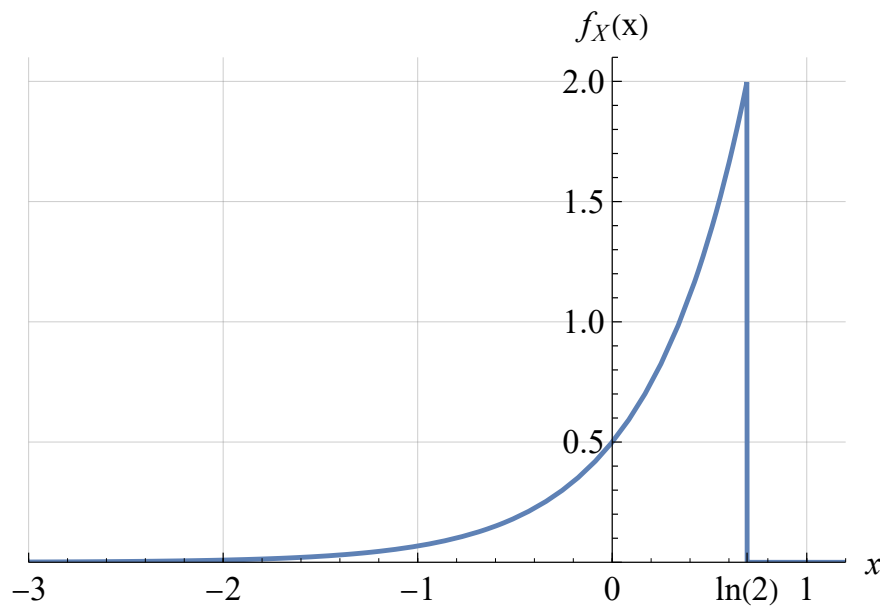
איור 5.9: שני המעריכים על פני  $k = 10^4$  חזרות של  $n = 10$  דגימות מ- $\mathcal{N}(0, 1)$  הם אינם העתקים מוזזים זה של זה אלא העתקים מוכפלים, מאחר ו- $s_{\text{unbiased}}^2 = \frac{n}{n-1} s_{\text{biased}}^2$ : ההיסטוגרמה ה-unbiased ממוקמת על  $\sigma^2$  אך רחבה יותר באותו יחס, והרוחב הנוסף הזה הוא שהופך את ה-MSE שלה לגדול מבין השניים.

(א) הממוצע של המעריך ה-unbiased הוא 1, ושל המעריך ה-biased הוא 0.9, החוסר של 10% שה-bias  $-\sigma^2/n$  חוזה.

(ב) למעריך ה-biased יש MSE קטן יותר, 0.19 לעומת 0.2222, נמוך בכ-14%. חלוקה ב- $n$  במקום ב- $n-1$  מקטינה מעט כל הערכה, והקטנת גודל חיובי רועש עולה מעט bias אך מסירה יותר שונות מכך. היעדר bias ודיוק הם שתי מטרות שונות, וכאן הן מושכות לכיוונים הפוכים.

(ג) עבור מספר בודד המדווה פעם אחת, המעריך ה-biased הוא ההימור הטוב יותר, והוא ברירת המחדל של הספר. המעריך ה-unbiased מצדיק את מקומו כאשר ה-bias היה מצטבר, ובעיקר בתוך רווחי הסמך של סעיף 5.2.2, שה-quantiles של  $t$  ושל  $\chi^2$  שלהם מחושבים עבור הנרמול ב- $n-1$ . שתי התופעות נעלמות ככל ש- $n$  גדל: עבור  $n = 100$  הפער בכל אחת מהעמודות הוא כ-1%.

## פתרון השאלה 6.1

איור 6.10: גרף של  $f_X(x)$  עבור השאלה 6.1.

$$\int_{-\infty}^{\infty} f_X(x) dx = 1$$

$$\Rightarrow \int_{-\infty}^{\ln(a)} \frac{1}{2} e^{2x} dx = \frac{1}{4} e^{2x} \Big|_{-\infty}^{\ln(a)} = \frac{1}{4} a^2 = 1 \Rightarrow |a| = 2$$

הערה:  $a$  לא יכול להיות שלילי, בגלל  $\ln(a)$ .

$$F_X(x) = \int_{-\infty}^x f_X(x) dx = \int_{-\infty}^x \frac{1}{2} e^{2x} dx = \begin{cases} \frac{1}{4} e^{2x} & x < \ln(2) \\ 1 & x > \ln(2) \end{cases}$$

$$\begin{aligned} \mathbb{E}[X] &= \int_{-\infty}^{\ln(2)} x \cdot \frac{1}{2} e^{2x} dx \\ &= \frac{1}{8} e^{2x} [2x - 1] \Big|_{-\infty}^{\ln(2)} \\ &= \frac{1}{8} (e^{2\ln(2)} = 4) [2\ln(2) - 1] \\ &= \ln(2) - \frac{1}{2} \approx 0.193147 \end{aligned}$$

$$\begin{aligned} \mathbb{E}[X^2] &= \frac{1}{2} \int_{-\infty}^{\ln(2)} x^2 e^{2x} dx \\ &= \frac{1}{2} \cdot e^{2x} \left[ \frac{x^2}{2} - \frac{2x}{2^2} + \frac{2}{2^3} \right] \Big|_{-\infty}^{\ln(2)} \end{aligned}$$

$$\begin{aligned}
&= \frac{1}{2} \cdot 4 \left[ \frac{\ln^2(2)}{2} - \frac{\ln(2)}{2} + \frac{1}{4} \right] \\
&= \ln^2(2) - \ln(2) + \frac{1}{2} \approx 0.287306 \\
\text{Var}[X] &= \mathbb{E}[X^2] - (\mathbb{E}[X])^2 = \frac{1}{4} \\
\Pr(X < 0) &= \int_{-\infty}^0 f_X(x) dx = \int_{-\infty}^0 \frac{1}{2} e^{2x} dx \\
&= \frac{1}{2} \cdot \frac{1}{2} e^{2x} \Big|_{-\infty}^0 = \frac{1}{4} (1 - 0) = \frac{1}{4} \\
F_X(x_0) &= \frac{1}{2} = \frac{1}{4} e^{2x_0} \Rightarrow e^{2x_0} = 2 \Rightarrow 2x_0 = \ln(2) \\
x_0 &= \frac{\ln(2)}{2} \approx 0.34657359
\end{aligned}$$

$$\begin{aligned}
\Pr(0 < X < 1) &= 1 - \Pr(X < 0) = \frac{3}{4} \\
&= \int_0^1 f_X(x) dx = \int_0^{\ln(2)} \frac{1}{2} e^{2x} dx \\
&= \frac{1}{2} \cdot \frac{1}{2} e^{2x} \Big|_0^{\ln(2)} = \frac{1}{4} (2^2 - 1) \\
&= \frac{n}{N} \\
n &= N \cdot \Pr(0 < X < 1) = 100,000 \cdot \frac{3}{4} = 75,000 \\
\hat{f}_X(0.5) &= \frac{n}{N} \frac{1}{\Delta x} = \Pr(0 < X < 1) \frac{1}{\Delta x} = \frac{3}{4} \cdot \frac{1}{1} = \frac{3}{4}
\end{aligned}$$

פתרון השאלה 6.2 נתונים:  $n_i = 80, \Delta x = 4, N = 500$

(א) (1) סוג count: הגובה הוא מספר התוצאות ב-bin,  $H_{\text{count}} = n_i = 80$

(2) סוג probability: הגובה הוא השכיחות היחסית.

$$H_{\text{prob}} = \frac{n_i}{N} = \frac{80}{500} = 0.16$$

(3) סוג pdf: מנרמלים את השכיחות היחסית ברוחב ה-bin.

$$H_{\text{pdf}} \approx f_X(x_i) = \frac{n_i}{N \cdot \Delta x} = \frac{80}{500 \cdot 4} = \frac{80}{2000} = 0.04$$

(ב) בגרף מסוג pdf, ההסתברות מיוצגת על ידי השטח. נחשב את שטח המלבן:

$$H_{\text{pdf}} \cdot \Delta x = 0.04 \cdot 4 = 0.16 = H_{\text{prob}}$$

משמעות: השטח (0.16) מייצג את ההסתברות המשווערת שהמשתנה האקראי  $X$  יפול בתוך ה-bin [10, 14].

הערה: ערך זה זהה לגובה שהתקבל בסוג probability, כצפוי.

(ג) נשתמש בנוסחה של היסטוגרמה מסוג pdf:

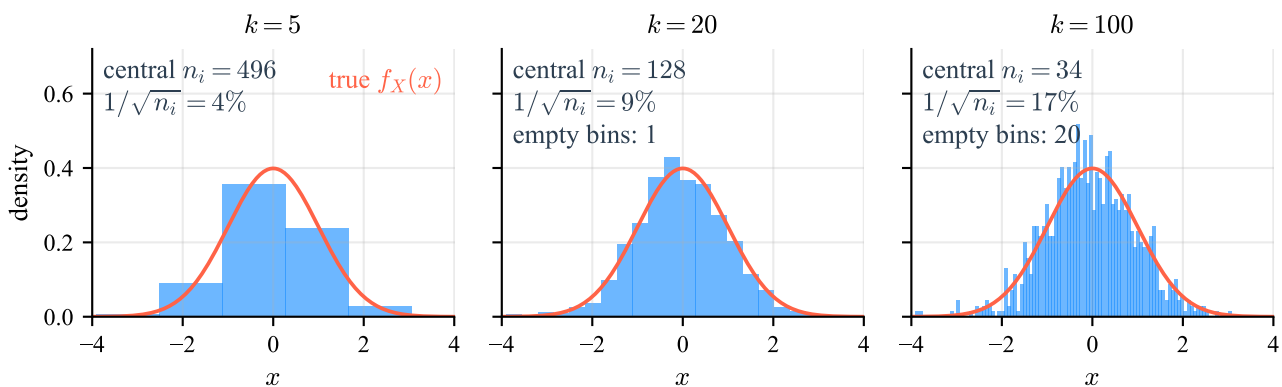
$$H_{\text{pdf}} = \frac{n_j}{N \cdot \Delta x} \Rightarrow 0.03 = \frac{n_j}{500 \cdot 4} = \frac{n_j}{2000} \Rightarrow n_j = 0.03 \cdot 2000 = 60$$

תשובה: ב-bin [20, 24] נספרו 60 תוצאות.

### פתרון השאלה 6.3

(א) ההגרלה משתרעת בערך על  $[-3.9, 3.1]$ , ולכן רוחב ה-bin הוא בערך  $7/k$ , ו-bin מרכזי מכיל בערך  $N f_X(0) \Delta x \approx 1000 \cdot 0.4 \cdot 7/k \approx 2800/k$  תצפיות. מספר התצפיות יורד אפוא כ- $1/k$ : כל חציה של רוחב ה-bin מחצה את הראיות שמאחורי כל עמודה. הרצה אחת נותנת

| $k$ | $\Delta x$ | $n_i$ מרכזי | $1/\sqrt{n_i}$ | bins מתחת ל-25 |
|-----|------------|-------------|----------------|----------------|
| 5   | 1.39       | 496         | 4.5%           | 1 מתוך 5       |
| 20  | 0.35       | 128         | 8.8%           | 9 מתוך 20      |
| 100 | 0.07       | 34          | 17.1%          | 86 מתוך 100    |



איור 6.11: אותה הגרלה של  $N = 1000$  תצפיות בשלושה רוחבי bin, עם הצפיפות האמיתית, כך שכל הבדל בין הגרפים נובע מ- $\Delta x$  בלבד. עבור  $k = 5$  העמודות יציבות אך צורת הפעמון נעלמה; עבור  $k = 100$  הצורה נכונה בממוצע אך העמודות הבודדות לא, ו-20 bins ריקים;  $k = 20$  מאזן בין השניים.

(ב)  $k = 20$ : ה-bins צרים מספיק כדי להפריד את השיא ורחבים מספיק כדי שכל עמודה תהיה אמינה עד כ-9%. שתי האחרות נכשלות מסיבות הפוכות. עבור  $k = 5$  העמודות יציבות, ושגויות בעליל: חמש עמודות אינן יכולות לייצג פעמון, וכל אחת ממצעת את הצפיפות על פני סטיית תקן שלמה, כך שהשיא משוטח והזנבות מוגזמים; השגיאה כאן אינה רעש אלא רוחב ה-bin. עבור  $k = 100$  העמודות עוקבות אחרי העקומה בממוצע אך מתפזרות סביבה בכחמישית מגובהן, וה-bins שבזנבות, עם תצפיות בודדות בכל אחד, מרמזים על מבנה שאינו קיים.

(ג) ה-bin המרכזי עובר את 25 התצפיות בכל שלושת המקרים, ולכן שיפוט היסטוגרמה לפי השיא שלה אינו מבחן כלל. העמודה האחרונה בטבלה היא החשובה: עבור  $k = 20$  כמעט מחצית מה-bins כבר נמצאים מתחת לכלל, ועבור  $k = 100$  גם 86 מהם, ו-20 אינם מכילים דבר. הזנבות תמיד הם הראשונים שנגמרות בהם התצפיות, ו- $N$  ו- $k$  חייבים לגדול יחד, מאחר והשגיאה היחסית תלויה רק ב- $n_i \approx N/k$ .

פתרון השאלה 7.1 הגרפים (a), (b) ו-(d) ישרים כולם, ולכן הם ממוינים לפי מידת הצמידות של הנקודות לקו; הגרף (c) גם מעוקל וגם המפוזר ביותר, ולכן הוא אחרון. הסדר הוא (b), (a), (d), (c), כאשר מקדמי הקורלציה הם:

(a)  $-0.991$ , (b)  $+0.997$ , (c)  $+0.835$ , (d)  $-0.941$

לסימן אין תפקיד במיון: לגרף (a) שיפוע שלילי, ובכל זאת הוא מקדים את הגרף (d), שגם שיפועו שלילי אך הפיזור שלו גדול יותר.

### פתרון השאלה 7.2

(א) נתונים נקיים. עם  $\bar{x} = 3.5$  ו- $\bar{y} = 63/6 = 10.5$

| $\sum$ | 6     | 5    | 4     | 3    | 2     | 1     | $k$             |
|--------|-------|------|-------|------|-------|-------|-----------------|
| 0      | 2.5   | 1.5  | 0.5   | -0.5 | -1.5  | -2.5  | $x_k - \bar{x}$ |
| 0      | 21.5  | 5.5  | -2.5  | -6.5 | -8.5  | -9.5  | $y_k - \bar{y}$ |
| 100.5  | 53.75 | 8.25 | -1.25 | 3.25 | 12.75 | 23.75 | מכפלה           |

$$r_{xy} = \frac{100.5}{\sqrt{17.5 \cdot 703.5}} = \frac{100.5}{110.96} \approx 0.906.$$

נתונים רועשים. החלפת  $y_3$  ו- $y_4$  אינה משנה את  $\bar{y}$  ואת  $\sum (y_k - \bar{y})^2$ , ולכן רק האיבר המעורב משתנה,

$$\sum_k (x_k - \bar{x})(y_k - \bar{y}) = 23.75 + 12.75 + 1.25 - 3.25 + 8.25 + 53.75 = 96.5, \quad r_{xy} = \frac{96.5}{110.96} \approx 0.870.$$

(ב) נתונים נקיים. גם  $x$  וגם  $y$  עולים ממש, ולכן שני וקטורי הדרגות הם 1, 2, 3, 4, 5, 6 והם זהים,  $r_s = 1$ .

נתונים רועשים. הדרגות של  $y$  הופכות ל-1, 2, 4, 3, 5, 6. עם  $R(x)$  שעדיין 1, 2, 3, 4, 5, 6, ולשני וקטורי הדרגות אותו פיזור  $\sum (R - \bar{R})^2 = 17.5$ .

$$\sum_k (R(x_k) - \bar{R})(R(y_k) - \bar{R}) = 6.25 + 2.25 - 0.25 - 0.25 + 2.25 + 6.25 = 16.5, \quad r_s = \frac{16.5}{17.5} \approx 0.943.$$

(ג) במקרה הנקי הקשר מושלם אך אינו לינארי, ושני המקדמים אומרים בדיוק את זה:  $r_s = 1$  מאשר את המונוטוניות המושלמת, בעוד ש- $r_{xy} = 0.906$  מדווח רק עד כמה קו ישר מקרב עקומת הכפלה. החוסר ב- $r_{xy}$  הוא אמירה על המודל, לא על הנתונים.

במקרה הרועש זוג אחד יצא מהסדר.  $r_s$  יורד ל-0.943, ויורד מהסיבה הנכונה: המונוטוניות אכן נשברה. גם  $r_{xy}$  יורד, ל-0.870, אך ערכו מערבב כעת שני פגמים שאינם קשורים, העקמומיות וההחלפה, ואינו יכול להפריד ביניהם. דיווח של שניהם הוא הדרך להבחין בין השניים:  $r_s$  גבוה עם  $r_{xy}$  נמוך יותר פירושו "מונוטוני, כדאי לנסות טרנספורמציה"; שניהם נמוכים פירושו שהקשר עצמו חלש.

### פתרון השאלה 7.3

(א) שלוש התיבות ממורכזות סביב אפס, כפי שהן חייבות להיות, כי כל ערך הוא רעש דגימה טהור סביב מקדם קורלציה אמיתי אפס. רק הרוחב שלהן משתנה: התיבה של  $M = 10$  רחבה מספיק כדי שה-whiskers שלה יחרגו מעבר ל- $\pm 0.6$ , ואילו התיבה של  $M = 100$  היא רצועה צרה סביב אפס, וכל ה-outliers שלה בתוך  $\pm 0.4$ .

(ב) סטיית התקן מתנהגת כ- $1/\sqrt{M-1}$ , שנותן 0.33, 0.19 ו-0.10 עבור  $M = 10$ , 30 ו-100, שלושת ערכי  $s$  שבאיור.

(ג) שני הקווים האדומים המנוקדים מסמנים את הסף  $|\hat{r}_{xy}| = 0.6$ . המספר האדום מתחת לכל  $s$  הוא החלק מתוך  $k$  ההרצות שנופל מחוץ לקווים אלה, כלומר ההסתברות שנתונים ללא שום קשר ייתנו מקדם קורלציה מעבר ל- $\pm 0.6$ : 6.7% עבור  $M = 10$ , 0.05% עבור  $M = 30$ , ופחות מ-0.01% עבור  $M = 100$ . שיעורים אלה מתאימים לערכים המדויקים 6.7%, 0.046% וכמעט אפס, המתקבלים מ- $t = \hat{r}\sqrt{M-2}/\sqrt{1-\hat{r}^2}$  עם  $M-2$  דרגות חופש.

בערך הרצה אחת מתוך חמש עשרה עבור  $M = 10$  נותנת מקדם קורלציה שהיה מתואר כחזק. עשר נקודות פשוט אינן מספיקות כדי ש- $r_{xy}$  יהיה בעל משמעות בפני עצמו. עבור  $M = 100$  אותו סף כמעט אינו בר השגה במקרה.

(ד) מקדם קורלציה אינו ניתן לפירוש ללא גודל המדגם שלצידו. יש לציין את  $M$  עם כל  $r_{xy}$ ; להתייחס ל- $|r_{xy}|$  הקטן מבערך  $2/\sqrt{M}$  כאל ערך שאינו נבדל מאפס; ולחשוד במיוחד במקדמים הגדולים ביותר שנמצאו במערכי הנתונים הקטנים ביותר, כי אלה בדיוק המקדמים שהסימולציה הזו מייצרת מרעש טהור.

זהו המנגנון של coincidence (צירוף מקרים) מסעיף 7.6, בצורה כמותית. כך גם נוצרות הקורלציות המדומות שאסף Tyler Vigen (ראו לקריאה נוספת): בין זוגות רבים של סדרות קצרות ובלתי קשורות, מעטים מתואמים מאוד במקרה בלבד. לפני ששואלים אם לקורלציה יש הסבר סיבתי, כדאי לשאול אם היא צריכה הסבר כלשהו.

פתרון השאלה 7.4 מה ה-boxplot מראה. שני דברים זזים בבת אחת כאשר  $M$  גדל, וה-boxplot מפריד ביניהם: התיבות נודדות כלפי מעלה, לעבר  $\sigma^2$ , והן נעשות צרות יותר. שניהם צפויים, ואף אחד מהם אינו מעיד על מודל גרוע יותר.

מזוע השגיאה עולה. ההתאמה מוציאה שתי דרגות חופש על מציאת  $w_0$  ו- $w_1$ , ולכן רק  $M-2$  מתוך  $M$  השגיאות חופשיות. לכן

$$\mathbb{E}[\text{SSE}] = (M-2)\sigma^2 \quad \Rightarrow \quad \mathbb{E}[\text{MSE}] = \frac{M-2}{M}\sigma^2 = \sigma^2\left(1 - \frac{2}{M}\right),$$

אותו חוסר  $(n-1)/n$  של השונות ה-biased בפרק ה-descriptive statistics, עם שני מקדמים מותאמים במקום ממוצע מותאם אחד. מספרית,

| $M$                               | 5    | 10   | 30   | 100  | 1000  |
|-----------------------------------|------|------|------|------|-------|
| $\mathbb{E}[\text{MSE}]/\sigma^2$ | 0.60 | 0.80 | 0.93 | 0.98 | 0.998 |

כלומר העלייה אינה רק סבירה, היא ניתנת לחיזוי עד שתי ספרות אחרי הנקודה, ויש לה תקרה: ה-MSE מטפס לעבר  $\sigma^2$  ונעצר שם, כי  $\sigma^2$  הוא החלק של  $y$  שאף מודל לינארי אינו יכול להגיע אליו.

מזוע  $M$  קטן נראה טוב. עם מעט נקודות, הקו יכול לעבור קרוב לכולן; עבור  $M = 2$  הוא עובר בדיוק דרך שתייהן, וה-MSE הוא אפס. זה אינו דיוק אלא אריתמטיקה: שתי נקודות קובעות קו. המצוינות לכאורה עבור  $M$  קטן היא ההתאמה שבולעת את הרעש יחד עם האות, וה- $0.60\sigma^2$  עבור  $M = 5$  מעריך בחסר את רמת השגיאה האמיתית ב-40%.

מזוע הפיזור מצטמצם. ל- $\text{SSE}/\sigma^2$  שונות  $2(M-2)$ , ולכן

$$\text{std}[\text{MSE}] = \frac{\sigma^2\sqrt{2(M-2)}}{M} \approx \sigma^2\sqrt{\frac{2}{M}},$$

הדעיכה המוכרת כ- $1/\sqrt{M}$ . זה מה שהופך את ה-boxplot לסיכום הנכון: MSE בודד עבור  $M = 10$  יכול ליפול בערך בכל מקום בטווח  $\pm 45\%$  מהממוצע שלו, כך שסריקה שמורצת פעם אחת, ללא 30 החזרות, הייתה מראה עקומה משוננת ומזמינה מסקנה שגויה.

מה לקחת מכאן. ה-MSE הנמדד על אותם נתונים ששימשו להתאמה הוא אומדן מוטה של שגיאת המודל, אופטימי בפקטור  $1 - 2/M$ , וההטיה גרועה ביותר בדיוק כאשר הנתונים מועטים. הפתרון הוא להעריך את המודל על נתונים שהתאמה לא ראתה.

# ב'. מילון

קיים בלבול מסויים בתרגום של מושגי היסוד לשפה העברית. להלן הטבלאות המסכמות עבור המושגים המופיעים בחוברת זו.

כאשר לא קיים בעברית מונח מקובל וחד-משמעי, המונח נשאר באנגלית, כמקובל בספרות.

## מושגי יסוד

| שם בעברית                                   | Abbr.     | English term                     |
|---------------------------------------------|-----------|----------------------------------|
| משתנה אקראי<br>משתנה מקרי                   | RV        | Random variable                  |
| משתנה אקראי בדיד                            |           | Discrete random variable         |
| משתנה אקראי רציף                            |           | Continuous random variable       |
| הסתברות                                     |           | Probability                      |
| התפלגות                                     |           | Distribution                     |
| פונקציית צפיפות<br>פונקציית צפיפות הסתברות  | PDF       | Probability density function     |
| פונקציית הסתברות                            | PMF (PDF) | Probability mass function        |
| פונקציית הצטברות<br>פונקציית התפלגות מצטברת | CDF       | Cumulative distribution function |
| תוחלת                                       |           | Expectation                      |
| ממוצע                                       |           | Average                          |
| שונות                                       |           | Variance                         |
| סטיית תקן                                   |           | Standard deviation               |
| מומנט                                       |           | Moment                           |
| חציון                                       |           | Median                           |
| שכיח                                        |           | Mode                             |
| משפט הגבול המרכזי                           | CLT       | Central limit theorem            |

## התפלגויות

| English term                   | Abbr.         | שם בעברית                                |
|--------------------------------|---------------|------------------------------------------|
| Bernoulli distribution         | Ber           | התפלגות Bernoulli                        |
| Binomial distribution          | Bin           | התפלגות בינומית                          |
| Geometric distribution         | Geo           | התפלגות גאומטרית                         |
| Exponential distribution       | Exp           | התפלגות אקספוננציאלית<br>התפלגות מעריכית |
| Uniform distribution           | U             | התפלגות אחידה                            |
| Gaussian (normal) distribution | $\mathcal{N}$ | התפלגות גאוסית<br>התפלגות נורמלית        |
| Chi-square distribution        | $\chi^2$      | התפלגות chi-square                       |
| Student t-distribution         | $t$           | התפלגות Student t                        |
| Degrees of freedom             | dof           | דרגות חופש                               |
| Quantile                       |               | כמותון                                   |
| Sample variance                |               | שונות מדגמית                             |
| Unbiased                       |               | חסר הטיה                                 |

## זוג משתנים אקראיים

| English term                                                                | Abbr.        | שם בעברית                 |
|-----------------------------------------------------------------------------|--------------|---------------------------|
| Joint distribution                                                          |              | התפלגות משותפת            |
| Joint probability density function                                          | Joint PDF    | פונקציית צפיפות משותפת    |
| Joint probability mass function                                             | Joint PMF    | פונקציית הסתברות משותפת   |
| Marginal distribution                                                       |              | התפלגות שולית             |
| Marginal probability density function                                       | Marginal PDF | פונקציית צפיפות שולית     |
| Statistical independence                                                    |              | אי תלות סטטיסטית          |
| Covariance                                                                  |              | שונות משותפת              |
| Correlation coefficient<br>(Pearson product-moment correlation coefficient) |              | מקדם קורלציה<br>מקדם מתאם |
| Un-correlated                                                               |              | בלתי מתואם<br>חסר קורלציה |
| Orthogonal                                                                  |              | אורתוגונליים              |
| Mixed moment                                                                |              | מומנט מעורב               |
| Jointly Gaussian                                                            |              | גאוסיים במשותף            |
| Bivariate Gaussian                                                          |              | גאוסי דו-ממדי             |

## חיזוי

| English term              | Abbr. | שם בעברית                     |
|---------------------------|-------|-------------------------------|
| Prediction                |       | חיזוי                         |
| Predictor                 |       | חזאי                          |
| Linear prediction         |       | חיזוי לינארי                  |
| Linear trendline          |       | קו מגמה לינארי                |
| Linear regression         |       | רגרסיה לינארית                |
| Mean square error         | MSE   | שגיאה ריבועית ממוצעת          |
| Minimum mean square error | MMSE  | שגיאה ריבועית ממוצעת מינימלית |
| Prediction error          |       | שגיאת חיזוי                   |
| Estimator                 |       | מעריך                         |

## סטטיסטיקה וגרפים סטטיסטיים

| English term                               | Abbr. | שם בעברית                  |
|--------------------------------------------|-------|----------------------------|
| Descriptive statistics                     |       | סטטיסטיקה תיאורית          |
| Sample mean                                |       | ממוצע מדגמי                |
| Population variance                        |       | שונות האוכלוסייה           |
| Bias<br>Biased                             |       | הטיה<br>מוטה               |
| Repeated experiments                       |       | ניסויים חוזרים             |
| Standard error                             | SE    | שגיאת תקן                  |
| Confidence bound<br>(confidence interval)  |       | רווח סמך                   |
| Order statistics                           |       | סטטיסטי סדר                |
| Multimodal                                 |       | רב-שיאית                   |
| Skewness                                   |       | צידוד                      |
| Kurtosis<br>Excess kurtosis                |       | גבנוניות<br>גבנוניות עודפת |
| Monte-Carlo simulation                     |       | סימולציית מונטה קרלו       |
| Law of large numbers                       |       | חוק המספרים הגדולים        |
| Seed<br>Pseudo-random                      |       | זרע<br>פסאודו-אקראי        |
| Quartile<br>Percentile                     |       | רבעון<br>אחוזון            |
| Interquartile range                        | IQR   | טווח בין-רבעוני            |
| Outlier                                    |       | ערך חריג                   |
| Boxplot (box-and-whisker plot)<br>Whiskers |       | דיאגרמת קופסה<br>שפמים     |
| Histogram                                  |       | היסטוגרמה                  |
| Bin                                        |       | מקטע                       |
| Violin plot                                |       | דיאגרמת כינור              |

## ריבועים פחותים וקורלציה

| English term                     | Abbr. | שם בעברית                       |
|----------------------------------|-------|---------------------------------|
| Least squares                    | LS    | ריבועים פחותים                  |
| Dataset                          |       | מערך נתונים                     |
| Database                         |       | מסד נתונים                      |
| Loss function<br>Cost function   |       | פונקציית הפסד<br>פונקציית מחיר  |
| Metric                           |       | מדד ביצועים                     |
| Sum of squared errors            | SSE   | סכום ריבועי השגיאות             |
| Root-mean-square error           | RMSE  | שורש השגיאה הריבועית<br>הממוצעת |
| Normal equations                 |       | המשוואות הנורמליות              |
| Sample covariance                |       | שונות משותפת מדגמית             |
| Pearson correlation coefficient  |       | מקדם הקורלציה של Pearson        |
| Coefficient of determination     | $R^2$ | מקדם הקביעה                     |
| Total sum of squares             | SST   | סכום הריבועים הכולל             |
| Sum of squares due to regression | SSR   | סכום הריבועים המוסבר            |
| Residual                         |       | שארית                           |
| Goodness of fit                  |       | טיב התאמה                       |
| z-score                          |       | ציון תקן                        |
| Spearman rank correlation        |       | מתאם הדרגות של Spearman         |
| Rank                             |       | דרגה                            |
| Non-parametric                   |       | א-פרמטרי                        |
| Association                      |       | קשר סטטיסטי                     |
| Causation<br>Reverse causation   |       | סיבתיות<br>סיבתיות הפוכה        |
| Confounder (confounding)         |       | גורם מבלבל                      |
| Selection bias                   |       | הטיית בחירה                     |
| Spurious correlation             |       | קורלציה מדומה                   |
| Causal inference                 |       | הסקה סיבתית                     |