נוסחאון

יחידה 5 – סטטיסטיקה תיאורית

סטטיסטיקה תיאורית (descriptive statistics) – ענף בסטטיסטיקה העוסק בשיטות לארגון, תיאור וסיכום הנתונים בטבלאות, בגרפים ועל ידי מדדים שונים בהתאם לסוג הנתונים ולשאלה שנשאלה.

משתנה בדיד (discrete variable) - משתנה כמותי היכול לקבל מספר סופי של ערכים בין כל שני ערכים אפשריים. למשל: מספר הנפשות במשפחה, מספר מכשירי T.V בדירה וכדומה.

משתנה רציף (continuous variable) - משתנה כמותי היכול לקבל אינסוף ערכים בין כל שני ערכים אפשריים. למשל: גובה, משקל, טמפרטורה וכדומה.

שכיחות (frequency) – מספר הפעמים שהתקבל ערך של המשתנה.
השכיחות מסומנת ב - $f(x)$

התפלגות שכיחויות (frequency distribution) - דרך לארגון נתונים בטבלה שבה מצוינים ערכי המשתנה ולצד כל ערך שכיחות הופעתו.

שכיחות יחסית (relative frequency) של מחלקה = שכיחות המחלקה מחולקת בגודל המדגם. $f(x)/n$

שכיחות מצטברת (cumulative frequency) - סה"כ השכיחות עד לגבול העליון של המחלקה. מסומנת ב- $F(x)$

צפיפות (density) – צפיפות היא השכיחות ליחידה של המשתנה הנחקר.
הצפיפות שווה לשכיחות המחלקה המחולקת ברוחבה. הצפיפות מסומנת ב- d_i
בהיסטוגרמה הצפיפות היא גובה המלבן.

דיאגרמת מקלות (bar diagram) – תיאור גרפי של התפלגות שכיחויות שאינה מקובצת למחלקות. בדיאגרמת מקלות ציר ה – x הוא ציר המשתנה וציר ה – y הוא ציר השכיחות. מעל ערכי ה - x השונים מעלים אנכים ("מקלות") בגובה השכיחות של ערכים אלו. גובה ה"מקל" מייצג את השכיחות.

היסטוגרמה (histogram) – הצגה גרפית של התפלגות שכיחויות עבור משתנה המקובץ במחלקות. הצגה זו בנויה ממלבנים כאשר המחלקה מיוצגת ע"י אורך של קטע – בסיס המלבן, והשכיחות מיוצגת ע"י שטח המלבן.

מדדי מיקום מרכזי (measures of central tendency)

שכיח (Mode) – ערך של המשתנה הנפוץ ביותר בהתפלגות. הערך ששכיחותו מקסימלית. מסומן ב- Mo

<u>חישוב השכיח בהתפלגות שכיחויות כשהמשתנה מקובץ במחלקות:</u>

כאשר רוחב המחלקות <u>שווה</u> – השכיח הינו <u>אמצע המחלקה</u> **<u>השכיחה</u>** ביותר.
כאשר רוחב המחלקות <u>אינו שווה</u> - השכיח הינו <u>אמצע המחלקה</u> **<u>הצפופה</u>** ביותר.

חציון (Median) – ערך המחלק את ההתפלגות לשניים: מחצית מהנתונים בהתפלגות (לפחות) קטנים ממנו או שווים לו ומחצית מהנתונים בהתפלגות (לפחות) גדולים ממנו או שווים לו. מסומן ב- Md.

<u>חישוב החציון:</u>

הערה: כדי לחשבו יש לסדר תחילה את הנתונים בסדר עולה.

<u>בסדרת ערכים או התפלגות שכיחויות של משתנה בדיד (לא מקובץ במחלקות):</u>

כשמספר הנתונים - N הוא **אי זוגי** , החציון הוא הערך המופיע במקום ה - $\frac{N+1}{2}$ בהתפלגות.

כשמספר הנתונים - N הוא **זוגי** , החציון הוא הממוצע החשבוני של שני הערכים המופיעים במקומות ה - $\frac{N}{2}$ וה - $\frac{N}{2}+1$ בהתפלגות.

<u>בהתפלגות שכיחויות של משתנה מקובץ במחלקות:</u>

$$Md=\frac{\frac{n}{2}-F(x_{m-1})}{f(x_m)}\cdot(L_1-L_0)+L_0$$

(הנוסחה מופיעה בעמוד 29)

- n - מספרם הכולל של המקרים
- x_m - המחלקה בה נמצא החציון.
- L_1 - גבול עליון אמיתי של מחלקה זו.
- L_0 - גבול תחתון אמיתי של מחלקה זו.
- x_{m-1} המחלקה הקודמת ל- x_m.
- $F(x_{m-1})$ השכיחות המצטברת עד למחלקה x_{m-1} (ועד בכלל).
- $f(x_m)$ שכיחות המחלקה x_m

הערות:

- בנוסחה הנ"ל אין הבחנה בין N זוגי לאי זוגי.
- כדי להשתמש בנוסחה יש לעבוד בגבולות אמיתיים.
- כאשר השכיחויות נתונות בערכים יחסיים נציב: N=1 והשכיחויות f,F יצוינו בשבר עשרוני.

ממוצע חשבוני (Aritmatic Mean) – סכום הנתונים מחולק במספר הנתונים.

נוסחאות חישוב:

<u>בסדרת ערכים:</u>

$$\bar{x} = \frac{x_1 + x_2 + \ldots + x_n}{n} = \frac{\sum_{i=1}^{n} x_i}{n} \qquad \sum_{i=1}^{n} x_i = n \cdot \bar{x}$$

<u>בהתפלגות שכיחויות של משתנה בדיד לא מקובץ:</u>

$$\bar{x} = \frac{x_1 \cdot f(x_1) + x_2 \cdot f(x_2) + \ldots + x_k \cdot f(x_k)}{N} = \frac{\sum_{i=1}^{k} x_i \cdot f(x_i)}{N}$$

כאשר $N = f(x_1) + f(x_2) + \ldots + f(x_k)$

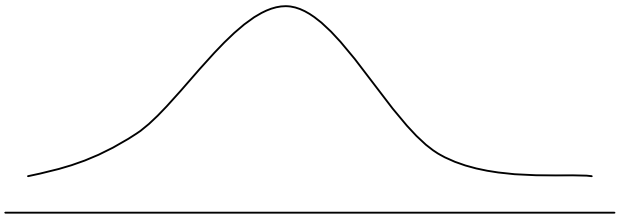
כאשר המשתנה מקובץ במחלקות מבטא ה- x בנוסחה הנ"ל את אמצע המחלקה.

אמצע הטווח – הממוצע בין התצפית הגדולה ביותר והקטנה ביותר בהתפלגות.

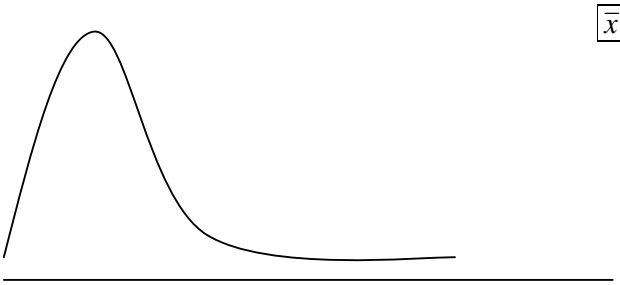
$$MR = \frac{X_{\max} + X_{\min}}{2}$$

<u>סוגי התפלגויות</u>

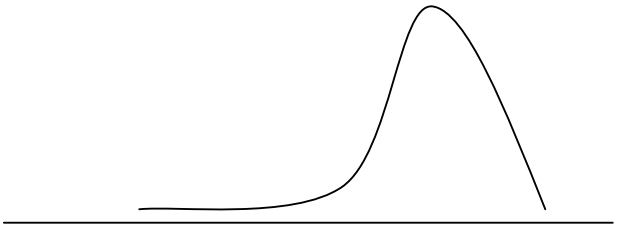
1. **התפלגות סימטרית פעמונית** – רוב התצפיות מרוכזות במרכז ההתפלגות ומעט מהן באיזורי השוליים. פיזור התצפיות סימטרי משני עברי המרכז. מיקום המדדים: כל מדדי המיקום המרכזי מתלכדים: $\bar{x} = Md = Mo$



2. **התפלגות אסימטרית חיובית (ימנית)** - מרבית התצפיות מרוכזות בקטע מסוים ויש <u>מעט</u> תצפיות חריגות כלפי ערכים גבוהים של x.
מיקום המדדים: $\bar{x} > Md$



3. **התפלגות אסימטרית שלילית (שמאלית)** - מרבית התצפיות מרוכזות בקטע מסוים ויש <u>מעט</u> תצפיות חריגות כלפי ערכים נמוכים של x.
מיקום המדדים: $\boxed{\bar{x} < Md}$



	ממוצע	חציון	שכיח
1	מבוסס על כל הנתונים בהתפלגות	אינו מבוסס על כל הנתונים בהתפלגות	אינו מבוסס על כל הנתונים בהתפלגות.
2	מושפע ע"י ערכים קיצוניים	מושפע מעט ע"י ערכים קיצוניים	אינו מושפע ע"י ערכים קיצוניים
3	סכום הסטיות מהממוצע שווה תמיד לאפס		
4	אינו ניתן לחישוב כשיש מחלקה/ות פתוחה/ות	לרוב ניתן לחישוב כשיש מחלקה/ות פתוחה/ות	לרוב ניתן לחישוב כשיש מחלקה/ות פתוחה/ות
5	טוב במיוחד באוכלוסיה **סימטרית** בה כל המדדים מתלכדים מאחר שהוא הנוח ביותר לחישוב	מועדף כשהתפלגות האוכלוסיה **אסימטרית** עם נטיה חזקה לכיוון מסוים.	מועדף כשרוצים להדגיש את הערך הטיפוסי והנפוץ ביותר בהתפלגות. מתאים גם עבור משתנה לא כמותי (מגדר, מצב משפחתי, צבע שיער וכו').

מדדי פיזור (measures of variability) - מדדים המתארים את מידת השוני או הגיוון בין הנתונים בהתפלגות. מתחלקים ל- 2 סוגים עיקריים:

א. מדדי פיזור אבסולוטיים המודדים את טווח הפיזור.
ב. מדדי פיזור אבסולוטיים המבוססים על סטיה מן המרכז.

- כל מדדי הפיזור מקבלים רק ערכים אי שליליים.
- מדדי הפיזור יקבלו ערך אפס כאשר כל הנתונים בהתפלגות זהים.
- ככל שמידת השוני בין הנתונים גדולה יותר, מדדי הפיזור יקבלו ערך גבוה יותר.
- הוספת גודל קבוע לכל הנתונים בהתפלגות (או החסרת קבוע) כמוה כהזזה של כל הנתונים בהתפלגות באותו גודל ואינה משפיעה על ערכם של מדדי הפיזור.

א. <u>מדדי פיזור המודדים את טווח הפיזור</u>

1. **טווח** (Range) – המרחק/ההפרש בין הערך הגבוה ביותר בהתפלגות לבין הערך הנמוך ביותר בהתפלגות. סימון: R

$$R = X_{\max} - X_{\min}$$

2. **טווח בין רבעוני**: תחום המשתרע מהרבעון התחתון (Q_1) ועד לרבעון העליון (Q_3). תחום זה כולל את ערכיהן של 50% הנתונים הנמצאים במרכז ההתפלגות סביב החציון. סימון: $IQR = Q_3 - Q_1$

$$Q_1 = \frac{\frac{n}{4} - F(x_{m-1})}{f(x_m)} \cdot (L_1 - L_0) + L_0$$

$$Q_3 = \frac{\frac{3n}{4} - F(x_{m-1})}{f(x_m)} \cdot (L_1 - L_0) + L_0$$

ב. <u>מדדי פיזור המבוססים על סטיה ממדד מרכזי</u>

שונות (variance) - ממוצע ריבועי הסטיות של כל הנתונים מממוצע ההתפלגות. השונות מסומנת ב - S_x^2.

חישוב השונות:

<u>בסדרת ערכים:</u>

$$S_x^2 = \frac{\sum_{i=1}^{n} x_i^2}{n} - \bar{x}^2 \qquad S_x^2 = \frac{\sum_{i=1}^{n} (x_i - \bar{x})^2}{n}$$

<u>בטבלת שכיחויות:</u>

$$S_x^2 = \frac{\sum_{i=1}^{k} x_i^2 \cdot f(x_i)}{n} - \bar{x}^2 \qquad S_x^2 = \frac{\sum_{i=1}^{k} f(x_i) \cdot (x_i - \bar{x})^2}{n}$$

הערה: השונות אינה נמדדת ביחידות המדידה המקוריות של המשתנה אלא ביחידות ריבועיות.

סטיית התקן (Standard Deviation) – השורש הריבועי של השונות. סימון: S_X
סטיית התקן הינה מדד הפיזור השימושי והנפוץ ביותר. היא מבטאת **בקירוב** את המרחק הממוצע של הנתונים מממוצע ההתפלגות.

<u>חישוב סטיית התקן</u>

כדי לחשב את סטיית התקן רצוי לחשב תחילה את השונות ומהתוצאה המתקבלת להוציא שורש ריבועי.

הערות:

א. סטיית התקן מתחשבת בסטיות כל הנתונים מן הממוצע ונותנת משקל יחסי מתאים לכל ערך בהתאם למרחקו מן הממוצע ובהתאם לשכיחותו. כל סטיה תורמת כריבוע גודלה, כך שסטיות גדולות יותר השפעתן על המדד גדולה יותר.

ב. סטיית התקן נמדדת באותן יחידות מידה כמו המשתנה.

השפעת טרנספורמציה לינארית על מדדי המיקום המרכזי ומדדי הפיזור

א. **הוספת/הפחתת קבוע** A ל/מ כל הערכים בהתפלגות גורמת להגדלת/להקטנת כל מדדי המיקום המרכזי **ב** – A.
<u>מדדי הפיזור **אינם מושפעים** מהוספת או הפחתת קבוע!</u>

ב. **הכפלת** כל הערכים בהתפלגות בקבוע חיובי B גורמת **להכפלת** כל מדדי המיקום המרכזי **ב-** B. גם מדדי הפיזור שצוינו לעיל, למעט השונות, יוכפלו **ב-** B.
השונות תוכפל **ב** – B^2 (כי היא נמדדת ביחידות ריבועיות).

ממוצע משוקלל ושונות מצורפת של שתי קבוצות

נתונות שתי קבוצות וידועים הגדלים הבאים:

I	II
n_1	n_2
$\bar{x}_1$	$\bar{x}_2$
s_1^2	s_2^2

<u>ממוצע משוקלל של שתי הקבוצות יחד:</u>

$$\bar{\bar{x}} = \frac{n_1 \cdot \bar{x}_1 + n_2 \cdot \bar{x}_2}{n_1 + n_2}$$

<u>שונות מצורפת של שתי הקבוצות:</u>

$$S_c^2 = \frac{n_1 \cdot \left(\bar{x}_1^2 + s_1^2\right) + n_2 \cdot \left(\bar{x}_2^2 + s_2^2\right)}{n_1 + n_2} - \bar{\bar{x}}^2$$

נוסחה חלופית:

$$S_c^2 = \frac{n_1 \cdot s_1^2 + n_2 \cdot s_2^2}{n_1 + n_2} + \frac{n_1 \cdot \left(\bar{x}_1 - \bar{\bar{x}}\right)^2 + n_2 \cdot \left(\bar{x}_2 - \bar{\bar{x}}\right)^2}{n_1 + n_2}$$

הערות

א. ניתן להכליל את הנוסחאות הנ"ל גם ל- 3 קבוצות ויותר בהתאמה.

ב. כאשר גודלי הקבוצות שווים ($n_1=n_2$), הממוצע המשוקלל הוא הממוצע הפשוט של שתי הקבוצות.

ג. כאשר ממוצעי שתי הקבוצות שווים, גם הממוצע המשוקלל שווה להם. השונות המצורפת תהיה במקרה זה שונות משוקללת: $S_c^2 = \frac{n_1 \cdot s_1^2 + n_2 \cdot s_2^2}{n_1 + n_2}$

מדדי מיקום יחסי

מאון[1] (Percentile) – המאון ה k – של הערך C מסומן ב - K_C ומוגדר **כאחוז** הנתונים בהתפלגות הקטנים מ- C או שווים לו.
הערך C שמתחתיו נמצאים K% מהאוכלוסיה מסומן ב - C_K.

<u>נוסחאות לחישוב (עבור טבלת שכיחויות של משתנה מקובץ במחלקות):</u>

$$C_K = \frac{\frac{n \cdot k}{100} - F\left(x_{m-1}\right)}{f\left(x_m\right)} \cdot \left(L_1 - L_0\right) + L_0$$

$$K_C = \left[\frac{C_K - L_0}{L_1 - L_0} \cdot f + F\right] \cdot \frac{100}{n}$$

שיטת החישוב בעזרת הנוסחה זהה לזו שבנוסחת החציון בטבלת שכיחויות של משתנה מקובץ (ראה לעיל).

ציון תקן (Standard score) – ציון התקן של X מסומן ב – Z_x ומבטא בכמה סטיות תקן רחוק X מממוצע ההתפלגות.

<u>נוסחה לחישוב</u>

$$Z_x = \frac{x - \bar{x}}{s_x}$$

תכונות:

1. ציון התקן חסר יחידות מידה (מספר טהור).
2. ממוצע ציוני התקן שווה לאפס , שונות ציוני התקן שווה ל – 1.
3. $Z_X>0$ **אם** $x > \bar{x}$

[1] נקרא גם אחוזון.

$Z_X=0$ **אם** $x=\bar{x}$

$Z_X<0$ **אם** $x<\bar{x}$

4. ניתן להשוות בעזרתו גדלים הנמדדים ביחידות שונות (למשל: משקל עם גובה).

מדדי קשר

מקדם המתאם של פירסון – r

$$r=\frac{\operatorname{cov}(x,y)}{s_x\cdot s_y}=\frac{\sum_{i=1}^{n}(x_i-\bar{x})(y_i-\bar{y})}{n\cdot s_x\cdot s_y}=\frac{\sum_{i=1}^{n}x_iy_i-n\cdot\bar{x}\cdot\bar{y}}{\sqrt{\left(\sum_{i=1}^{n}x_i^2-n\bar{x}^2\right)\left(\sum_{i=1}^{n}y_i^2-n\bar{y}^2\right)}}=$$

$$\frac{n\cdot\sum_{i=1}^{n}x_iy_i-\left(\sum_{i=1}^{n}x_i\right)\cdot\left(\sum_{i=1}^{n}y_i\right)}{\sqrt{\left[n\cdot\sum_{i=1}^{n}x_i^2-\left(\sum_{i=1}^{n}x_i\right)^2\right]\left[n\cdot\sum_{i=1}^{n}y_i^2-\left(\sum_{i=1}^{n}y_i\right)^2\right]}}$$

<u>תכונות</u>

1. $-1\le r\le+1$
2. $r=+1$ הקשר בין x ל- y חיובי ומושלם.
3. $r=-1$ הקשר בין x ל- y שלילי ומושלם.
4. ככל ש- r מתקרב בערכו המוחלט ל -1 עוצמת הקשר הלינארי בין שני המשתנים עולה.
5. <u>השפעת טרנספורמציה לינארית על מקדם המתאם:</u>

נתונים שני משתנים X ו- Y. ומקדם המתאם ביניהם: $r_{X,Y}$

נגדיר : $X'=b\cdot X+a$ ו - $Y'=c\cdot Y+d$ (טרנספורמציה לינארית)

אזי מקדם המתאם ביניהם (לאחר הטרנספורמציה\יות) יהיה: $r_{X'Y'}=\frac{b\cdot c}{|b\cdot c|}\cdot r_{X,Y}$

השונות המשותפת של x ו- y - $\operatorname{cov}(x,y)$

<u>נוסחאות חישוב:</u>

$$\operatorname{cov}(x,y)=\frac{\sum_{i=1}^{n}(x_i-\bar{x})\cdot(y_i-\bar{y})}{n}=\frac{\sum_{i=1}^{n}x_iy_i}{n}-\bar{x}\cdot\bar{y}$$

תכונות:

1. **אם** $\operatorname{cov}(x,y)>0$ **אזי** $r>0$

   אם $\operatorname{cov}(x,y)<0$ **אזי** $r<0$

   אם $\operatorname{cov}(x,y)=0$ **אזי** $r=0$
2. $\operatorname{cov}(x,x)=S_X^2$

יחידה 6 – הסתברות במרחב בדיד

תורת ההסתברות (Probability Theory) – תורה מתמטית לטיפול בתופעות הכרוכות באי וודאות, כגון: הטלת מטבע, הטלת קוביה וכדומה.

ניסוי מקרי (Random experiment) – ניסוי שבו אנו מכירים את כל התוצאות האפשריות, אך לא ניתן לדעת מראש איזו מן התוצאות האפשריות תתרחש בפועל.

מרחב המדגם (Sample space) – אוסף כל התוצאות האפשריות של ניסוי מקרי. (סימון: Ω)

מאורע (Event) – קבוצה חלקית כלשהי של תוצאות ניסוי. (סימון: A, B,...)

<u>פעולות בין מאורעות</u>

נתונים שני מאורעות A ו- B. נגדיר:

איחוד המאורעות – יסומן ב - $A \cup B$, הוא המאורע המורכב מכל תוצאות הניסוי השייכות ל - A ו/או ל – B. כלומר מורכב מתוצאות השייכות לפחות לאחד משני המאורעות A , B.

חיתוך המאורעות - יסומן ב - $A \cap B$, הוא המאורע המורכב מכל תוצאות הניסוי המשותפות לשני המאורעות A , B.

המשלים של מאורע A - יסומן ב - $\overline{A}$, הוא המאורע המורכב מכל תוצאות הניסוי שאינן **ב-** A.

מאורעות זרים - A ו- B ייקראו מאורעות זרים אם ורק אם אין להם אף תוצאה משותפת. כלומר החיתוך שלהם ריק $A \cap B = \Phi$.
(מאורעות זרים אינם יכולים להתרחש בו זמנית באותו ניסוי, הם מאורעות המוציאים זה את זה)

מאורעות מוכלים ומאורעות זהים

A מוכל ב – B אם כל תוצאות הניסוי הכלולות ב- A , כלולות ב - B .
סימון: $A \subset B$

A זהה ל– B אם כל תוצאות הניסוי הכלולות ב- A , כלולות ב – B , ואם כל תוצאות הניסוי הכלולות ב- B , כלולות ב – A.
סימון: אם $A \subset B$ ו- $B \subset A$ אז $A = B$

חוקי דה מורגן:

"משלים האיחוד שווה לחיתוך המשלימים" - $\boxed{A^C \cap B^C = (A \cup B)^C}$

"משלים החיתוך שווה לאיחוד המשלימים" - $\boxed{A^C \cup B^C = (A \cap B)^C}$

הסתברות של מאורע

פונקציית ההסתברות מתאימה לכל מאורע A מספר ממשי $P(A)$ המבטא את מידת הסבירות (הסיכוי) שהמאורע A יתרחש כתוצאה מניסוי הסתברותי.

אקסיומות פונקציית ההסתברות והחוקים הנגזרים מהן:

1. $\boxed{0 \le P(A) \le 1}$ לכל A
2. $\boxed{P(\Omega) = 1}$
3. חוק החיבור: אם A ו- B מאורעות **<u>זרים</u>** אז $\boxed{P(A \cup B) = P(A) + P(B)}$

<u>חוקים נגזרים</u>

4. הכללה של חוק החיבור: אם $A_1, A_2, ..., A_n$ מאורעות זרים (בזוגות) אז מתקיים: $\boxed{P(A_1 \cup A_2 \cup ... \cup A_n) = P(A_1) + P(A_2) + ... + P(A_n)}$
5. חוק המשלים: $\boxed{P(A^C) = 1 - P(A)}$
6. $P(\Phi) = 0$
7. אם $A \subseteq B$ אז $P(A) \le P(B)$
8. חוק האיחוד לכל שני מאורעות A ו- B: $\boxed{P(A \cup B) = P(A) + P(B) - P(A \cap B)}$
9. חוקי ההפחתה\ההפרש :

$$\boxed{P(A \setminus B) = P(A \cap B^C) = P(A) - P(A \cap B)}$$

10. חוקי דה מורגן:

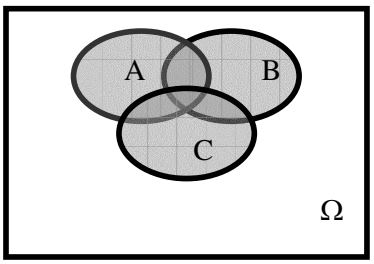
$$\boxed{P(A^C \cap B^C) = P(A \cup B)^C = 1 - P(A \cup B)}$$

$$\boxed{P(A^C \cup B^C) = P(A \cap B)^C = 1 - P(A \cap B)}$$

תרשים 2X2

A/B	A	A^C	
B	$P(A\cap B)$	$P(A^C\cap B)$	$P(B)$
B^C	$P(A\cap B^C)$	$P(A^C\cap B^C)$	$P(B^C)$
	$P(A)$	$P(A^C)$	**1**

נוסחת האיחוד של שלושה מאורעות



$$P(A\cup B\cup C)=P(A)+P(B)+P(C)-P(A\cap B)-P(A\cap C)-P(B\cap C)+P(A\cap B\cap C)$$

נוסחה זו תקפה **לכל** שלושה מאורעות (השרטוט נועד רק להמחשת מקרה פרטי)

מרחב הסתברות אחיד (סימטרי)

מרחב מדגם ייקרא אחיד (סימטרי) אם הוא מכיל מס' סופי של תוצאות ולכולן סיכוי שווה לקרות (שוות הסתברות).

$$\Omega=\{\omega_1,\omega_2,....,\omega_n\}$$

$$P(\omega_1)=P(\omega_2)=...=P(\omega_n)=\frac{1}{n}$$

<u>הסתברות של מאורע במרחב הסתברות אחיד</u>

$$P(A)=\frac{n(A)}{n(\Omega)}$$

$n(A)$ – מס' התוצאות הכלולות ב- A

$n(\Omega)$ – מס' התוצאות הכלולות ב- Ω

קומבינטוריקה

עקרון הכפל

אם ניתן לבצע ניסוי ב – K שלבים בזה אחר זה כך ש:

בשלב מס' 1 יש n_1 תוצאות אפשריות
בשלב מס' 2 יש n_2 תוצאות אפשריות
.
.
.
בשלב מס' k יש n_k תוצאות אפשריות
אז לניסוי ה – k שלבי כולו יש בסך הכל $\boxed{n_1 \cdot n_2 \cdot \ldots \cdot n_k}$ תוצאות אפשריות שונות.

(תוצאה אפשרית של הניסוי ה –k שלבי היא רשימה סדורה של תוצאות כל k השלבים.)

דגימה מאוכלוסיה סופית

נתונה אוכלוסיה בת N איברים שונים נבחר מתוכה k איברים – מתקבל **מדגם** בגודל k מהאוכלוסיה.

מדגם סדור

מדגם סדור בגודל k מאוכלוסיה בת n איברים שונים הוא שורה מסודרת של k איברים שנבחרו מאיברי האוכלוסיה.

במדגם סדור סדר הוצאתם של איברי המדגם משנה, כלומר 2 תוצאות באותו הרכב אך בסדר שונה נחשבות לתוצאות שונות (למשל ab ו- ba 2 תוצאות שונות)

מדגם לא סדור

מדגם לא סדור בגודל k הנבחר ללא החזרה מאוכלוסיה בת n איברים שונים הוא קבוצה חלקית של k איברים שנבחרו מתוך איברי האוכלוסיה.

במדגם לא סדור סדר הוצאתם של איברי המדגם אינו משנה, כלומר 2 תוצאות באותו הרכב אך בסדר שונה נחשבות לתוצאה אחת (למשל ab ו- ba תוצאה אחת)

דגימה עם\ללא החזרה

אם כל איבר שנבחר למדגם **מוחזר** לאוכלוסיה בטרם נבחר האיבר הבא (ולכן יכול לחזור ולהיבחר שנית) נאמר שהמדגם הוצא **עם החזרה**.

אם איבר שנבחר למדגם **אינו מוחזר** לאוכלוסיה בטרם נבחר האיבר הבא (ולכן אינו יכול להיבחר שנית) נאמר שהמדגם הוצא **ללא החזרה**.
(במקרה כזה כל איברי המדגם שונים זה מזה)

מספר המדגמים השונים בגודל k שניתן להוציא מאוכלוסיה בת N איברים:

	מדגמים סדורים	**מדגמים לא סדורים**
עם החזרה	$\underbrace{N\cdot N\cdot\ldots\cdot N}_{k}=N^k$	לא נלמד בקורס
ללא החזרה $(k\le n)$	$(N)_k=N\cdot(N-1)\cdot(N-2)\cdot\ldots\cdot(N-K+1)$ $=\frac{N!}{(N-K)!}$	$\binom{N}{K}=\frac{(N)_K}{K!}=\frac{N!}{K!\cdot(N-K)!}$

(תזכורת: $n!=n\cdot(n-1)\cdot\ldots\cdot2\cdot1$ וכן $0!=1$)

תמורות

נתונה אוכלוסיה בת N איברים שונים.
כל סידור של N האיברים בשורה נקרא תמורה.

מספר הסידורים השונים של N האיברים השונים בשורה (מס' התמורות)

הוא: $\boxed{n\cdot(n-1)\cdot\ldots\cdot2\cdot1=n!}$

כל שתי תמורות מתוך n! התמורות האפשריות שונות זו מזו רק בסדר האיברים ולא בזהותם.

תמורות במעגל - מספר התמורות של n איברים שונים במעגל הוא $\boxed{(n-1)!}$

תמורות עם איברים זהים

מספר האפשרויות לסדר n איברים בשורה שמתוכם n_1 איברים זהים מסוג ראשון, n_2 איברים זהים מסוג שני, ... , n_k איברים מסוג k, (כאשר $n_1+n_2+\ldots+n_k=n$) הוא:

$$\boxed{\frac{n!}{n_1!\cdot n_2!\cdot\ldots\cdot n_k!}}$$

מקרה פרטי: מספר הסידורים השונים בשורה של n איברים שמהם k זהים מסוג I ו- n-k זהים מסוג II הוא: $\binom{n}{k}=\frac{n!}{(n-k)!\cdot k!}$

הפילוג ההיפר- גיאומטרי

באוכלוסיה בת N פרטים, שמהם R פרטים המקיימים תכונה מסוימת (שייקראו להלן "מיוחדים"), מוציאים מהאוכלוסיה ללא החזרה n פרטים ובודקים את מספר ה"מיוחדים" שהתקבלו.

ההסתברות שהתקבלו בדיוק k "מיוחדים" (k=0,1,..n) היא: $\boxed{\frac{\binom{R}{k}\cdot\binom{N-R}{n-k}}{\binom{N}{n}}}$

יחידה 7 - הסתברות מותנית, נוסחת ההסתברות השלמה ונוסחת בייז

הסתברות מותנית (Conditional probability)

יהיו A ו- B שני מאורעות ב - Ω. ההסתברות המותנית של מאורע A כאשר ידוע שמאורע B קרה מסומנת ב - $P(A \mid B)$ ומוגדרת: $\boxed{P(A \mid B) = \frac{P(A \cap B)}{P(B)}}$ $(P(B) > 0)$

(B מהווה מרחב מדגם מצומצם $B = \Omega'$)

הערות

א. $0 \le P(A \mid B) \le 1$

ב. $P(A \mid B) + P(A^C \mid B) = 1$

ג. $P(A \mid B)$ יכולה להיות גדולה, קטנה או שווה ל - $P(A)$

אם: $P(A \mid B) = P(A)$ נאמר שהמאורעות A ו- B בלתי תלויים. (ראה בהמשך)

ד. הקשר $P(A \mid B) = P(B \mid A)$ מתקיים רק במקרים הבאים:

$P(A) = P(B)$ או A ו- B זרים. כאשר A ו- B זרים מתקיים: $P(A \mid B) = P(B \mid A) = 0$

נוסחת המכפלה (עקרון החישוב בשלבים)

מתוך הגדרת ההסתברות המותנית מתקבלת נוסחה לחישוב הסתברות של חיתוך שני מאורעות (או יותר). נוסחה זו שימושית במיוחד כאשר ניתן להתייחס לניסוי המקרי כמו לניסוי המתבצע בשלבים.

$$\boxed{P(A \cap B) = P(A) \cdot P(B \mid A)}$$

$$\boxed{P(A \cap B) = P(B) \cdot P(A \mid B)}$$

כלל השרשרת

את נוסחת המכפלה ניתן להכליל ליותר משני מאורעות:

$$\boxed{P(A_1 \cap A_2 \cap ... \cap A_n) = P(A_1) \cdot P(A_2 \mid A_1) \cdot P(A_3 \mid A_1 \cap A_2) \cdot ... \cdot P(A_n \mid A_1 \cap ... \cap A_{n-1})}$$

נוסחת ההסתברות השלמה (Total probability formula)

אם מרחב המדגם Ω מתפרק לאיחוד של מאורעות זרים $A_1, A_2, ... A_n$ $\left(\sum_{i=1}^{n} P(A_i) = 1\right)$ אז לכל מאורע B מתקיים:

$$\boxed{P(B) = \sum_{i=1}^{n} P(A_i) \cdot P(B \mid A_i) = P(A_1) \cdot P(B \mid A_1) + P(A_2) \cdot P(B \mid A_2) + ... + P(A_n) \cdot P(B \mid A_n)}$$

ההסתברות של B ניתנת כממוצע משוקלל של ההסתברויות המותנות - $P(B \mid A_i)$, כאשר המשקלות הן הסתברויות המאורעות A_i המתאימים.

אסטרטגיית עץ לחישוב הסתברויות בניסוי בשלבים

א. משורש העץ מוציאים מספר ענפים כמספר האפשרויות השונות בשלב א' של הניסוי. על כל ענף רושמים את ההסתברות לקבלת התוצאה המתאימה.

ב. מקצה כל ענף שהתקבל בשלב א', מוציאים שוב ענפים כמספר האפשרויות השונות בשלב ב'. על כל ענף כזה רושמים את ההסתברות **המותנית** לקבלת תוצאה זו בשלב ב', לאור התוצאה שאליה הגענו בשלב א'. כך ממשיכים עבור כל שלבי הניסוי.

ג. "גובה" העץ הוא כמספר שלבי הניסוי. כל מסלול מייצג תוצאה בודדת של הניסוי המשולב כולו. לפי חוק המכפלה, ההסתברות של כל תוצאה כזאת שווה **למכפלת** כל ההסתברויות הרשומות על הענפים לאורך המסלול המתאים לתוצאה זו.

ד. ההסתברות של כל מאורע היא **סכום** ההסתברויות של כל התוצאות (מסלולים) המתאימים למאורע זה.

נוסחת בייז (Bayes formula)

אם מרחב המדגם Ω מתפרק לאיחוד של מאורעות זרים $A_1, A_2, ...A_n$ $\left(\sum_{i=1}^{n} P(A_i) = 1\right)$ ו - B מאורע כלשהו **ב-** Ω **אזי מתקיים:**

$$P(A_k \mid B) = \frac{P(A_k \cap B)}{P(B)} = \frac{P(A_k) \cdot P(B \mid A_k)}{P(B)} = \frac{P(A_k) \cdot P(B \mid A_k)}{\sum_{i=1}^{n} P(A_i) \cdot P(B \mid A_i)}$$

ההסתברויות $P(A_1), P(A_2), ..., P(A_k)$ של השלב הראשון בניסוי נקראות **הסתברויות אפריוריות**.
ההסתברויות $P(A_1 \mid B), P(A_2 \mid B), ..., P(A_k \mid B)$ של השלב הראשון בניסוי לאור תוצאה מסוימת B בשלב השני נקראות **הסתברויות אפוסטריוריות**.

נוסחת בייס משמשת לעדכון ההסתברויות $P(A_i)$ לאור המידע שהמאורע B התרחש. לחישוב ההערכות החדשות של $P(A_i \mid B)$ משתמשים בהערכות הראשונות של ההסתברויות, דהיינו ב- $P(A_i)$ - ים ובהסתברויות המותנות $P(B \mid A)$

אי תלות של מאורעות

מאורעות בלתי תלויים (Independent events) – שני מאורעות A ו- B נקראים בלתי תלויים (ב"ת) אם: $\boxed{P(A\cap B)=P(A)\cdot P(B)}$

אם התנאי הנ"ל לא מתקיים נאמר שהמאורעות תלויים.

הגדרה חלופית: נאמר שמאורע A בלתי תלוי ב – B אם הידע בדבר התרחשות (או אי התרחשות) המאורע B אינו משנה (משפיע) את ההסתברות שהמאורע A יתרחש. כלומר: $P(A\mid B)=P(A)$

הערות

א. אם A בלתי תלוי ב – B אזי גם B בלתי תלוי ב – A. (תנאי סימטרי)

ב. אם A ו - B שני מאורעות בלתי תלויים אזי כל אחד מהזוגות:

A , B^C

B , A^C

B^C , A^C

הם מאורעות בלתי תלויים.

הקשר בין מאורעות בלתי תלויים למאורעות זרים

A ו- B **הם מאורעות זרים אם** $A\cap B=\Phi$ **ואז** $P(A\cap B)=0$

A ו- B **הם מאורעות בלתי תלויים אם** $P(A\cap B)=P(A)\cdot P(B)$

אם A ו - B הם שני מאורעות **זרים** אז הם בהכרח **תלויים**.
אם A ו - B הם שני מאורעות **בלתי תלויים** אז הם בהכרח **לא זרים**.

יחידה 8 - משתנה מקרי בדיד

משתנה מקרי בדיד (Discrete random variable)

משתנה מקרי (מ"מ) הוא פונקציה מספרית המוגדרת על מרחב המדגם.
טווח של מ"מ x: קבוצת הערכים שהמ"מ x יכול לקבל.

רשימת הערכים הנמצאים בטווח של משתנה מקרי בדיד X בצירוף ההסתברויות לכל ערך נקרא **פונקצית הסתברות של משתנה מקרי (מ"מ) X.**

<u>תכונות פונקצית ההסתברות של מ"מ בדיד X</u>

X	x_1	x_2	⋯	⋯	x_k	
$P(x)$	$P(x_1)$	$P(x_2)$			$P(x_k)$	**1**

$I \quad 0 \le P(x_i) \le 1$

$II \quad \sum_{i=1}^{k} P(x_i) = 1$

תוחלת של משתנה מקרי בדיד

תוחלת היא **ממוצע** הערכים שהמשתנה X מקבל המשוקללים בהתאם להסתברויותיהם.
סימון: $E(X)$

<u>נוסחה לחישוב התוחלת:</u>

$$E(X) = \sum_{i=1}^{k} x_i \cdot P(x_i) = x_1 \cdot P(x_1) + x_2 \cdot P(x_2) + \ldots + x_k \cdot P(x_k)$$

<u>תכונות התוחלת:</u>

1. $E(a) = a$ (תוחלת של משתנה המקבל סיכוי קבוע a בוודאות הינה a)
2. $E(X + a) = E(X) + a$ (תוספת קבועה לכל אחד מערכי המשתנה תגרום לתוספת זהה לתוחלת)
3. $E(b \cdot X) = b \cdot E(X)$ (הכפלת כל ערכי המשתנה בקבוע b תגרום להכפלת התוחלת בקבוע b)
4. $E(bX + a) = bE(X) + a$ (תכונת הלינאריות)
5. $E(X + Y) = E(X) + E(Y)$ (תוחלת של סכום שני מ"מ שווה לסכום התוחלות שלהם)

שונות של מ״מ

השונות של מ״מ X מסומנת[1] ב -$V(X)$ ומחושבת בעזרת אחת הנוסחאות הבאות:

$$V(X)=E(X^2)-[E(X)]^2 \qquad or \qquad V(X)=\sum_{i=1}^{k}(x_i-EX)^2\cdot P(x_i)$$

כאשר $E(x^2)=\sum_{i=1}^{k}x_i^2\cdot P(x_i)$ *(תוחלת של פיזורי סביבי המשתנה)*

סטיית התקן

סטיית התקן של מ״מ X היא השורש הריבועי של השונות $\sigma_x=\sqrt{V(x)}$.

<u>תכונות השונות:</u>

1. $V(X)\geq 0$ **לכל מ״מ** X.
2. $V(a)=0$ *(שונות של משתנה המקבל ערך קבוע* a *בוודאות היא 0)*
3. $V(X+a)=V(X)$ *(תוספת קבועה לכל סביבי המשתנה לא משנה את השונות)*
4. $V(b\cdot X)=b^2\cdot V(X)$ *(הכפלת סביבי המשתנה בקבוע* b *שווה להכפלת השונות ב-* b^2*)*
5. $V(bX+a)=b^2V(X)$ *(תכונת הלינאריות)*
6. $V(X\pm Y)=V(X)+V(Y)$ *לכל שני מ״מ בלתי תלויים*
7. $V(X_1+X_2+\ldots+X_n)=V(X_1)+V(X_2)+\ldots+V(X_n)$ *לכל* n *מ״מ <u>בלתי תלויים</u>*

[1] יש המסמנים שונות ב - VAR(X) או ב - σ^2

התפלגויות בדידות

ההתפלגות	סימון מקובל	פונקציית ההסתברות	הערכים האפשריים	תוחלת	שונות
בינומית	$X \sim B(n,p)$ $n \geq 1,\ 0 < p < 1$	$p(k) = \binom{n}{k} p^k (1-p)^{n-k}$	$k = 0,1,...,n$	$n \cdot p$	$n \cdot p \cdot (1-p)$
גיאומטרית	$X \sim G(p)$ $0 < p < 1$	$p(k) = (1-p)^{k-1} \cdot P$	$k = 1,2,...$	$\frac{1}{p}$	$\frac{1-p}{p^2}$
פואסון	$X \sim P(\lambda)$ $\lambda > 0$	$p(k) = e^{-\lambda} \cdot \frac{\lambda^k}{k!}$	$k = 0,1,...$	λ	λ
אחידה	$X \sim U(a,b)$ שלם $b-a$	$p(k) = \frac{1}{b-a+1}$	$= a, a+1, .., b$	$\frac{a+b}{2}$	$\frac{(b-a+1)^2 - 1}{12}$

משתנה מקרי דו-ממדי

התפלגות משותפת

הגדרה: פונקציית ההסתברות המשותפת של משתנה מקרי דו-ממדי (X,Y) ניתנת על ידי:

$P(x,y)=P(X=x,Y=y)$ לכל ערך אפשרי x של X ו – y של Y.

טבלת ההסתברות המשותפת

X/Y	x_1	x_2	.....	x_i	x_n	$P(Y=y)$
y_1						$P(y_1)$
y_2						$P(y_2)$
....						
y_j				$P(X=x_i \cap Y=y_j)$		$P(y_j)$
y_m						$P(y_m)$
$P(X=x)$	$P(x_1)$	$P(x_2)$		$P(x_i)$	$P(x_2)$	1

הסתברויות שוליות של y

הסתברויות שוליות של X

<u>תכונות פונקציית ההסתברות המשותפת</u>

א. $0 \le P_{XY}(x_i, y_j) \le 1$ לכל זוג ערכים אפשרי (x_i, y_j)

ב. סכום ההסתברויות של כל התאים בטבלה שווה ל – 1 כלומר $\sum_i \sum_j P_{XY}(x_i, y_j) = 1$

<u>משתנים מקריים בלתי תלויים</u>

הגדרה: שני מ״מ בעלי התפלגות משותפת P_{XY} ייקראו משתנים **בלתי תלויים** (בקיצור ב״ת) אם לכל (x_i, y_j) מתקיים: $\boxed{P(X=x_i, Y=y_j) = P(X=x_i) \cdot P(Y=y_j)}$

אחרת נאמר שהמשתנים **תלויים**.

הערות

- להוכחת תלות בין X ו- Y די למצוא ערך אחד x של X וערך אחד y של Y שעבורם $P(X=x_i, Y=y_j) \neq P(X=x_i) \cdot P(Y=y_j)$

- כשנתונה טבלת התפלגות משותפת של המשתנים, הוכחת תלות זהה למציאת "**תא אחד**" בטבלה, שהסתברותו **שונה** ממכפלת שתי ההסתברויות השוליות המתאימות. להוכחת אי תלות יש להראות שהטבלה היא **כולה** מכפלתית.
- אם לפחות באחד מהתאים בטבלת ההתפלגות המשותפת מופיעה הסתברות **אפס**, אזי המשתנים **תלויים** .

<u>פונקציית ההסתברות המותנית של Y בהינתן $X = x$</u>

ההסתברות המותנית שהמשתנה Y יקבל ערך y כלשהו בהינתן שהמשתנה X קיבל ערך מסוים x ניתנת על ידי הנוסחה:

$$\boxed{P(Y = y \mid X = x) = \frac{P(X = x, Y = y)}{P(X = x)}}$$

כאשר עוברים על כל הערכים האפשריים y של Y מקבלים פונקציית הסתברות לכל דבר, הנקראת **פונקציית ההסתברות המותנית של Y בהינתן $X = x$.**

<u>תוחלת מותנית</u>

התוחלת המותנית של Y בהינתן $X = x$ מסומנת $E(Y \mid X = x)$ וניתנת על ידי הנוסחה:

$$\boxed{E(Y \mid X = x) = \sum_{y} y \cdot P(Y = y \mid X = x)}$$

שונות משותפת

הגדרה: השונות המשותפת של שני מ"מ X ו- Y בעלי תוחלת סופית מסומנת על ידי $\mathrm{cov}(X,Y)$

ומוגדרת כך: $\mathrm{cov}(X,Y) = E(X \cdot Y) - E(X) \cdot E(Y)$

<u>תכונות השונות המשותפת</u>

1. צורה אחרת לכתיבת $\mathrm{cov}(X,Y)$ היא: $\mathrm{cov}(X,Y) = E[(X - EX)(Y - EY)]$
2. $\mathrm{cov}(X,X) = V(X)$ *(השונות המשותפת של מ"מ עם עצמו היא השונות הרגילה שלו)*
3. השונות המשותפת היא סימטרית, כלומר $\mathrm{cov}(X,Y) = \mathrm{cov}(Y,X)$
4. אם (X,Y) מ"מ דו-ממדי. a,b,c,d מספרים קבועים אזי:
$\mathrm{cov}(aX + b,\ cY + d) = a \cdot c \cdot \mathrm{cov}(X,Y)$
5. אם X,Y,Z מ"מ המוגדרים על אותו מרחב מדגם Ω אז:

$$\mathrm{cov}(X, Y + Z) = \mathrm{cov}(X,Y) + \mathrm{cov}(X,Z)$$

<u>תוחלת ושונות של סכום והפרש של משתנים מקריים</u>

יהיו X ו Y – מ"מ המוגדרים על אותו מרחב מדגם Ω אז:

$$E(X \pm Y) = E(X) \pm E(Y)$$

$$V(X \pm Y) = V(X) + V(Y) \pm 2COV(X,Y)$$

<u>משתנים מקריים בלתי מתואמים</u>

יהיו X ו Y – מ"מ המוגדרים על אותו מרחב מדגם Ω ייקראו **בלתי מתואמים** אם מתקיימת אחת משלוש התכונות (ואז בהכרח כולן מתקיימות):

$$E(X \cdot Y) = E(X) \cdot E(Y)$$

$$\text{cov}(X,Y) = 0$$

$$V(X + Y) = V(X) + V(Y)$$

<u>הקשר בין תלות ומתאם בין 2 מ"מ</u>

שני המושגים תלות ומתאם אינם זהים!!!

אי תלות גוררת חוסר מתאם כלומר: אם X ו Y – מ"מ **ב"ת** אזי הם בהכרח גם **בלתי מתואמים.**

ולכן אם X ו Y – מ"מ **בלתי מתואמים** אזי הם בהכרח גם **תלויים**.

שים לב שמשתנים תלויים יכולים להיות גם בלתי מתואמים!

מאחר שמשתנים בלתי תלויים הם בהכרח גם בלתי מתואמים ניתן לגזור את הנוסחה הבאה:

שונות הסכום של **2** מ"מ **<u>בלתי תלויים</u>**: $V(X + Y) = V(X) + V(Y)$

ובהכללה: **לכל n** מ"מ **<u>בלתי תלויים</u>**: $V(X_1 + X_2 + ... + X_n) = V(X_1) + V(X_2) + ... + V(X_n)$

<u>מקדם המתאם בין 2 מ"מ X ו Y –</u>

מקדם המתאם הלינארי הוא מדד לעצמת הקשר הלינארי בין שני משתנים תוך שימוש בהתפלגות המשותפת שלהם.

הגדרה: מקדם המתאם בין 2 מ"מ X ו- Y מסומן ב- $\rho(X,Y)$ ומוגדר על ידי:

$$\boxed{\rho(X,Y) = \frac{\text{cov}(X,Y)}{\sigma(X) \cdot \sigma(Y)}}$$

<u>תכונות מקדם המתאם</u>

1. סימטריה: $\rho(X,Y)=\rho(Y,X)$
2. לכל X ו- Y מתקיים: $-1\le\rho(X,Y)\le 1$
3. התנאי $\rho(X,Y)=0$ שקול ל- $\mathrm{cov}(X,Y)=0$ כלומר אם X ו – Y מ״מ **בלתי מתואמים**
4. אם $Y=\beta X+\alpha$ אזי:
   אם $\beta>0$, $\rho(X,Y)=1$
   אם $\beta<0$, $\rho(X,Y)=-1$
5. ערכו של מקדם המתאם אינו תלוי בקנה המידה ובהזזות. לשון אחר: טרנספורמציה לינארית על המשתנה\ים אינה משפיעה על ערכו המוחלט של מקדם המתאם. כיוונו יכול להשתנות בהתאם לסימן המכפלה של $\beta\cdot\gamma$

עבור $\beta\cdot\gamma>0$ מתקיים: $\rho(\beta\cdot X+\alpha\ ,\ \gamma\cdot Y+\delta)=\rho(X,Y)$

עבור $\beta\cdot\gamma<0$ מתקיים: $\rho(\beta\cdot X+\alpha\ ,\ \gamma\cdot Y+\delta)=-\rho(X,Y)$