**ל.ע קורלציה**

**קורלציה – גודל האפקט נומרי על נומרי**

עד לנקודה הזאת דיברנו על משתנים קטגוריאליים. איך הם מתקשרים למשתנים רציפים, או למשתנים קטגוריאליים אחרים, אבל מה לגבי הקשר בין משתנה נומרי אחד למשתנה נומרי נוסף?

גודל האפקט נומרי על נומרי נקרא קורלציה. בקורלציה אנחנו נסתכל על שני משתנים נומריים, ונבדוק עד כמה הם קשורים באופן לינארי זה לזה – עד כמה הם משתנים בקצב קבוע – ככל שמשתנה אחד עולה גם השני עולה, או ככל שמשתנה אחד עולה השני יורד.

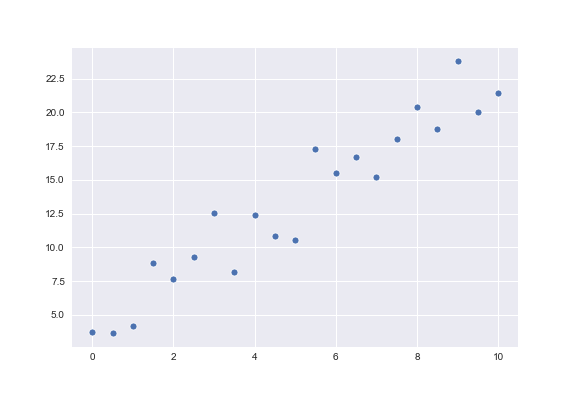
חשוב לציין שהקשר שאנו בודקים הוא דו כיווני ולא סיבתי, זה לא שמשתנה אחד גורם לשינוי במשתנה השני, אלא ששניהם גורמים לשינויים זה אצל זה. לדוגמה, אם קיים קשר בין המשתנים הרציפים משקל וגובה, אפשר להגיד שככל שהמשקל עולה גם הגובה עולה, ואפשר גם להגיד מהכיוון השני שככל שהגובה עולה גם המשקל עולה.

\*נבצע קורלציה אך ורק כשהמשתנים שלי הם ברמת מדידה אינטרוולית או יחסית.

\*הערה נוספת – כשנבצע קורלציה באמצעות pearson’s r נעדיף שהמידע יהיה כמה קרוב לנורמליות, אומנם אפשר אחרת אבל ככל שהמידע יהיה קרוב לנורמלי כך התוצאה שנקבל תהיה אמינה יותר.

**ויזואליזציות**

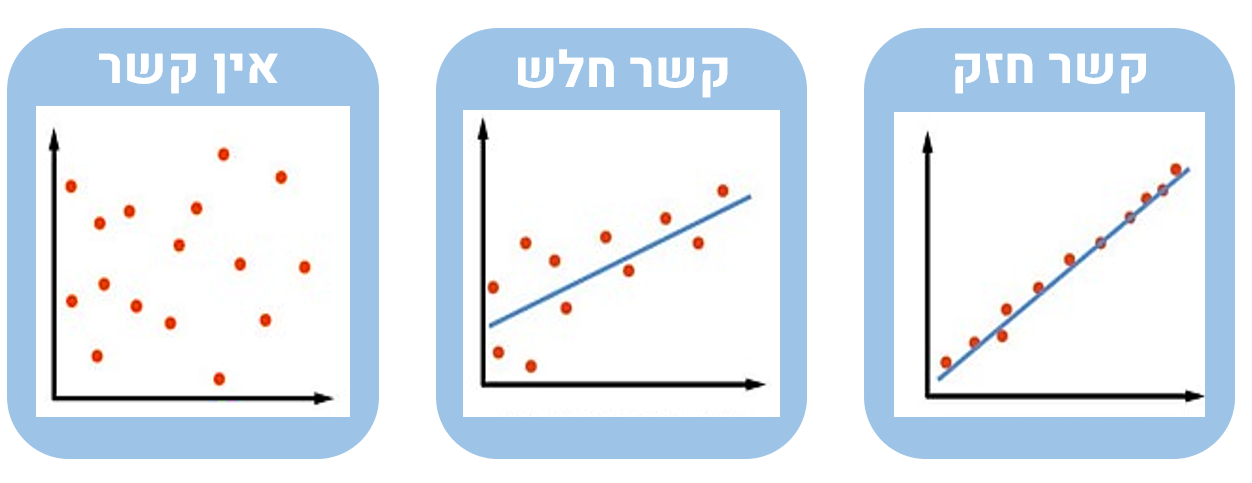
הוויזואליזציה המרכזית שתשמש אותנו בחקירת הקשר בין שני משתנים רציפים היא תרשים פיזור (scatter plot). התרשים נראה כך:



**כל תצפית מיוצגת ע"י נקודה אשר ערך ה-X שלה הוא הערך של משתנה א', וערך ה-Y שלה הוא הערך של משתנה ב'.**

שימו לב - כאשר חוקרים השפעה של משתנה אחד על משתנה אחר עם כיוון מסויים, X ייצג את המשתנה הבלתי תלוי (המשפיע), ו-Y ייצג את המשתנה התלוי (המושפע). במקרה של קורלציה אנחנו חוקרים קשר ללא הנחת כיוון ספציפי, הקשר שלנו הוא דו כיווני, לכן אין חשיבות לאיזה משתנה מיוצג ע"י איזה ציר.

תרשים הפיזור ייתן לנו אינדיקציה ויזואלית על טיב הקשר בין המשתנים מכיוון שקורלציה בעצם אומרת לנו עד כמה ההצלבה בין הנתונים שלנו מתאימה לקו ישר. ככל שהנקודות שלנו קרובות יותר להיות מונחות על קו ישר, כך הקשר בין המשתנים יהיה יותר משמעותי. ניתן לראות זאת בדוגמה הבאה:



גם כאן, כמו בשאר גדלי האפקט, נעבוד באותו תהליך עבודה – נבדוק האם קיים קשר בעזרת מציאת ערך ה-p ונמצא מה חוזק הקשר.

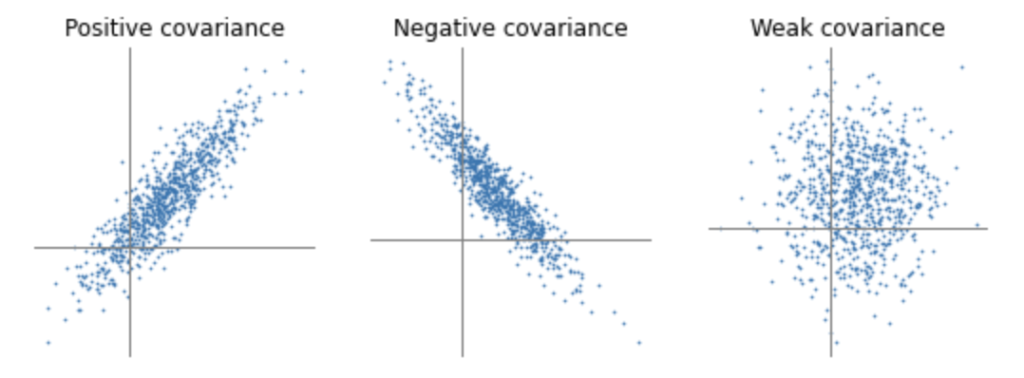
**מציאת חוזק הקשר**

אנחנו יודעים כיצד מוצאים חוזק קשר עבור שני הסוגים האחרים של גודל האפקט, אך איך נעשה זאת עבור קורלציה?

**שונות משותפת** היא מדד סטטיסטי לקשר בין שני משתנים הבודק את המידה שבה שני משתנים עוברים שינוי יחד.

חישוב השונות המשותפת מחזיר מספר בטווח שבין מינוס אינסוף לאינסוף. כאשר השונות המשותפת היא **חיובית** זה מעיד על כך ששני המשתנים (משתנה א' ומשתנה ב') משתנים **באותו הכיוון, בקשר ישיר** (ככל שמשתנה א' גדל, גם משתנה ב' גדל).

כאשר השונות המשותפת **שלילית**, זה מעיד על כך שהמשתנים משתנים **בכיוונים שונים, בקשר הפוך** (ככל שמשתנה א' גדל, משתנה ב' קטן). כאשר השונות המשותפת שווה ל-0, זה אומר שלא קיים קשר בין המשתנים.



אך מה אם אנחנו רוצים להגדיר בבירור מהו חוזק הקשר, ולא רק מה הכיוון שלו?

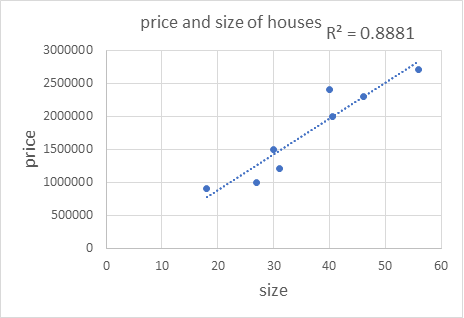
כאשר אנו מקבלים ערך מסוים בחישוב השונות - ערך חיובי או שלילי (משתנים הנעים לאותו כיוון או להיפך), כיצד נדע מה זה מעיד על הקשר? האם הקשר חזק? חלש? מדד השונות פחות אינטואיטיבי להבנת נקודה זו, כיוון שאינו נמצא על סקאלה של מספרים מסוימים, הוא לא נותן לנו "ציון" ברור למידע.

*ניקח דוגמה:*

יש בידינו מידע על בתים – גודל הבית במטר רבוע ומחירו בשקלים.

המידע מונח על גרף פיזור באופן הבא:

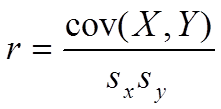
|  |  |
| --- | --- |
| **size** | **price** |
| 18 | 900000 |
| 27 | 1000000 |
| 31 | 1200000 |
| 30 | 1500000 |
| 40.5 | 2000000 |
| 46 | 2300000 |
| 56 | 2700000 |
| 40 | 2400000 |



כאשר נחשב את השונות המשותפת על ידי נוסחה באקסל (**covariance**) נקבל את הערך: **6815625**.

ערך זה אינו מעיד שום דבר ברור על עוצמת הקשר בין המשתנים, רק על כך ששניהם נוטים לאותו הכיוון. המספר הגדול שהתוצאה נותנת לנו אינו בהכרח מעיד לנו על חוזקו של הקשר, אולי הוא קשור למשתנים בעלי הערך הגדול משמעותית של המחיר? אין לנו דרך לדעת האם הקשר הזה נחשב קשר חלש, בינוני או חזק, חלש או בינוני – מכיוון שהמספר הזה לא מנורמל.

לכן, עלה הצורך **לנרמל** את השונות המשותפת, ונוצר מדד נוסף הנקרא **מדד פירסון Pearson’s r**. הנרמול התבצע על ידי חלוקה של השונות המשותפת במכפלת סטיות התקן של כל אחד מן המשתנים.

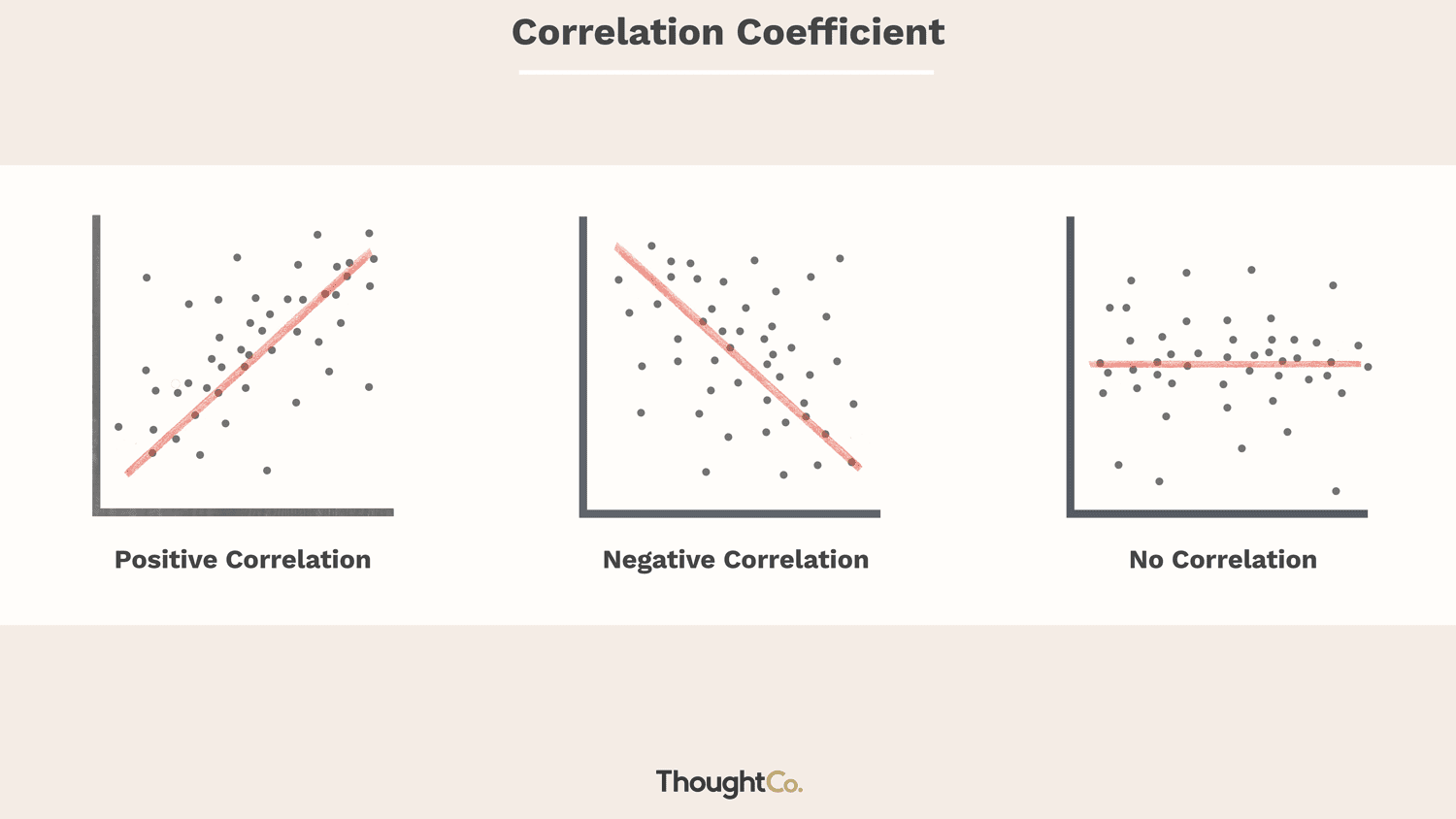


מדד פירסון הוא המדד שנשתמש בו לבדיקת הקורלציה הליניארית בין המשתנים.

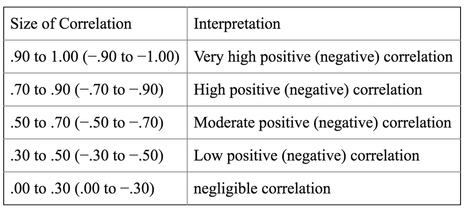
r (תוצאת המדד) נע בין 1- ל -1, כאשר 1 מיצג קורלציה ליניארית מושלמת, 0 מיצג אי-קורלציה (חוסר קשר) מושלמת, ו- 1- מיצג קורלציה שלילית מושלמת. אין צורך להיכנס לעומק לנוסחה משום **שנחשב אותה באמצעות ההרחבה לאנליסטים באקסל (ראו דוגמה בסוף).[[1]](#footnote-2)**

**פירוש התוצאה**

הערך של r נע בין 1 ל 1-, כאשר הסימן (פלוס או מינוס) מייצג את סוג הקשר (חיובי או שלילי), כלומר האם ככל ש-X עולה Y עולה, או שככל ש-X עולה Y יורד.   
ככל ש-r מתרחק מ-0 ומתקרב ל-1 או ל 1-, הקשר שבין X ו-Y חזק יותר עד שב-1 וב1- הקשר מושלם.   
חשוב לזכור ש-r לא אומר לנו כלום על אופי הקשר בין X ו-Y (דברים כגון שיפוע הקו האופטימלי או נקודת המפגש עם הצירים), אלא רק מה חוזק הקשר.



כאשר אנחנו מנסים לפרש את הערך של r מבחינת חוזק הקשר יש להיעזר בטבלאות שניתן למצוא באינטרנט, בעזרתן נצמיד מסקנה לערכי r:

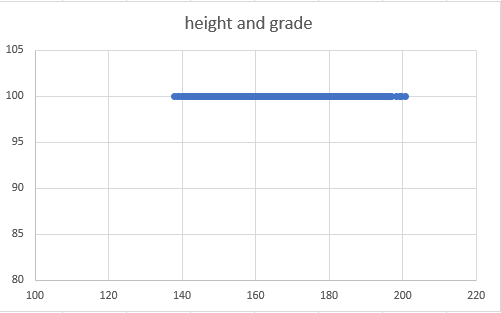


\* מספרים שמתחילים בנקודה הכוונה היא במספרים עשרוניים, לדוגמה: 0.30.

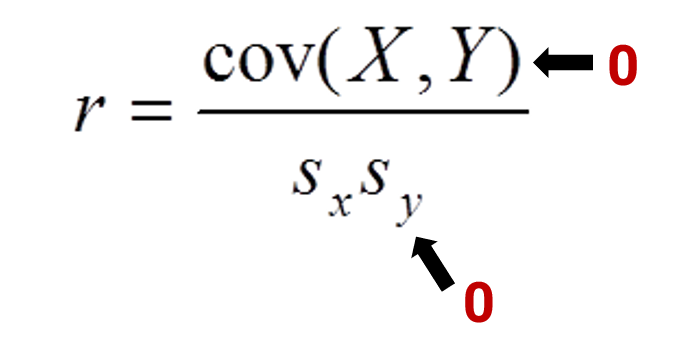
**מקרים יוצאי דופן – דגשים לגבי מדד פירסון**

חישוב מדד פירסון לא תקף עבור זוג משתנים בהם X או Y בעלי ערך בודד. לכאורה, הקורלציה אמורה להיות מושלמת, מאחר שכל הנקודות מתכנסות לאותו קו, המקביל לציר ה-X (במקרה וכל ערכי Y זהים), או לציר ה-Y (במקרה וכל ערכי X זהים). לדוגמה:

יש לי מידע על גבהים וציונים של תלמידים ואני רוצה לראות האם יש קשר בין שני המשתנים. אני רואה שכל הנקודות נמצאות על קו ישר בצורה מושלמת, כי כל התלמידים קיבלו 100.



אם נסתכל על הנוסחה של המדד לחוזק הקשר r, שונות משותפת חלקי מכפלת סטיות התקן:



גם השונות המשותפת של שני המשתנים תהיה שווה ל-0 וגם סטיית התקן של המשתנה הקבוע תהיה שווה 0 (אין לו פיזור, הוא קבוע בערך אחד), כך שהנוסחה לחישוב המדד תיתקל במקרים אלו בשגיאה שנובעת מהניסיון לחלוקה ב-0. אם נחשוב על זה, המדד באמת אינו משקף מקרים כאלו, מאחר שהוא מודד את השינוי ב-Y כפונקציה של השינוי ב -X (או להפך), וכאשר אין שינוי ב-X או ב-Y, הנוסחה נכשלת. במקרים אלו, נתייחס אל הקשר כאקראי, משמע r = 0. באופן כללי, כאשר קו המגמה מקביל לאחד הצירים r = 0 – נשתדל להימנע מחישוב של קורלציה.

**צפו בסרטון הבא** - <https://www.youtube.com/watch?v=DAH8DyLXdjM>

**ערך r בריבוע**

כאשר נעלה את הערך של pearson’s r בריבוע, נקבל את מקדם ההסבר, מדד סטטיסטי שאומר **עד כמה המשתנים יכולים להסביר אחד את השני**. ההעלאה בריבוע גורמת לתוצאה להיות בין 0 ל-1, ונקבל דרך פשוטה יותר להסביר את הקשר בין המשתנים – בעזרת R בריבוע בעצם נוכל להגיד בכמה אחוזים משתנה ב' מסביר את השונות של משתנה א''.

בכך נוכל להצמיד מסקנות מילוליות לערכי r, למשל עבור r = 0.9, נוכל להגיד ש81% מהשונות במשתנה Y (0.9\*0.9 = 0.81) מוסברים ע"י משתנה X.

* תזכורת: שונות = סטיית תקן בריבוע, מדד שמצביע על פיזור הערכים סביב הממוצע.

אם המשתנים מסבירים זה את זה ב-100%, הם מסבירים בצורה מושלמת אחד את השני וכל הנקודות יהיו על קו ישר.

ככל שהמספר שנקבל יהיה גבוה יותר - כך ניתן להבין כי בעקבות משתנה א' הרבה מן התצפיות במשתנה ב' משתנות במרחב באופן מסוים. כך אנחנו רואים עד כמה משתנה אחד מסביר את ההבדלים במשתנה אחר.  
כך נוכל גם לאמוד את ההבדלים האמיתיים בין ערכי r, למשל r = 0.9 מסביר 81 אחוזים, אבל r = 0.6 מסביר רק 36 אחוזים.

ומה לגבי שאר השונות שאינה מוסברת ע"י המשתנה? קיימות שתי אופציות:

1. השונות נגרמת ע"י רנדומליות טבעית במידע - גם אם נפתח מודל מדויק שחוזה גודל של תפוחים, עדיין יש כנראה אלמנט של רנדומליות שמשפיע על הגודל שלהם, ולכן לעולם לא נצליח להסביר 100% מהשונות.
2. השונות נגרמת ע"י גורם שאינו מוכר לנו, או שלא התייחסנו אליו במחקר שלנו (ייתכן שאחוז הילודה באוזבקיסטן משפיע על גודל התפוחים, ומאחר שאינו מיוצג במדד שלנו, לא נוכל להסביר את השונות שהוא גורם).

לרוב ניתקל בשילוב של שני המקרים, שכן נדיר שאין בכלל אלמנט של רנדומליות במידע (שכולו נשלט ע"י גורמים מסוימים), וככל הנראה לא תהיה לנו גישה לכל המידע המשפיע, או בכלל מושג לגבי כל הגורמים המשפיעים.[[2]](#footnote-3)

**ערך p בקורלציה**

המשמעות של ערך ה-p בקורלציה היא **ההסתברות שנתונים רנדומליים יציגו קשר בעל חוזק דומה או גדול לקשר שמצאנו**. כלומר, אם ההסתברות לקחת נתונים רנדומליים מהמידע ולקבל ביניהם קשר כמו זה שקיבלנו או חזק יותר היא גבוהה, אנחנו יודעים שכנראה קיבלנו את הקשר הזה במקרה ושאי אפשר להסתמך עליו – הוא לא מובהק סטטיסטית ולא נסיק לגביו מסקנות.

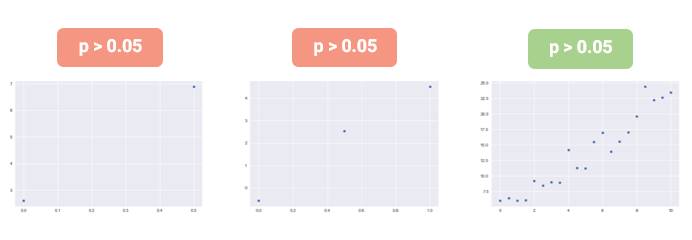
גם במקרה הזה, ערך ה-p המקסימלי עבורו נוכל להתייחס לקשר שמצאנו הוא 0.05, כלומר שעבור 20 ניסויים של קשרים רנדומליים, רק אחד יציג קשר זהה או חזק יותר, וזה היחס המקסימלי שמקובל לאפשר.

**על ערך ה-p משפיעים שני מאפיינים:**

1. **r** – מאחר שהסיכויים של מידע רנדומלי להציג קשר זהה או חזק יותר קטנים יותר ככל שהקשר חזק יותר, ככל שערכו של r עולה, ערכו של p יורד.
2. **N** – ככל שיש יותר נקודות מידע, ההסתברות שמידע רנדומלי יציג קשר חזק יורדת. ניקח לדוגמה מידע בעל 2 נקודות. כל מידע בעל שתי נקודות יציג קשר מושלם, אבל מאחר שיש רק 2 נקודות, p יהיה גבוה מאוד (p = 1 למעשה), ולכן לא נוכל להתייחס לקשר.

**ככל ש-r ו-n גבוהים יותר, כך p נמוך יותר.**

ניקח דוגמה:



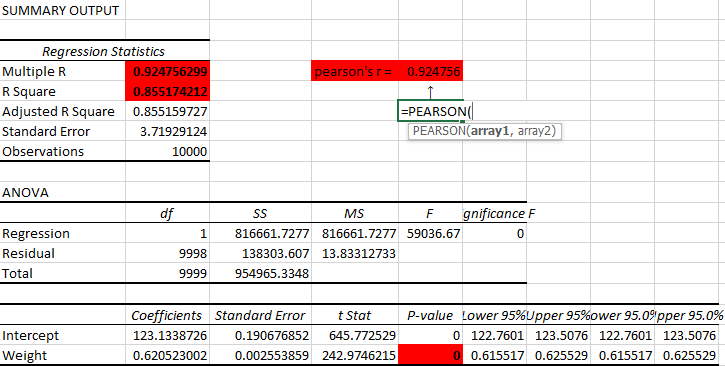
יש לנו 3 תרשימי פיזור, בעבור משתנים רציפים שכבר מצאנו שיש בהם קורלציה. אולם, כמות התצפיות משפיעה על האם אנחנו יכולים להסתמך על הקשר הזה או לא. אפשר לראות בשני הגרפים השמאליים יש 2 או 3 תצפיות, מה שגורם לחשוב שהתצפיות נעות באותה המגמה, אך זה רק בגלל המחסור בתצפיות. אי אפשר להסתמך על כמות כזו קטנה של ערכים ולהגיד שקיים קשר מסויים, לכן ערך הp יתחשב בכך ויגיד לנו שהקשר לא מובהק סטטיסטית.

בתרשים הימני לעומת זאת, יש הרבה תצפיות, אז ערך ה-p יראה שאנחנו כן יכולים להסתמך על הקשר שמצאנו ולהגיד שקיים קשר בחוזק מסוים.

**גם את ערך ה-p נחשב דרך ההרחבה לאנליסטים באקסל****.**

**הסבר – חישוב pearson’s r, r squared, p value בהרחבה לאנליסטים:**

* נפתח את ההרחבה ונבחר באופציה של regression.
* נסמן את שתי העמודות שלנו – שימו לב לא לסמן את העמודה כולה כ-A:A למשל, זה עלול להקפיץ לכם שגיאה, תסמנו כל עמודה במדוייק עם קיצור המקלדת- לחיצה על כותרת העמודה, ולאחר מכן על המקשים: ctrl + shift + ↓
* נקבל מספר תוצאות בטבלה:



התוצאות המסומנות באדום הן מה שרלוונטי לנו-

**R square**- מראה לנו את תוצאת r בריבוע

**Multiple r**- מראה לנו את תוצאת pearson’s r- ההרחבה הזו מחשבת את ערך זה על ידי שורש של r square, כך שזה מראה לנו את חוזק הקשר ללא כיוון הקשר (חיובי או שלילי). כדי שנקבל תוצאה בין -1 ל1 , עם הכיוון, **נשתמש בנוסף בנוסחה המובנית באקסל PEARSON**.

**p-value-** מראה לנו את ערך המובהקות, נשתמש בערך שבשורה השנייה.

**הרחבת גרף הפיזור:**

לפעמים תרשים הפיזור יראה כל כך מכווץ שלא נוכל להבין מה המגמה במידע. כדי להתמודד עם כך, נוכל לשנות מעט את גרף הפיזור, אחרת פשוט לא נשיג מגרף הפיזור את המטרה שלו.

1. הצגה אוטומטית: לפעמים לא יהיה לנו צורך לעשות כל שינוי בגרף הפיזור, כיוון שהאקסל באופן אוטומטי יזיז את הצירים קרוב יותר לריכוז הנתונים. כך, ראשית הצירים לא תהיה ב -0, אלא במספרים שמתאימים את עצמם לנתונים. זוהי פשוט הסתכלות קרובה יותר על הפיזור.

עם זאת, יש לזכור שבמקרים אחרים לא נציג את פיזור הנתונים בצורה הזו, זה לא עומד בסטנדרטים לוויזואליזציה מתאימה. הגרף הזה מאפשר לנו לראות את המגמה הכללית של המידע, אך רק במקרים האלה נשתמש בגרף פיזור בצורה הזו.

1. נרמול מינימום – מקסימום: אם אנחנו רואים שגרף הפיזור מאוד מקובץ(האקסל לא תמיד יתאים את עצמו למידע, תקראו את הלמדריך) נוכל להרחיב אותו ע"י נרמול כל אחד מהמשתנים עפ"י מינימום מקסימום.

נרמול כזה שומר על היחסים בין המרחקים שבין הנקודות, לכן הפיזור יהיה אותו הפיזור, רק באופן מורחב.

האקסל לא יזיז לנו באופן אוטומטי את הפיזור וישאיר אותו מאוד מכווץ במקרה שיש חריגים.

גם אם ננרמל את המשתנים במקרה הזה, הפיזור עדיין יהיה מאוד מכווץ. בשביל לראות כמו שצריך איך המידע מתנהג מבחינת הלינאריות שלו, נוריד את החריגים באופן זמני, ואז נוכל להשתמש באחת הדרכים המצוינות למעלה כדי להרחיב את הפיזור. או שנבחר להעלות את שני המשתנים על גרף פיזור והאקסל באופן אוטומטי יזיז את הצירים, או שננרמל את המשתנים.

שימו לב – אנחנו תמיד נעדיף לראות את גרף הפיזור כאשר ראשית הצירים היא 0.0 כי כך אנו רואים אותו בצורה המייצגת ביותר על ציר המספרים. אנחנו נעשה את ההזזה של הצירים והנרמול אך ורק במקרה בו הפיזור מאוד מקובץ ואנחנו לא מצליחים להסיק ממנו מסקנות.

***שאלות הבנה***

1. הסבר מהי קורלציה.
2. מה ההבדל בין קורלציה לבין גודלי אפקט אחרים שקשורים במשתנים קטגוריאליים?
3. באיזו וויזואליזציה נשתמש בעיקר בקורלציה, ומה מייצגת כל תצפית? הסבר.
4. איך ניתן לפרש את תוצאת מדד פירסון?
5. מה המשמעות של r בריבוע?

**רגרסיה לינארית**

**מה זה רגרסיה?**

רגרסיה היא שם כולל למשפחה של מודלים סטטיסטיים שהמשותף ביניהם הוא שיש משתנה מוסבר (המשתנה התלוי) ומשתנה מסביר (משתנה תלוי - אחד או יותר), כאשר המטרה היא לחזות את הערך של המשתנה המוסבר בהסתמך על המשתנה המסביר (או מספר משתנים מסבירים). כלומר, בהינתן הערכים של המשתנים המסבירים, נוכל לדעת מה יהיו הערכים של המשתנה המוסבר. למשל, בהינתן משקל של בן אדם, נוכל לדעת מהו הגובה שלו.

ישנם סוגים רבים של מודלים לרגרסיה - אפשר להשתמש ברגרסיה בעבור מערכות יחסים לינאריות של משתנים (של קו ישר) ובעבור מערכות יחסות לא לינאריות של משתנים (כמו רגרסיה לוגיסטית למשל), אך סוג הרגרסיה הנפוץ ביותר אשר נלמד עליו היום הוא **רגרסיה לינארית**, שמתעסקת במשתנים עם קשר לינארי, אשר יוצרים קו ישר. חשוב לציין שנלמד על רגרסיה לינארית בין 2 משתנים בלבד – משתנה תלוי ובלתי תלוי, ולא על סיטואציה בה יש מספר משתנים בלתי תלויים.

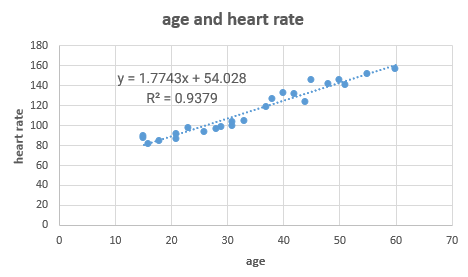
**הרגרסיה הלינארית**

עד לנקודה זו עסקנו בקורלציה לינארית – קשר דו צדדי בין שני משתנים רציפים, אשר יוצרים יחד קו ישר אם הקשר מספיק חזק. כעת, לאחר שראינו שיש קשר בין המשתנים, נשתמש ברגרסיה לינארית על מנת להבין את המשמעות של קו המגמה שהמשתנים יוצרים.

ברגרסיה לינארית y הוא המשתנה התלוי (המוסבר) ו-x הוא המשתנה הבלתי תלוי (המסביר). ברגרסיה לינארית מתאימים למשתנים האלה קו ישר עם המשוואה **y = ax + b,** על פי עיקרון מינימום הריבועים– שאומר כי סכום ריבועי המרחקים של הנקודות מהקו הוא הקטן ביותר. עבור הסכום הקטן ביותר, הישר הוא הישר האופטימלי, ומשוואתו היא משוואת הרגרסיה הליניארית של הנתונים. הקו הישר הזה נקרא גם **קו הרגרסיה**. הקו הזה יכול לאפשר לנו לחזות את ערכי משתנה y בעזרת ערכי משתני x.

דוגמה:

יש לנו מידע על גיל של בן אדם בשנים וקצב הלב שלו, ומצאנו כי קיים קשר בין שני המשתנים. הוספנו לתרשים הפיזור את קו המגמה (קו הרגרסיה), את משוואת הרגרסיה, ואת ערך r בריבוע (שלמדנו עליו מקודם).



שימו לב: בקורלציה אין חשיבות לאיזה משתנה נשים על איזה ציר, כי הקשר שאנו בודקים הוא דו כיווני, אך ברגרסיה יש חשיבות לכך - רגרסיה היא חד כיוונית, ומתאימה ערך למשתנה התלוי (המשתנה בציר ה-y) על סמך המשתנה הבלתי תלוי (המשתנה בציר ה-x).

**מה אפשר להסיק מכך?**   
השימוש ברגרסיה מאפשר לנו לחזות בעזרת המשתנה המסביר את המשתנה המוסבר. למשל, אם ידוע שגיל הבן אדם הוא 45, תצפית שאין לנו במידע הקיים, אפשר להציב במשוואת הרגרסיה שx = 45 ונקבל 133.87 y =, ולכן ניתן להסיק שאם גיל הבן אדם הוא 45 כנראה שקצב הלב שלו יהיה בערך 133.87. עד כמה זה יהיה באמת קרוב למספר הזה? זה תלוי בערך הקורלציה הליניארית שחישבנו, או יותר נכון בערך ה-r בריבוע.

* עבור r בריבוע גדול - סימן שאחוז גדול מהשונות מוסבר ע"י משתנה X, ולכן משתנה X מסביר באופן משמעותי את השונות של משתנה ב', ומכאן שטווח הטעות שלנו בחיזוי Y קטן.
* עבור r בריבוע קטן - X מסביר לנו רק חלק קטן השונות במשתנה Y, ומאחר שההשפעה הגדולה היא של גורמים שאנחנו לא מודעים אליהם, נוכל לצפות לטווח טעות גדול בתחזית שלנו של משתנה Y באמצעות משתנה X.

**הסבר לחישוב רגרסיה לינארית באקסל**

איך עושים זאת? נציג את תרשים הפיזור של הנתונים, ואז נלחץ על אחת מהנקודות לחצן ימני, ונבחר ב"הוספת קו מגמה". מיד יופיע לנו קו הרגרסיה. יש להקפיד שהאופציה שנבחרה היא "רגרסיה ליניארית" ולא אף אחת מהאופציות האחרות. כדי להוסיף את משוואת הרגרסיה, נלחץ לחצן ימני על קו הרגרסיה, נבחר ב"עיצוב קו מגמה", ונסמן את התיבה "הצג משוואה בתרשים" בסרגל העיצוב שנפתח.

בהמשך, תלמדו איך ליצור תרשים פיזור ולהוסיף את קו הרגרסיה באמצעות ספריות בפייתון.

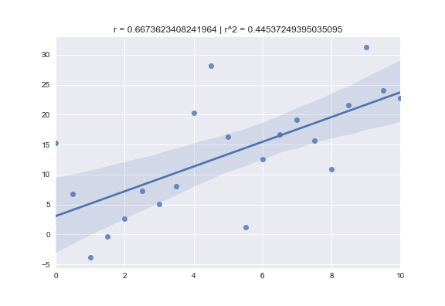
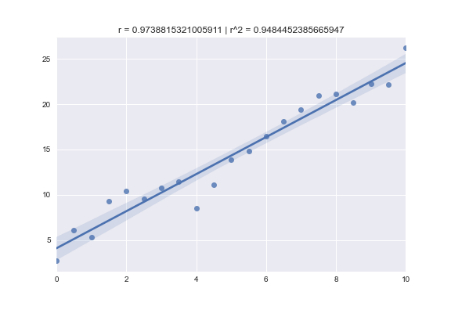
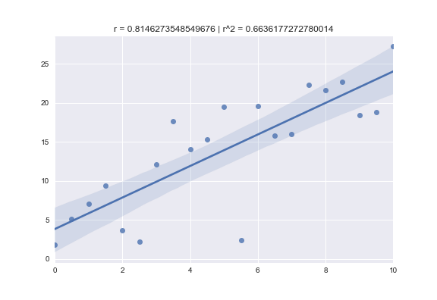
**תנאים לרגרסיה לינארית**

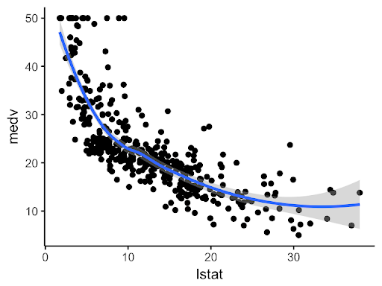
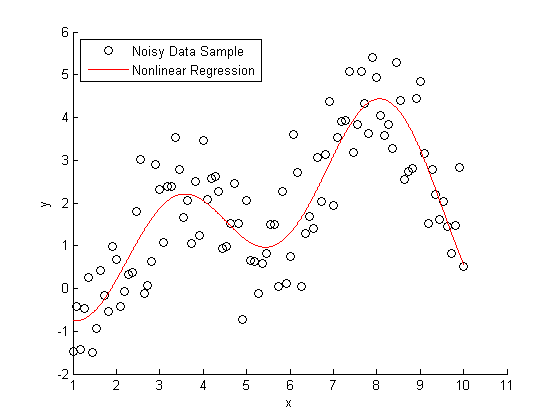
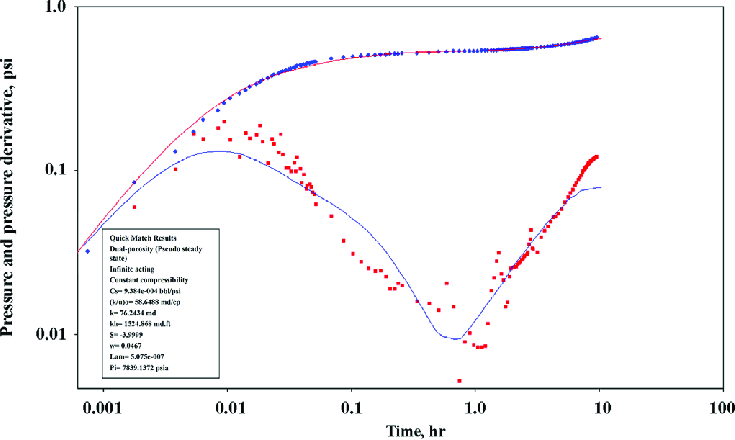
עד כה עבדנו עם צורות מידע ליניאריות אשר מאפשרות תחזית באמצעות רגרסיה, אבל ישנן צורות מידע שאינן מאפשרות את זה, וחשוב להכיר אותן, על מנת שלא להתאים מודלים שאינם מתאימים ולהטעות את הלקוח.

ישנם שלושה תנאים שחייבים להתקיים כדי שנוכל להשתמש במודל הרגרסיה הלינארית ולהפיק ממנו מסקנות אמיתיות:

1. לינאריות כללית:

על מנת שהרגרסיה תספק לנו תחזית מדויקת ואחידה בדיוקה, על הנתונים להסתדר בצורה שדומה לליניארית, למשל:

    
במקרים בהם הנתונים נראים רנדומליים, או כאשר הם עוקבים אחרי תבנית שאינה ליניארית, לא נוכל להסתמך על חישוב הרגרסיה כדי להפיק מסקנו אמיתיות.

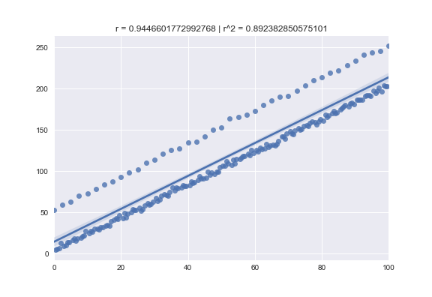
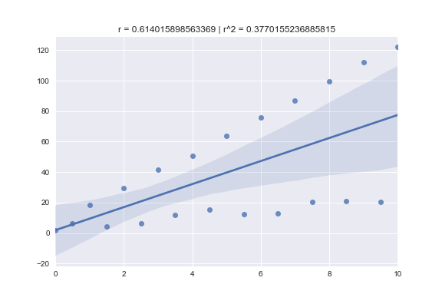
דוגמה למשתנים שכאלו:  
  

חשוב לזכור גם שככל שהקשר יהיה חזק יותר – כך הרגרסיה תהיה מדויקת יותר. כך שאם אנחנו משתמשים בקו הרגרסיה לחיזוי, עלינו לחשוב עד כמה הוא מדויק.

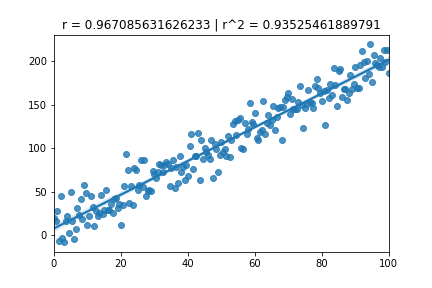
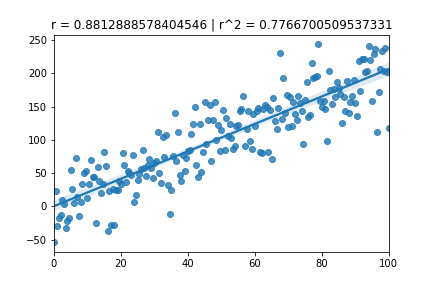
1. פיזור נורמלי סביב קו הרגרסיה:

ייתכנו מקרים בהם הנתונים מפוזרים בצורה תבניתית מסוימת סביב קו הרגרסיה.

שימו לב לתבניות הבאות:



על מנת שנוכל להשתמש ברגרסיה ליניארית בצורה שמשקפת את המידע, התצפיות צריכות להתפזר סביב קו הרגרסיה כך שהמרחקים של התצפיות מקו הרגרסיה יתפלגו נורמלית. אין צורך ממש לבדוק את הנורמליות של המרחקים, פשוט חשוב לשים לב שהנקודות מפוזרות בצורה הומוגנית סביב קו הרגרסיה, כאשר ככל שמתקרבים לקו יש יותר ויותר נקודות, משהו שנראה כך:

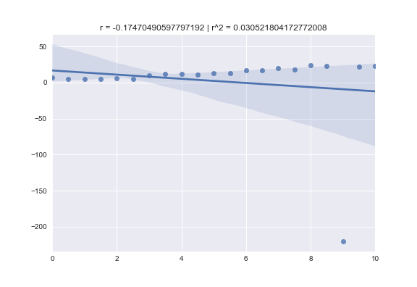
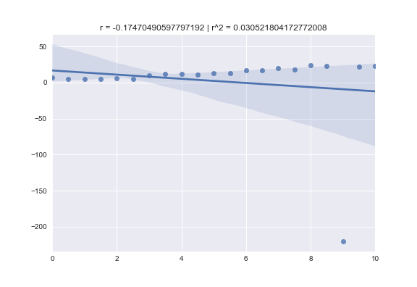
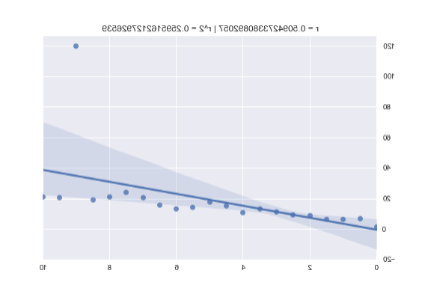


אז מה עושים אם הנתונים לא מתפזרים נורמלית סביב קו הרגרסיה?

* כשמזהים כמה תבניות ליניאריות שמתנגשות, ייתכן שיש גורמים נוספים שמשפיעים על הנתונים (כנראה קטגוריאליים), ונוכל לבדוק את הנתונים עבור כל קטגוריה בנפרד.
* במקרה והנתונים מתאימים לרגרסיה ליניארית רק בטווח מסוים ואז מתחילים להתפזר בצורה שאינה מתאימה, נוכל להשתמש ברגרסיה ופשוט להתייחס רק לטווח המתאים במסקנות שלנו.
* מקרה נוסף הוא שהנתונים פשוט אינם מתאימים לחיזוי באמצעות רגרסיה ליניארית, ונצטרך להמשיך הלאה בתהליך הניתוח.

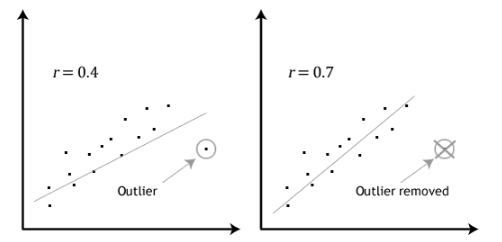
1. ללא חריגים משמעותיים:

מאחר שחישוב הרגרסיה הוא חישוב ערכי, הוא מאוד פגיע לחריגים. אם נחשב רגרסיה לקשר ליניארי עם חריגים משמעותיים, **נקבל תחזית מוסטת!** לדבר כזה יכולות להיות תוצאות הרסניות על המסקנות.

חריגים משמעותיים במידע נראים כך:

מעבר לרגרסיה, גם בקורלציה גודל הקשר יכול להיות מושפע מאוד מחריגים - המדד של pearson’s r מאוד רגיש לחריגים ועלול להביא תוצאות מטעות במקרים כאלה. באותה המידה, יכול להיות שיהיה לנו מידע שאין בו קשר בכלל בין שני המשתנים אך בגלל חריג אחד גדול מאוד המדד יראה שקיים קשר, וזה שגוי.

לדוגמה:



מה נוכל לעשות במקרה כזה?

יהיה לכם לע על התמודדות עם חריגים בהמשך שבו תרחיבו על הנושא.

מה שחשוב לזכור כרגע, זה שחריגים יכולים מאוד להשפיע על הקשר שאנו בודקים ועל הרגרסיה, אך כשהם לא שגיאה, דרך ההתמודדות היא גמישה. על מנת להתמודד איתם בצורה הטובה ביותר, עלינו להתחשב בהרבה גורמים, כמו האירוע שהם מייצגים, צורך המחקר, העדפת הלקוח וכו'.

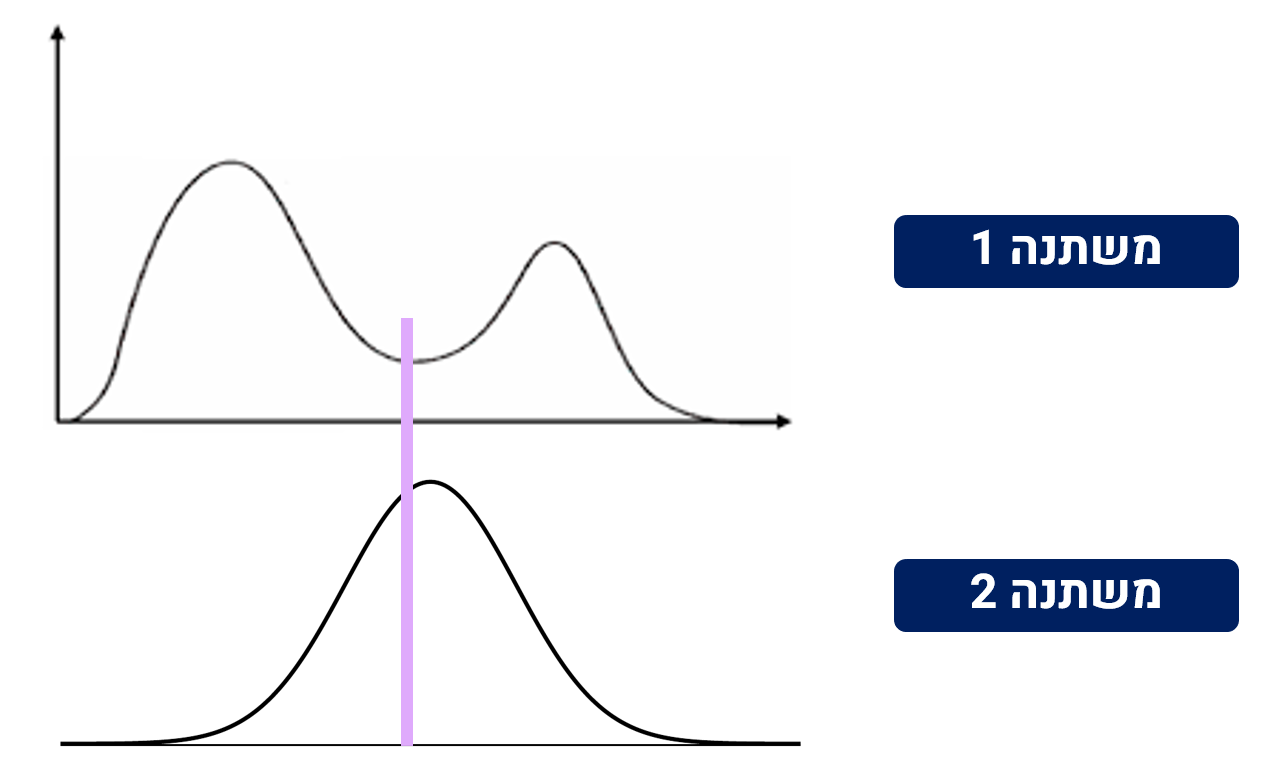
כמובן שגם אם נשמיט חריגים מהמידע –ננסה לשאול את עצמנו מה גורם לחריגים כאלו: האם זו פשוט שגיאה במידע, או אולי תופעה שיש לחקור אותה? לא נתעלם מקיומם כי יכול להיות שנצליח למצוא תובנות מעניינות מחריגויות.

*שאלות הבנה*

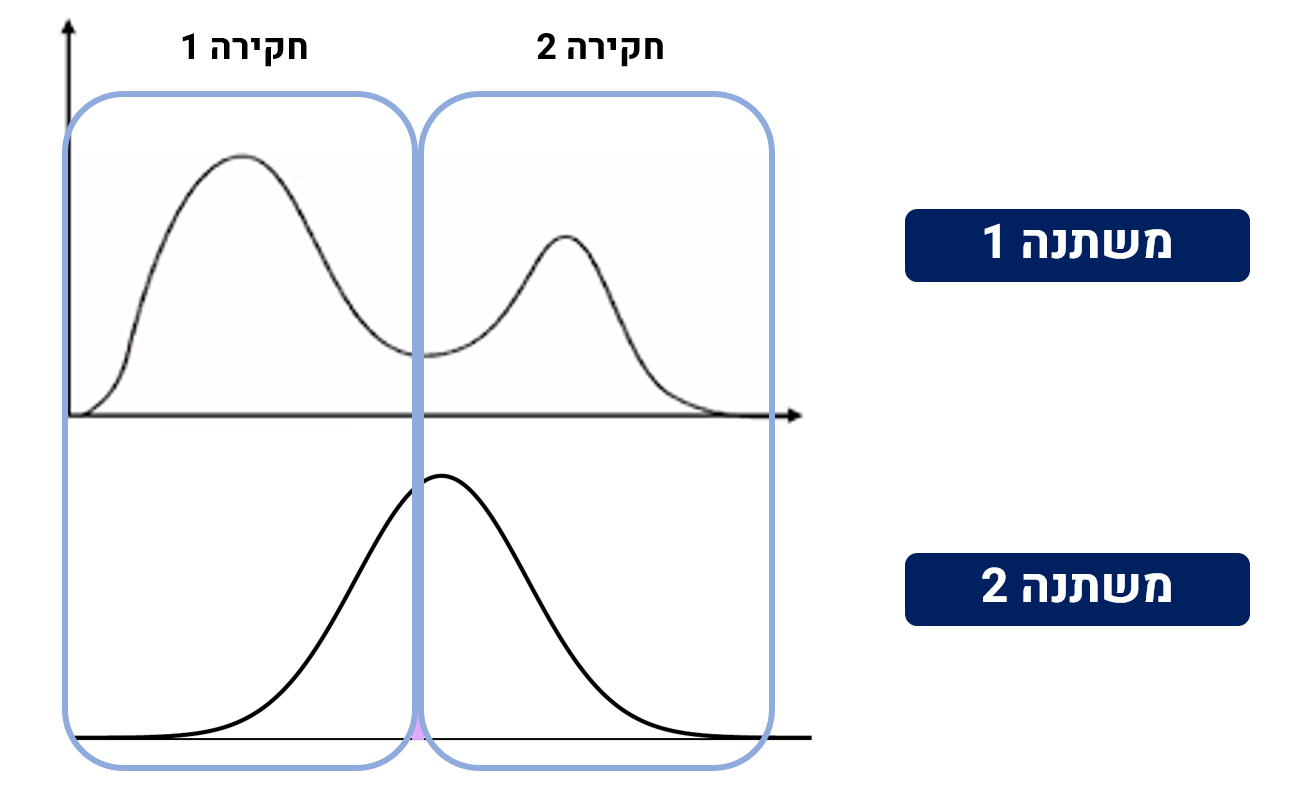
1. מה ההבדל בין המשמעות של r  לבין המשמעות של r בריבוע?
2. מהי רגרסיה?
3. למה חשוב להסתכל על תרשים הפיזור לפני שמוסיפים את קו הרגרסיה? הסבר בהרחבה.
4. מהו התהליך העבודה שאני צריך לעשות כאשר אני רוצה למצוא קשר בין שני משתנים רציפים? (כתוב את השלבים)

**התמודדות עם מידע דו מודלי**

כפי שלמדנו בשיעורים הראשונים כשיש מידע דו מודלי – נפצל את השיאים. נמצא את הנקודה בה ניתן לחלק את ההתפלגות לשני שיאים מוגדרים. בנקודת החיתוך נחתוך גם את המשתנה השני.



כעת יש לנו שני חלקים: משתנה א' וב' לפני נקודת החיתוך ואחרי נקודת החיתוך. בין כל חלק של א' ו-ב' בהתאמה נבצע קורלציה ונראה את הקשר. נעשה חקירה נפרדת עבור כל שיא על מנת "לנטרל" את ההשפעה של הגורם השלישי שכנראה יצר את המודליות הזו.



נסיק מסקנות עבור כל שיא.

**תהליך העבודה**

* העלאת כל אחד מהמשתנים על היסטוגרמה (אם המשתנה נומרי רציף ולא בדיד), תיאור המשתנים וההתפלגות.
* העלאת המידע על תרשים פיזור.
* בחינת חריגויות והסבר על דרך הפעולה הנבחרת
* בדיקת חוזק הקשר (ערך r), r בריבוע וp-value (רק אם קטן מ0.05 נמשיך בתהליך).
* מתן הסבר מילולי - פירוש חוזק הקשר והרמה שבה המשתנים מסבירים זה את זה.
* אם ניתן והמידע עומד בתנאים- נוסיף את קו הרגרסיה ומשוואת הרגרסיה לתרשים (אם ננסה לבצע חיזוי - נשים לב שהמשתנה שאנחנו רוצים לחזות על ציר y, והמשתנה המסביר על ציר x).



1. אקסטרה – סרטון קצר שמסביר על נוסחת פירסון - <https://www.youtube.com/watch?v=2B_UW-RweSE> [↑](#footnote-ref-2)
2. ניתן לייחס רנדומליות לגורמים שאנחנו לא מכירים [↑](#footnote-ref-3)