תמציות מרכז ופיזור, נקודות שבירה, ציוני תקן וקו הרגרסיה — פתרון מלא

פתרון מפורט עם הסבר לכל מעבר + תיבות 💡 על הטריקים והרעיונות החשובים

ההגדרות שנעבוד איתן (אלו ההגדרות הסטנדרטיות בקורס):
תמצית למרכז: $T: \mathbb{R}^n \to \mathbb{R}$ כך שלכל $a > 0,\ b \in \mathbb{R}$ מתקיימת שקילות למיקום ולסקלה: $T(aX + b\cdot 1) = a\cdot T(X) + b$.
תמצית לפיזור: $T(X) \geq 0$, אדישה להזזה וכופלת ב-$|a|$: לכל a, b: $T(aX + b\cdot 1) = |a|\cdot T(X)$.
נקודת שבירה: $\varepsilon^*_n = m^*/n$, כאשר m^* הוא המספר הקטן ביותר של תצפיות שיריב יכול להחליף בערכים שרירותיים כך שהתמצית "התפוצץ" (תשאף ל-$\pm\infty$). **נקודת שבירה אסימפטוטית:** הגבול כאשר $n \to \infty$. תזכורת לגבי שברונים: לפי ההגדרה $x_p = \inf\{x : \hat{F}(x) \geq p\}$, השברון האמפירי הוא סטטיסטי הסדר $x_p = X_{(k)}$ עם $k = \lceil np \rceil$.

שאלה 2 — האם התמציות הבאות חוקיות? ומהן נקודות השבירה?

2.1 המקסימום $T(X) = X_{(n)}$ — תמצית למרכז? כן

בדיקת ההגדרה: עבור $a > 0$ הפונקציה $x \mapsto ax+b$ עולה ממש, ולכן היא **שומרת על הסדר** בין התצפיות — האיבר הגדול ביותר נשאר הגדול ביותר:

$$T(aX+b) = \max_i (aX_i + b) = a\cdot\max_i(X_i) + b = a\cdot T(X) + b \ \checkmark$$

כלומר המקסימום מקיים את ההגדרה הפורמלית — הוא תמצית למרכז חוקית (וגם מקיים $X_{(1)} \leq T \leq X_{(n)}$ באופן טריוויאלי). זה לא אומר שהוא תמצית *טובה*.

נקודת שבירה: מספיק לקלקל תצפית **אחת** — מזיזים אותה ל-$+\infty$ והמקסימום שואף ל-∞. לכן $\varepsilon^*_n = 1/n$, ואסימפטוטית: $1/n \to \mathbf{0}$.

💡 **הטריק/הלקח (וזו כל הפואנטה של התרגיל):** ההגדרה של תמצית למרכז תופסת רק **שקילות לטרנספורמציה לינארית עולה** — היא לא תופסת "אמצעיות". לכן דברים מוזרים כמו המקסימום עוברים אותה. ההבחנה בין תמציות מגיעה מנקודת השבירה: מקסימום חוקי אבל שביר לחלוטין (0), לעומת חציון עם נקודת שבירה ½ — הרובסטית ביותר האפשרית. המעבר המרכזי בהוכחות מהסוג הזה: **פונקציה עולה שומרת על סטטיסטי הסדר**, כלומר $(aX+b)_{(i)} = aX_{(i)}+b$ עבור $a>0$.

2.2 ממוצע משוקלל $T(X) = \Sigma w_i X_i / \Sigma w_i$, $w_i \geq 0$ — תמצית למרכז? כן

בדיקת ההגדרה (בהנחה $\Sigma w_i > 0$, אחרת לא מוגדר) — לינאריות פשוטה של הסכום:

$$T(aX+b) = \Sigma w_i(aX_i+b)/\Sigma w_i = (a\cdot\Sigma w_i X_i + b\cdot\Sigma w_i)/\Sigma w_i = a\cdot T(X) + b \ \checkmark$$

בנוסף, זהו צירוף קמור של התצפיות (מקדמים אי-שליליים שסכומם 1), ולכן תמיד $X_{(1)} \leq T(X) \leq X_{(n)}$ — עוד חיזוק לכך שזו תמצית מרכז סבירה. הממוצע הרגיל הוא המקרה הפרטי $w_i = 1$.

נקודת שבירה: קיים לפחות i אחד עם $w_i > 0$. היריב מחליף דווקא את התצפית הזו ב-$M \to \infty$, והתמצית מכילה את האיבר $w_i M/\Sigma w_j \to \infty$. לכן $\varepsilon^*_n = 1/n \to \mathbf{0}$ — בדיוק כמו הממוצע הרגיל.

💡 **חשוב ללמוד:** שקלול **לא** קונה רובסטיות: כל עוד למישהו יש משקל חיובי, קלקול שלו מפוצץ את התמצית. נקודת השבירה מוגדרת לפי **המקרה הגרוע ביותר** (היריב בוחר את מי לקלקל) — טעות נפוצה היא לחשוב שאם יש תצפיות עם משקל 0 אז התמצית מוגנת.

2.3 הטווח $T(X) = X_{(n)} - X_{(1)}$ — תמצית לפיזור? כן

אי-שליליות: תמיד $X_{(n)} \geq X_{(1)}$, ולכן $T \geq 0$ ✓.

שקילות לסקלה: נבדוק את שני המקרים. עבור $a > 0$ הסדר נשמר:

$$T(aX+b) = (aX_{(n)}+b) - (aX_{(1)}+b) = a(X_{(n)}-X_{(1)}) = |a|\cdot T(X) \ ✓$$

עבור $a < 0$ הסדר **מתהפך**: המקסימום החדש מגיע מהמינימום הישן ולהפך — $\max(aX+b) = aX_{(1)}+b$ ו-$\min(aX+b) = aX_{(n)}+b$:

$$T(aX+b) = (aX_{(1)}+b) - (aX_{(n)}+b) = a(X_{(1)}-X_{(n)}) = (-a)(X_{(n)}-X_{(1)}) = |a|\cdot T(X) \ ✓$$

נקודת שבירה: קלקול תצפית אחת ל-$+\infty$ שולח את המקסימום, ולכן את הטווח, ל-∞: $\varepsilon^*_n = 1/n \to \mathbf{0}$.

💡 **טריק נפוץ בהוכחות פיזור:** תמיד לפצל ל-$a>0$ ו-$a<0$, ולזכור שכפל במספר שלילי **הופך את הסדר**: $(aX)_{(i)} = aX_{(n+1-i)}$. שכחת המקרה $a<0$ היא הורדת הנקודות הקלאסית בשאלות האלה. וגם: ההזזה b תמיד מתבטלת בהפרש — לכן תמציות פיזור בנויות מהפרשים.

2.4 העשירון התחתון $T(X) = X_{0.1}$ — תמצית למרכז? כן

בדיקת ההגדרה: $X_{0.1} = X_{(k)}$ עם $k = \lceil 0.1n \rceil$ — סטטיסטי סדר. עבור $a > 0$ הסדר נשמר, ולכן:

$$T(aX+b) = (aX+b)_{(k)} = aX_{(k)} + b = a\cdot T(X) + b \ ✓$$

וגם $X_{(1)} \leq X_{(k)} \leq X_{(n)}$. כלומר, שוב — לפי ההגדרה הפורמלית זו תמצית למרכז חוקית, למרות שאינטואיטיבית העשירון התחתון בכלל לא "באמצע".

נקודת שבירה: כדי לשלוח את $X_{(k)}$ ל-$-\infty$ צריך שלפחות k תצפיות יהיו ב-$-\infty$, ומספיק לקלקל בדיוק k. (הכיוון השני, ל-$+\infty$, יקר יותר: דרושות $n-k+1 \approx 0.9n$ תצפיות.) לכן:

$$\varepsilon^*_n = \lceil 0.1n \rceil / n \quad \longrightarrow \quad \text{אסימפטוטית } \mathbf{0.1}$$

💡 **חשוב ללמוד (הכלל הכללי):** לשברון ה-p יש נקודת שבירה אסימפטוטית $\min(p, 1-p)$ — הצד הזול לקלקול קובע. מציבים $p=0.5$ ומקבלים את התוצאה המפורסמת: החציון עם נקודת שבירה ½, המקסימלית האפשרית. $p=0.1$ נותן 0.1. נוסחה ששווה לשנן — היא עונה על משפחה שלמה של שאלות.

2.5 $T(X) = X_{0.9} - X_{0.1}$ (הטווח הבין-עשירוני) — תמצית לפיזור? כן

אי-שליליות: $X_{0.9} = X_{(\lceil 0.9n \rceil)} \geq X_{(\lceil 0.1n \rceil)} = X_{0.1}$ כי $\lceil 0.9n \rceil \geq \lceil 0.1n \rceil$, ולכן $T \geq 0$ ✓.

הזזה: שני השברונים זזים ב-b וההפרש מבטל אותה ✓. **סקלה:** עבור $a>0$ שני השברונים נכפלים ב-a $\Longleftarrow$ $T(aX)=aT(X)$. עבור $a<0$ הסדר מתהפך והשברונים **מתחלפים בתפקידים** — העשירון העליון של הנתונים החדשים מגיע מהעשירון התחתון של הישנים:

$$(aX)_{0.9} = a\cdot X_{0.1}\ ,\quad (aX)_{0.1} = a\cdot X_{0.9} \implies T(aX) = a\cdot X_{0.1} - a\cdot X_{0.9} = (-a)(X_{0.9}-X_{0.1}) = |a|\cdot T(X)\ \checkmark$$

(עד כדי דקויות של ההגדרה עם inf בקצוות — ברמת הקורס זו התשובה.)

נקודת שבירה: הדרך הזולה ביותר לפוצץ: לשלוח את $X_{0.1}$ ל-$-\infty$ (עולה $[0.1n]$ קלקולים) או את $X_{0.9}$ ל-$+\infty$ (עולה $n-[0.9n]+1 \approx 0.1n$). בשני המקרים כ-10% מהתצפיות:

$$\varepsilon^*_n \approx 0.1 + O(1/n) \quad \longrightarrow \quad \text{אסימפטוטית } \mathbf{0.1}$$

השוואה ששווה לזכור: הטווח הרגיל (2.3) והטווח הבין-עשירוני מודדים "אותו דבר", אבל הראשון נשבר עם תצפית אחת (0) והשני דורש 10% מהנתונים (0.1). קיצוץ הזנבות הוא בדיוק המנגנון שקונה רובסטיות — אותו רעיון כמו IQR (עם 0.25) וממוצע קטום. שאלת השוואה כזו היא חומר קלאסי למבחן.

2.6 מרכז הטווח פחות החציון — תמצית לפיזור? לא

$$T(X) = (X_{(1)}+X_{(n)})/2 - \text{median}(X)$$

התמצית אמנם אדישה להזזה (שני האיברים שקולים למיקום וההפרש מבטל את b), אבל היא נכשלת בשתי דרישות:

1) אי-שליליות נכשלת — מספיקה דוגמה נגדית אחת: ניקח $X = (0, 9, 10)$. אז מרכז הטווח $= (0+10)/2 = 5$, החציון $= 9$, ולכן:

$$T(X) = 5 - 9 = -4 < 0\ \checkmark$$

2) גם שקילות הסקלה נכשלת: עבור $a < 0$ שני הרכיבים שקולים למיקום, ולכן $T(aX) = a\cdot T(X)$. כאשר $T(X) \neq 0$ זה שווה $-|a|T(X) \neq |a|T(X)$ ✗.

נקודת שבירה (לשם שלמות): גם אילו הייתה חוקית — קלקול תצפית אחת מפוצץ את מרכז הטווח, כלומר $1/n \to 0$.

מה ללמוד מכאן: הפרש בין **שתי תמציות מרכז** הוא לא תמצית פיזור — הוא מודד **א-סימטריה** (צידוד): חיובי כשהזנב הימני ארוך, שלילי כשהשמאלי ארוך. והטריק הטכני החשוב: כדי להפריך, לא מוכיחים — **בונים דוגמה נגדית מספרית קטנה** (3 תצפיות מספיקות). לחפש דוגמה נגדית לפני שמנסים להוכיח זה הרגל ששווה זהב במבחני נכון/לא נכון.

שאלה 1 — לציוני תקן ממוצע 0 וסטיית תקן 1

ציון התקן: $z_i = (x_i - \bar{x})/s$, כאשר s סטיית התקן המדגמית (ומניחים $s > 0$).

ממוצע: מלינאריות הסכום — הקבועים יוצאים החוצה, וסכום הסטיות מהממוצע הוא אפס:

$$\bar{z} = (1/n)\cdot\Sigma_i\,(x_i-\bar{x})/s = (1/s)\cdot[(1/n)\Sigma x_i - \bar{x}] = (1/s)\cdot(\bar{x} - \bar{x}) = 0$$

סטיית תקן: מציבים $\bar{z} = 0$ בהגדרת השונות המדגמית, ומזהים בפנים בדיוק את השונות של x:

$$s^2_z = (1/(n-1))\cdot\Sigma_i(z_i - \bar{z})^2 = (1/(n-1))\cdot\Sigma_i\, z_i^2 = (1/s^2)\cdot(1/(n-1))\cdot\Sigma_i(x_i-\bar{x})^2 = s^2/s^2 = 1$$

ולכן $s_z = 1$. ∎ (דרך שקולה: $z = (1/s)\cdot x - \bar{x}/s$ היא טרנספורמציה לינארית, ולפי הזהויות מהדף: הממוצע עובר לינארית והשונות נכפלת ב- $(1/s)^2$.)

> 💡 **חשוב ללמוד:** תקנון הוא בדיוק "להשתמש בכך שממוצע הוא תמצית למרכז וסטיית תקן תמצית לפיזור": מחסרים את המרכז ומחלקים בפיזור ⇐ מקבלים גודל **חסר יחידות** עם מרכז 0 ופיזור 1. שני המעברים שחוזרים בכל הוכחה כזו: $\Sigma(x_i-\bar{x}) = 0$ ולינאריות. זה גם הבסיס לסעיף הבא ולמשפט הגבול המרכזי.

שאלה 2 — קו הרגרסיה על ציוני תקן הוא $z_Y = r\cdot z_X$

מריצים רגרסיה פשוטה של z_Y על z_X ומשתמשים בנוסחאות הריבועים הפחותים שפותחו בדף (עבור נתונים כלשהם: שיפוע $b = s_{XY}/s^2_X$ וחותך $a = \bar{Y} - b\bar{X}$), כשמציבים את נתוני ה-z:

שלב 1 — השיפוע. המונה הוא השונות המשותפת של ציוני התקן. מכיוון ש-$\bar{z}_X = \bar{z}_Y = 0$ (שאלה 1), היא בדיוק ההגדרה של מקדם המתאם של פירסון (כפי שמופיע בדף — r הוא "הממוצע המתוקן של מכפלות ציוני התקן"):

$$s_{z_X z_Y} = (1/(n-1))\cdot\Sigma_i(z_{X_i}-\bar{z}_X)(z_{Y_i}-\bar{z}_Y) = (1/(n-1))\cdot\Sigma_i ((X_i-\bar{X})/s_X)\cdot((Y_i-\bar{Y})/s_Y) = r$$

המכנה הוא שונות ציוני התקן של X, ששווה 1 (שאלה 1). לכן:

$$b = s_{z_X z_Y} / s^2_{z_X} = r / 1 = r$$

שלב 2 — החותך. קו הריבועים הפחותים עובר תמיד דרך נקודת הממוצעים $(\bar{z}_X, \bar{z}_Y) = (0,0)$:

$$a = \bar{z}_Y - b\cdot\bar{z}_X = 0 - r\cdot 0 = 0$$

ולכן קו הרגרסיה: $\hat{z}_Y = r\cdot z_X$ ∎

> 💡 **למה זו תוצאה חשובה (ולא רק תרגיל):** (1) היא נותנת את **הפירוש האמיתי של r**: סטייה של סטיית תקן אחת ב-X מנבאת סטייה של r סטיות תקן ב-Y. (2) מכיוון ש-$|r| \le 1$, הניבוי תמיד **קרוב יותר לממוצע** מהתצפית — זו בדיוק "הנסיגה אל הממוצע" (regression to the mean), ומכאן השם רגרסיה! (3) חוזרים ליחידות המקוריות ומקבלים מיד את השיפוע הכללי: $b = r\cdot s_Y/s_X$. שלישיית עובדות שחוזרת במבחנים שוב ושוב.

טבלת סיכום

תמצית	חוקית?	נק׳ שבירה ε^*_n	אסימפטוטית
מקסימום $X_{(n)}$ (מרכז)	כן	1/n	0
ממוצע משוקלל (מרכז)	כן	1/n	0
טווח $X_{(n)}-X_{(1)}$ (פיזור)	כן	1/n	0
עשירון תחתון $X_{0.1}$ (מרכז)	כן	$\lceil 0.1n \rceil/n$	0.1
$X_{0.9} - X_{0.1}$ (פיזור)	כן	≈ 0.1	0.1
מרכז הטווח – חציון (פיזור)	**לא** (עלול להיות שלילי)	—	—

סיכום — הטריקים שחוזרים

- **$a>0$ שומר סדר, $a<0$ הופך סדר:** $(aX+b)_{(i)} = aX_{(i)}+b$ או $aX_{(n+1-i)}+b$ בהתאמה — המעבר המרכזי בכל הוכחות השקילות.
- **ההגדרות פורמליות, לא אינטואיטיביות:** מקסימום ועשירון עוברים את הגדרת ה"מרכז". מה שמבדיל תמצית טובה מרעה זו נקודת השבירה.
- **נקודת שבירה של שברון p:** $\min(p, 1-p)$; חציון = ½ (המקסימום האפשרי); ממוצע/מקסימום/טווח = 0.
- **הפרכה = דוגמה נגדית קטנה,** לא הוכחה. 3 תצפיות בדרך כלל מספיקות.
- **ציוני תקן:** $\Sigma(x_i-\bar{x})=0$ + לינאריות ⇐ ממוצע 0, ס״ת 1; ומהם $\hat{z}_Y = r \cdot z_X$ ⇐ הנסיגה אל הממוצע והשיפוע $r \cdot s_Y / s_X$.