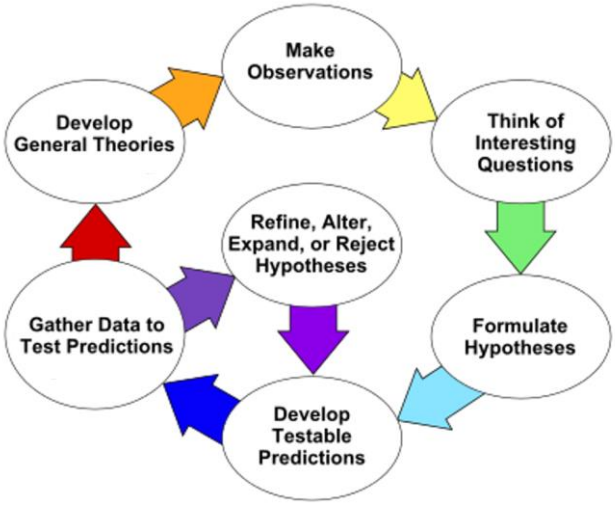
אנליזת נתונים

תוכן

1 השיטה המדעית (Scientific Method)

השיטה המדעית היא אוסף של מתודות וקריטריונים לביצוע מחקר מדעי. השיטה התפתחה לאורך השנים וכיום מהווה דרך לביסוס או תיקון תיאוריות מדעיות קיימות וצבירת ידע מדעי חדש. על מנת להתקבל בקרב הקהילה המדעית, מחקרים מודרניים מתבצעים לפי השיטה המדעית.



איור 1: השיטה המדעית כתהליך מחזורי [1]

השיטה המדעית ([2]) כוללת היקף רחב של שיטות מדידה וניתוח ממצאים, לצד כלים מחשבתיים ודדוקטיביים (Deduction) לניסוח ובדיקת התיאוריה. שימוש במתודות הללו מאפשר לבצע מחקר תוך הימנעות מכשלים לוגיים[1], כשלים בביצוע הניסוי או פירוש מוטעה של תוצאות ניסוי, העלולים להביא למסקנות שקריות. כמו כן, תכנון וביצוע קפדני של ניסוי ופיענוח נכון של תוצאותיו לפי השיטה מאפשרים להכריע בין רעיונות, השערות ותפיסות שונות ולקבוע מי מהן מתארת את המציאות בצורה הטובה ביותר.

השיטה המדעית כוללת סדרת שלבים וניתן להתייחס אליה גם כתהליך מחזורי השולח את החוקרים שוב ושוב לבדוק את השערותיהם (ראה איור 1). שלבי השיטה מפורטים להלן.

תצפית (observation)

החוקר בוחר תופעה או משפחה של תופעות דומות ואוסף תצפיות שלהן ושל הקשרים ביניהן. התצפית יכולה להתבצע באמצאות כלי מדידה או באמצעות החושים האנושיים וכוללת איסוף של מידע קודם על התופעה. תיעוד התצפיות מוביל בהכרח לצורך בתיאור מדויק שלהן, כזה שכל המדענים יבינו באותו אופן, ולעיתים יש צורך בהמצאת מינוח חדש וספציפי עבור אותן תופעות. בעת תיעוד תצפית, יש חשיבות גבוהה לתעד אותה מיד ולא להסתמך על הזיכרון האנושי. כמו כן, יש להימנע ככל הניתן מהטיה (bias) של התצפית כתוצאה מדעות ורעיונות קודמים של החוקר.

ניסוח שאלה

החוקר מנסח שאלת חקר המבטאת קשר של סיבה ותוצאה בין המשתנים הנצפים וממקדת את תחום המחקר. על השאלה להיות ממוקדת מספיק כדי להבהיר את התנאים ואת התופעות הספציפיות בהן מדובר ואת כיוון המחקר. עם זאת, שאלה טריוויאלית או כזו שניתן לענות עליה במשפטים ספורים לא יכולה להוות בסיס למחקר.

היפותזה (השערה)

החוקר בונה היפותזה המסבירה את התופעה ומאפשרת חיזוי תוצאות של מצב מוגדר חדש, שלא נכלל בתצפיות. החשיבות הפרקטית של ההיפותזה גלומה בכך שהיא מאפשרת לשלוט על מערכות או לחזות את ההתנהגות שלהן. ישנם סוגים רבים של היפותזות, בעלות רמות שונות של מורכבות, ואחת מהן היא בניית מודל מתמטי. המודל הוא פישוט של המציאות לכדי מספר מינימאלי סופי של פרמטרים המשפיעים על התופעה, ולא מהווה תיאור שלם של המצב בטבע. לפיכך, המודל כולל תנאים והנחות המצדיקים פישוט זה. הקביעה אילו גורמים **אינם** משפיעים על התופעה עלולה להיות מסובכת ויש לבסס אותה על התצפיות.

תכנון ניסוי

[1] הסתמכות על אינטואיציה ללא ביסוס או אימות, למשל.

החוקר מתכנן ניסוי כך שתוצאותיו יאמתו או ישללו את ההיפותזה. על הניסוי לקיים את תנאי המודל ולכן בעת התכנון יש להקפיד שהניסוי יאפשר דיוק מספק על מנת להפיק תוצאות משמעותיות ביחס להיפותזה. הניסוי צריך לבודד במידת האפשר את התופעה הנחקרת מגורמים חיצוניים למודל ולאפשר למדוד אותה בצורה מבוקרת. כמו כן הניסוי חייב להיות מתוכנן כך שניתן יהיה לחזור עליו ולצפות לתוצאות דומות (**Reproducibility**).

ביצוע מדידות

החוקר בונה את מערכת הניסוי ומבצע מדידות של התופעה תוך הקפדה על התנאים שהוגדרו במודל והגדלת דיוק המדידות ככל האפשר. בשלב זה יש יתרון ברור לטכניקות מדידה מתקדמות, מכשור מדויק, ושיקולים הנדסיים שונים אשר קובעים בפועל את הדיוק שניתן לייחס לתוצאות המדידה. מבלי להתחשב בכל אלה, פירוש התוצאות יכול להוביל למסקנה שגויה או שהמדידות יתגלו כחסרות משמעות. היכולת לבצע מדידות מדויקות זו מיומנות נרכשת והיעדר יכולת זו עלול לעכב קבלה של תיאוריה מדעית נכונה[2].

אנליזה של המדידות והשוואה עם ההיפותזה

החוקר מעבד את תוצאות המדידה הגולמיות בעזרת כליים מדעיים וסטטיסטיים. יש להתאים את אופן הניתוח למאפייני המדידות, כגון גודל המדגם, התפלגות פיזור המדידות, אי-הוודאות היחסית של המדידות וכו'. מאפיינים אלו ועוד קובעים אילו סוגי אנליזה רלוונטיים ותקפים עבור ניתוח התוצאות של הניסוי. יש להימנע מביצוע אנליזה שמציגה מצג שווא[3] של התוצאות ומובילה למסקנות שאין להן בסיס לפי הניסוי. כמו כן, עיבוד התוצאות צריך להימנע ככל האפשר מהטיה שלהן, כגון השמטת תוצאות חריגות ללא קריטריון אובייקטיבי.

לאחר שלב העיבוד, החוקר משווה את תוצאות המדידה עם התוצאות החזויות לפי ההיפותזה. השוואה זו חייבת להתבצע לפי קריטריון השוואה מוגדר ויש להצדיק לוגית את עצם ביצוע ההשוואה. כמו כן, גם את תוצאת ההשוואה יש לנתח ולהסביר מה הביא לקבלת תוצאה זו.

הסקת מסקנות

החוקר מנסח מסקנות מהניסוי על נכונות ההיפותזה, תיקונים למודל שעלו מתוך התוצאות והרחבה אפשרית שלו. המסקנות צריכות להיגזר באופן ישיר מתוצאות הניסוי ולהיות כמותיות ומנוסחות בצורה ברורה. על המסקנות להתייחס גם למידת הוודאות שהניסוי מספק והביטחון בתוקף התוצאות שלו.

לסיכום, השיטה המדעית מבהירה את המקום והחשיבות של החלקים השונים במחקר, ואת הסטנדרטים המדעיים המגדירים את איכות הידע שנצבר מהתהליך. עבור החוקר הניסיונאי, יש חשיבות גבוהה לחלקים ספציפיים בשיטה, כגון בניית מודל מתמטי פשוט ככל שניתן, ביצוע מדידות מדויקות והפעלת לוגיקה מתקדמת בעת ההשוואה של המדידות עם המודל והסקת המסקנות.

<u>רשימת מקורות:</u>

[1] מקור התמונה שייך לפרופ' תאודור גרלנד, אוניברסיטת קליפורניה, Riveland, 2015.
http://idea.ucr.edu/documents/flash/scientific_method/story.htm

[2] הסבר מפורט של בסיס השיטה המדעית מופיע

Wilson, E. Bright. An Introduction to Scientific Research (McGraw-Hill, 1952), chapter 3.

[2] Dayton Miller 1921-1926 experiment על יחסות פרטית

[3] לדוגמא: חישוב סטיית תקן על מדגם קטן מידי, הצגת גרפים עם סקאלות צירים לא מותאמות לנתונים, ביצוע רגרסיה לינארית לעקום לא לינארי.

2 מדידה ואי-ודאות המדידה

כפי שצוין בפרק "השיטה המדעית", מדידה של גדלים בניסוי היא חלק אינטגראלי מהתפתחות המדע ונחשבת למקור מידע חשוב עבור שיטות המחקר. עם זאת, קיים קושי עקרוני בביצוע מדידה. אם נניח כי קיים **ערך אמיתי (true value)** לגודל מסוים, אין לחוקר אף דרך למדוד את הערך הזה במדויק בניסוי.

בבסיסה של כל מדידה קיימת אי-ודאות בערך הנמדד, ויש לייצג אי-ודאות זו כאשר מתעדים את תוצאת המדידה. ייצוג אי-הוודאות נעשה על ידי קביעת **שגיאה** (**Error**) לכל ערך שנמדד לפי אופן המדידה ושיקולים נוספים. השגיאה מגדירה טווח ערכים עבור הגודל הנמדד שמתחייב להכיל את הערך האמיתי.

הערכת השגיאה במדידה

גורם אחד לאי-ודאות במדידה הוא מכשירי המדידה, אשר תמיד מאפשרים למדוד ערכים עד כדי דיוק מסוים, ולא מעבר לכך[4]. למכשירי המדידה יכולה להיות גם **שגיאה שיטתית**, כזו שגורמת לתוצאת המדידה להיות מוסטת בחוקיות ספציפית מהערך האמיתי. ניתן לפחית את השגיאה של מכשיר המדידה על ידי כיולו. הדיוק שמיוחס למכשיר מתאר את אי-הוודאות של המדידה שהוא מאפשר וקובע את השגיאה של הערכים הנמדדים במכשיר.

סיבה נוספת לאי-הוודאות היא אלמנט האקראיות הקיים בטבע. ככל שמגדילים את הדיוק של המדידות, ניתן להבחין ב**התפלגות (distribution)** של הערכים המתקבלים במדידה. חזרה על המדידה תפיק בכל פעם ערך שונה ולאחר כמות מספקת של חזרות מתקבלת התפלגות של המדידות סביב ערך ממוצע. לפי ההתפלגות ניתן להעריך את ההסתברות לקבל ערכים שונים לפי מרחקם מהממוצע ולאפיין את האקראיות של הגודל הנמדד. ללא שגיאה שיטתית במדידה, הערך הממוצע עבור אינסוף מדידות צפוי להתלכד עם הערך האמיתי של הגודל. בפועל, מבצעים את המדידה מספר סופי של פעמים ולכן התוצאה תמיד תכיל **שגיאה אקראית** כלשהי.

הקביעה מהי השגיאה של ערך מדוד משקללת לתוכה את השגיאות השיטתיות והאקראיות על סוגיהן השונים. טכניקת המדידה (Measurement Procedure) והמכשור בו משתמשים משפיעים במידה עצומה על השגיאה, ותכנון נכון ומוקפד של המדידה יכול להקטין את השגיאה בסדרי גודל רבים, אך לעולם לא להעלים אותה.

סוגי מדידה

טכניקות המדידה מתחלקות לשני סוגים: **מדידה ישירה ומדידה עקיפה**. מדידה ישירה היא השוואה של הגודל אותו מעוניינים למדוד עם מכשיר מכויל, בעוד שבמדידה עקיפה מודדים גדלים אחרים ישירות ומחשבים מתוכם את הגודל הרצוי לפני נוסחה או מודל תיאורטי[5]. השגיאה של ערכים שנמדדו ישירות נקבעת לפי אופן המדידה שלהם, בעוד שהשגיאה של ערך שנמדד בצורה עקיפה "נגררת" מהשגיאות של המדידות הישירות. חישוב השגיאה הנגררת מתבצע לפי הנוסחה המקשרת בין הערכים שנמדדו ישירות לערך הרצוי.

היכולת לבצע מדידה ישירה מתבסס על קיומם של **סטנדרטים** (**Standard,** Measuring Gauge) של היחידות הבסיסיות[6] ומכשירים המכוילים לפיהם. המדע העוסק במדידת וקביעת סטנדרטים אלו נקרא **מטרולוגיה** (**Metrology**), ובמסגרתו מוגדרים גם הקשרים בין היחידות השונות לאלו הבסיסיות. על מנת שתוצאות ניסוי של צוות מחקר מסוים יהיו תקפות עבור ניסוי שמבצע צוות מחקר אחר, שני הניסויים חייבים להתבצע באמצעות מכשירים שכוילו לפי אותם הסטנדרטים.

---

[4] לדוגמה, צפיפות השנתות על הסרגל קובעת את הדיוק שהוא מאפשר למדוד, או כמות הפיקסלים בתמונה קובעת את הדיוק של המידע המופיע בה.

[5] לדוגמה, על מנת למדוד אורך יש להשוות אותו עם סרגל וזו מדידה ישירה. מהירות ממוצעת של גוף ניתן למדוד רק בצורה עקיפה: יש למדוד את המרחק שהגוף עבר בסרגל ואת משך התנועה בעזרת שעון עצר ואז לחלק את הגדלים (שנמדדו ישירות) לקבלת המהירות הממוצעת.

[6] היחידות הבסיסיות תלויות במערכת היחידות בה משתמשים. ב-SI היחידות הבסיסיות במכניקה הן המטר, הקילוגרם והשנייה.

<u>כתיב תקין של ערך מדוד (דוגמה)</u>

לסיכום, כדי שתוצאת מדידה תכיל את כל המידע הדרוש על מנת ליחס לה ערך מדעי, היא צריכה להירשם בפורמט של

$$\text{גודל} = \text{ערך מספרי} \pm \text{שגיאה} \left[\text{יחידות}\right]$$

<u>לדוגמה:</u> $g = 9.7949 \pm 0.0001 \frac{m}{sec^2}$

ברישום זה מובע המידע הבא: הערך שנמדד עבור קבוע הכבידה בניסוי הוא 9.7949. השגיאה 0.0001 מציינת שלפי הניסוי, הערך האמיתי של קבוע הכבידה נמצא בטווח $9.7948 - 9.7950$. היחידות של הערך שנמדד הן מטר לשנייה בריבוע וכדי לרשום את קבוע הכבידה ביחידות אחרות יש לבצע המרת יחידות, לפי הקשרים המוסכמים בין היחידות.

רשימת ערך ללא כל החלקים המצוינים להלן אינה מייצגת מידע שלם, והשמטה של אחד מן הפרטים **מבטלת** את הערך המדעי של התוצאה. כמו כן, באופן עקרוני יש בעיה לקבוע את השגיאה במדויק, ולכן לא ניתן למדוד שגיאה אלא רק להעריך את גודלה. המוסכמה להערכת השגיאה הוא לעגל אותה לספרה הדומיננטית, ובהתאם לשגיאה לעגל את הערך המדוד.

מדידה גולמית	ייצוג המדידה לפי המוסכמה	הערות
$X = 1.23456 \pm 0.1234 Joule$	$X = 1.2 \pm 0.1 Joule$	השגיאה עוגלה לספרה בודדת והערך עוגל עד לספרה של השגיאה.
$X = 20.03041 \pm 0.00256 kg$	$X = 20.030 \pm 0.003 kg$	הספרה 5 עוגלה כלפי מעלה בשגיאה, והערך רשום עד לספרת השגיאה, גם אם הוא 0.
$X = 123456.78912 \pm 0.0023 m$	$X = 123456.789 \pm 0.002 m$	ניתן לרשום הרבה ספרות בערך, אם הן מוצדקות על ידי השגיאה.
$X = 0.00004523 \pm 0.000007 \Omega$	$X = 45 \pm 7 \mu\Omega$ או $X = (45 \pm 7) \cdot 10^{-6} \Omega$	כדאי להשתמש בקיצורים מקובלים כדי לייצג סדרי גודל.

טבלה 1: דוגמאות לעיגול תוצאות ממדידה.

<u>רשימת מקורות:</u>

[1] הסבר נוסף לטבען של שגיאות מדידה מופיע

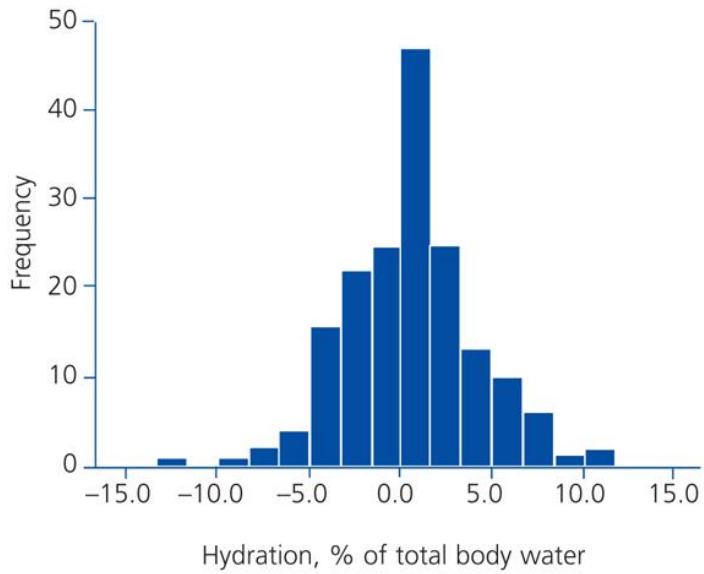
G. L. Squires, "Practical Physics", Cambridge University Press, 4th edition (2001), Ch. 2

3 סטטיסטיקה והתפלגות

כפי שהוזכר בפרק "מדידה ואי-וודאות המדידה", בכל ניסוי קיים אלמנט של אקראיות המשפיע על תוצאות המדידה. לכן, אם נבצע מדידה של אותו גודל n פעמים נצפה לקבל n תוצאות קרובות אך לא זהות. על מנת לפענח את המשמעות של התוצאות הנ"ל יש להשתמש בכלים סטטיסטים.

3.1 היסטוגרמה

אחת מן הצורות הגרפיות להציג תוצאות של מדידות חוזרות היא **היסטוגרמה (Histogram).** תיאור מפורט תהליך בניית היסטוגרמה מופיע ב-[1]. ראו לדוגמה את ההיסטוגרמה באיור 2.



איור 2: התפלגות השינויים של כמות המים בגוף במצבי התייבשות באחוזים מכמות המים הכוללת בגוף, מתוך [2].

ההיסטוגרמה מציגה את **התפלגות המדידות** (**Distribution**) ומכמתת את השכיחות של הערכים שנמדדו. באופן זה היא מאפיינת את האקראיות של הגודל הנמדד ומאפשרת לחזות מה ההסתברות (סיכוי) למדוד כל ערך במדידה עתידית. אם לאחר מספר רב של מדידות, עשירית מסך המדידות היו בתחום מסוים, נסיק שההסתברות למדוד ערך בתחום זה היא 0.1. בחירת מספר הקטעים (נקראים bins) שאליו מחלקים את התחום היא שרירותית ונעשית לפי התכונות של המדידות אותן נרצה להציג באופן ברור.

3.2 גדלים סטטיסטיים

ניתן לייצג את תכונות ההתפלגות באמצעות הגדרת שני גדלים סטטיסטיים: ממוצע ו**סטיית תקן (Standard Deviation)**. גדלים אלו מאפשרים לנתח את השגיאה האקראית של מדידה בודדת ושל אוסף המדידות יחדיו.

הממוצע של סט מדידות (מדגם, Sample) $x_1, x_2, \dots x_i \dots x_n$ מוגדר להיות:

$$\langle x \rangle = \frac{1}{n}\sum_{i=1}^{n} x_i$$

וסטיית התקן של סט מדידות מוגדרת לפי המרחק של כל המדידות מהממוצע:

$$s = \sqrt{\frac{1}{n-1}\sum_{i=1}^{n}(x_i - \langle x \rangle)^2}$$

יש לשים לב שבסטיית התקן המקדם המופיע במכנה הוא $n-1$ במקום n. זהו תיקון להגדרה של סטיית התקן הנקרא "תיקון בסל" (**Bessel's Correction**) והוא מתקן את ההטיה הנובעת מכך שהסטייה מחושבת על מדגם של תוצאות ולא על התוצאות כולן. סטיית התקן של "האוכלוסייה השלמה" של המדידות (ולא מדגם שלה) מסומנת ב-σ.

סטיית התקן מייצגת את רוחב ההתפלגות סביב הממוצע ולכן מתארת את **השגיאה האקראית של מדידה בודדת**. ככל שסטיית התקן קטנה ההתפלגות צרה יותר והסיכויים למדוד ערכים קרובים לממוצע גבוהים יותר (השגיאה האקראית קטנה). סטיית תקן גדולה מתארת ההתפלגות רחבה ובמקרה זה נייחס לגודל הנמדד שגיאה אקראית גדולה יותר.

אם אנו מודדים m מדידות, **השגיאה של ממוצע המדידות**[7] תחושב לפי:

$$s_m = \frac{s}{\sqrt{m}}$$

לפי הנוסחה ניתן לראות שככל שחוזרים על הניסוי יותר פעמים, השגיאה של הממוצע קטנה. באופן זה ניתן להשיג שגיאה אקראית קטנה כרצוננו, בהתאם למידת ההשקעה שאנו מעוניינים להקדיש לניסוי וחשיבות התוצאות שלו.

3.3 פונקציית צפיפות ההסתברות

במקרה אידיאלי, כאשר מכשיר המדידה אינו מוגבל בדיוק, אם מבצעים מספר גדול מאוד של מדידות ההתפלגות שלהן תשאף לעקום רציף (ולא "מדרגות" כמו באיור 2). כדי לתאר התפלגות כזו נגדיר את **פונקציית צפיפות ההסתברות (Probability Density Function)** ונסמן אותה ב-$f(x)$. פונקציית צפיפות ההסתברות מאפשרת לחשב את ההסתברות למדוד ערך הנמצא בטווח מסוים באופן הבא:

$$\mathrm{P}(x_1 < X < x_2) = \int_{x_1}^{x_2} f(x)dx$$

כאשר P היא קיצור של Probability ולכן הביטוי באגף השמאלי מציין את "ההסתברות ש-X נמצא בין x_1 ל-x_2". עבור תחום קטן מאוד, מ-x ל-$x+dx$, ההסתברות למדוד את הערך היא פשוט $f(x)dx$ (ללא האינטגרל).

3.4 התפלגות נורמלית

פעמים רבות התוצאות של ניסויים שונים מתפלגות **התפלגות נורמלית** (Normal), הידועה גם בשם "התפלגות גאוסיאנית" (Gaussian) או "עקומת פעמון" (Bell Curve). הסיבה לכך שההתפלגות הזו נפוצה כל כך מוסברת על ידי **משפט הגבול המרכזי** [3] בתורת ההסתברות. פונקציית הצפיפות של ההתפלגות הנורמלית היא:

$$f(x) = \frac{1}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

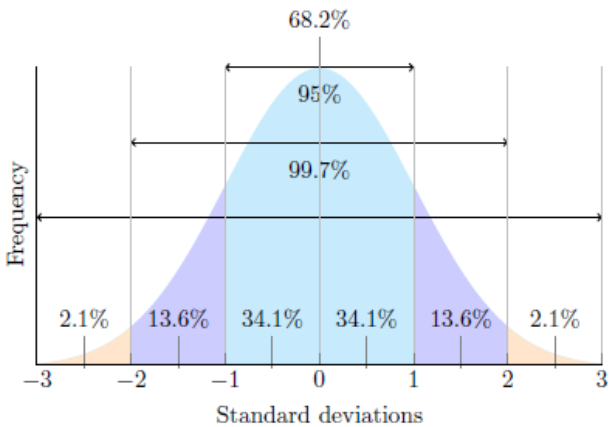
[7] ניתן להוכיח נוסחה זו לפי חישוב שגיאה נגררת עבור הממוצע.

כאשר σ^2 זו סטיית התקן בריבוע הנקראת גם שונות (Variance) ו-μ זהו הממוצע של ההתפלגות הנורמלית. הרישום המקוצר המציין שהגודל X מתפלג נורמלית עם ממוצע μ וסטיית תקן σ הוא:

$$X \sim N(\mu, \sigma^2)$$

(N מציין *"Normal"*).

מבחינה פרקטית, להתפלגות הנורמלית (ראו איור 3) יש מספר תכונות "יפות" שהופכות אותה מתאימה לתיאור התוצאות של רוב הניסויים[8]. יש להתפלגות שיא בודד המהווה את הערך הסביר ביותר והוא גם הממוצע. ההתפלגות סימטרית סביב הממוצע, ובכך מראה שההסתברות למדוד ערך גבוה מהממוצע שווה להסתברות למדוד ערך קטן ממנו. כמו כן, לפי ההתפלגות, ההסתברות למדוד ערך מסוים קטנה ככל שערך זה רחוק מהממוצע, ורוב הערכים שימדדו יהיו קרובים לממוצע (זאת בניגוד, למשל, להתפלגות אחידה).



איור 3: צפיפות ההתפלגות נורמלית כתלות במרחק מהממוצע ביחידות של סטיית תקן [3]. האחוזים מציינים את הכמות היחסית של המדידות הצפויות להתקבל מאוסף נתונים שמפולג נורמלית בכל תחום.

כדי לבנות היסטוגרמה תיאורטית של התפלגות נורמלית, יש לחשב את ההסתברות לכל קטע, כלומר להכפיל את פונקציית הצפיפות ברוחב הקטע dx, ולכפול במספר המדידות n לחישוב מספר המדידות שיתקבלו בכל קטע:

$$Histogram = \frac{n \cdot dx}{\sqrt{2\pi\sigma^2}} e^{-\frac{(x-\mu)^2}{2\sigma^2}}$$

בסטטיסטיקה מגדירים פונקציה נוספת בשם **"Error Function"**, או בקיצור "erf()"[9]. פונקציה זו מחשבת את ההסתברות למדוד ערך בקטע $[-x, x]$ אם הגודל X מתפלג לפי $N\left(0, \frac{1}{2}\right)$. עבור התפלגות נורמלית כללית $N(\mu, \sigma^2)$

$$P(\mu - x < X < \mu + x) = \mathrm{erf}\left(\frac{|x|}{\sigma\sqrt{2}}\right)$$

את הערכים של ה-Error Function ניתן למצוא בטבלאות או לחשב בתוכנה על ידי קריאה לפונקציית *erf()*

3.5 התפלגות CHI SQUARED

התפלגות שימושית נוספת נקראת χ^2 (Chi Squared). זו ההתפלגות של סכום הריבועים של משתנים נורמלים סטנדרטיים, כלומר היא מוגדרת באופן הבא:

נסמן ב-$Z_1, Z_2, \dots, Z_k$ סט של k משתנים אקראיים בלתי תלויים, כאשר כל אחד מהמשתנים מתפלג לפי התפלגות נורמלית סטנדרטית $Z_i \sim N(0,1)$. אזי אם נסמן את סכום הריבועים של המשתנים ב-T:

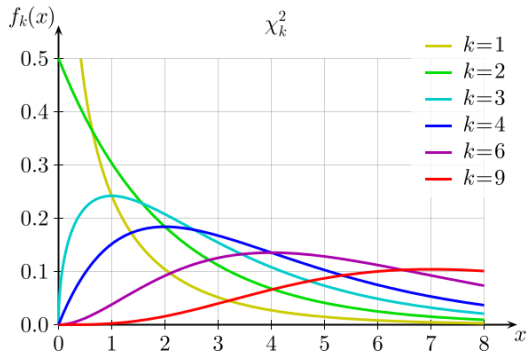
$$T \equiv \sum_{i=1}^{k} Z_i^2$$

[8] זאת אף מבלי להוכיח שתנאיי משפט הגבול המרכזי מתקיימים עבור מערכת הניסוי.
[9] השם הזה מקובל ומוכר בשפות התכנות, כגון Matlab, Mathematica (Wolfram Alpha),Excel.

נקבל (מהגדרה) שהמשתנה T מתפלג

$$T \sim \chi_k^2$$

כאשר k הוא מספר דרגות החופש של ההתפלגות. אם סט המשתנים $Z_1, \dots, Z_k$ **אינם בלתי תלויים** אלא יש משוואות שהם מקיימים ביחד (הגבלות), אז כל משוואה מורידה את מספר דרגות החופש ב-1.



איור 4: פונקצית ההתפלגות של χ^2 עבור מספר דרגות חופש k משתנה. (נלקח מויקיפדיה)

באיור 4 מופיעות פונקציות התפלגות של χ^2 עבור דרגות חופש שונות. ניתן לראות שעבור $k = 1$ ההסתברות של המשתנה T להיות שווה ל-0 גבוהה מאוד וההסתברות שהוא רחוק מ-0 נמוכה. הדבר מתאים לכך ש-Z_1 מתפלג נורמלית סביב 0 ולכן נצפה ש-Z_1^2 יהיה גם כן קרוב ל-0. לעומת זאת, ככל שמספר דרגות החופש עולה, יותר סביר לקבל ערכים גבוהים ב-T, וההסתברות שהוא יהיה למשל 7 כבר לא זניחה.

3.5.1 מבחן *Chi Squared* להתאמה של גרף

אחד מהשימושים של התפלגות χ^2 היא לכמת את מידת ההתאמה של עקום או פונקציה לסט של נקודות מדודות ולהגיד עד כמה המודל שמיוצג בפונקציה סביר בתור הסבר למדידות, או שצריך לדחות את המודל ולחפש פונקציה אחרת יותר מתאימה. מבחן כזה נקרא goodness of fit והוא עונה בדיוק על הצורך הזה.

נניח שיש לנו מודל המגדיר את הקשר בין שני משתנים x, y כ-$y = f(x)$. כפי שכבר צויין בפרק "מדידה ואי וודאות המדידה", בכל מדידה יש שגיאה מסוימת ולכן המודל המלא של מדידה ספציפית הוא

$$y_i = f(x_i) + e_i$$

כעת עלינו לקבל את ההנחה שהשגיאה e_i היא בלתי תלויה בשגיאות של מדידות אחרות, והיא מתפלגת נורמלית עם סטיית תקן מסויימת סביב 0:

$$e_i \sim N(0, \sigma_i^2)$$

מכך נובע שניתן לרשום

$$Z_i \equiv \frac{y_i - f(x_i)}{\sigma_i} \sim N(0,1)$$

כלומר, בנינו משתנה שמחושב מתוך המדידה וההתפלגות שלו (לפי ההנחה שהנחנו על השגיאה) היא התפלגות נורמלית סטנדרטית.

עבור סט של N מדידות הכולל גם הערכה של השגיאה בכל מדידה $\{x_i, y_i, e_i\}_{i=1}^N$ ופונקצית f נתונה, נשתמש בשגיאה e_i בתור סטיית התקן במודל הסטטיסטי σ_i ונחשב את T:

$$T = \sum_{i=1}^{N} \left(\frac{\left(y_i - f(x_i)\right)}{e_i} \right)^2$$

בשלב השני נחשב את דרגות החופש: אם הפונקציה f מכילה m פרמטרים שמחושבים מ- N המדידות, מספר דרגות החופש (שנהוג לסמן אותו ב-d עבור degrees of freedom) הוא

$$d = N - m$$

כדי להחליט האם המודל מהווה הסבר **סביר** למדידות, נמצא מהי ההסתברות לקבל את הערך של T שחישבנו או ערך גבוה (קיצוני) יותר ממנו, בהינתן שההתפלגות שלו היא χ^2_d. ההסתברויות האלו מופיעות בטבלאות מכיוון שלא קיימת פונקציה קדומה להתפלגות χ^2. ההסתברות הזאת נקראת "**ערך *p***" של התאמת המודל (*"p-value"*).

כלל ההחלטה: אם **הערך *p*** גבוה מ-5% נאמר שיש הסתברות לא מבוטלת לקבל את הדגימות שמדדנו מהמודל הנבדק, ולכן סביר שהוא אכן המודל המתאים למדידות. אם הערך קטן מ-5% נאמר שלא סביר שהמודל שלנו ייתן את המדידות שמדדנו ולכן הוא לא מתאים לנתונים. הבחירה ב-5% היא רמת המובהקות של המבחן. רמות נהוגות בקהילה המדעית הן 1% או 5%, כאשר ככל שרמת המובהקות נמוכה יותר, נקבל התאמה של המודל יותר "בקלות" (תנאים פחות מחמירים).

נדגיש כי מבחן χ^2 עונה על השאלה: ***האם סביר לטעון שהמודל מתאים למדידות?***

דוגמה לשימוש במבחן χ^2

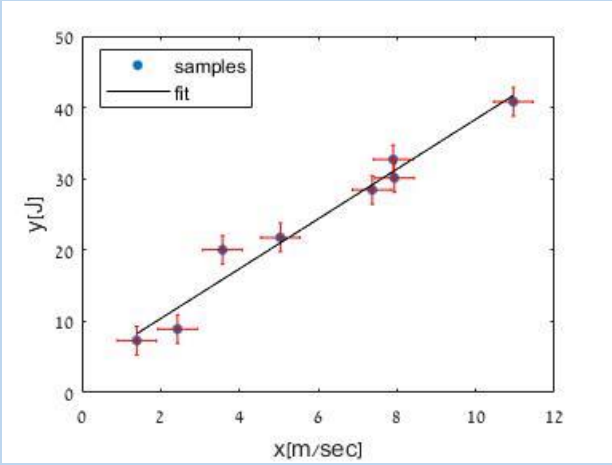
נניח שמדדנו את הנתונים הבאים:

x[m/sec]	1.4	2.43	3.57	5.04	7.91	7.37	7.94	10.96
y[J]	7.3	8.92	20.03	21.76	32.71	28.43	30.11	40.79

והערכנו שהשגיאה ב-*x* היא $\delta x = 0.5\,\frac{\mathrm{m}}{\mathrm{sec}}$ והשגיאה ב-*y* היא $\delta y = 2\mathrm{J}$. המודל שלנו (הפונקציה f) הוא לינארי:

$$y = ax + b$$

כאשר $a\left[\frac{\mathrm{J}\cdot sec}{m}\right], b[\mathrm{J}]$ הם פרמטרי המודל. לאחר ביצוע רגרסיה לינארית התקבל הגרף הבא:



ופרמטרי הרגרסיה ששוערכו הם: $\hat{a} = 3.5 \pm 0.3\,\frac{\mathrm{J}\cdot s}{m}$ ו- $\hat{b} = 3 \pm 2\mathrm{J}$. השגיאה הנגררת המחושבת לפי המודל היא

$$e = \sqrt{(\delta y)^2 + (\hat{a}\cdot\delta x)^2} = 2.66\mathrm{J}$$

והיא תקפה לכל הנקודות. אם הייתה לנו הערכת שגיאה שונה לכל נקודה אז היינו צריכים לחשב כל אחת בנפרד.

חישוב של T לפי הנוסחה נותן כי

$$\mathrm{T} = \sum_{i=1}^{N}\left(\frac{\left(y_i - \left(\hat{a}x_i + \hat{b}\right)\right)}{e_i}\right)^2 = 4.63$$

$$d = N - 2 = 6$$

לפי הטבלה של התפלגות χ^2 (ראה בנספחים), ערך *p* של T הנ"ל הוא בין 0.5 ל-0.75, כלומר גדול מ-0.05. לפיכך ניתן לאמר שהמודל מתאים במידה סבירה לנקודות שמדדנו.

**הטבלה והקוד ב-MATLAB להכנת וניתוח המדידות מצורפים בסוף המסמך בנספחים.*

3.6 מבחני השוואה וקריטריונים סטטיסטיים

3.6.1 מבחן להשוואה בין שני גדלים לפי מובהקות סטטיסטית

במקרים רבים נרצה להשוות תוצאות של שתי שיטות של מדידות או חישובים ולקבוע האם הן זהות. לא נצפה לקבל בדיוק את אותם ערכים בשתי המדידות ולכן השאלה שנשאל היא **האם ההבדל בין הערכים שמדדנו מוסבר על ידי השגיאה האקראית בכל מדידה?**

לשם כך נמצא עבור כל מדידה טווח ערכים (קטע) אשר יכיל את הערך האמיתי במידת ביטחון של 95%. המשמעות של הקטע הזה היא שאם היינו חוזרים על המדידה שוב ושוב, ב-95% מהמקרים היינו מקבלים מדידות שמוכלות בקטע. טווח הערכים הנ"ל נקרא "**רווח בר-סמך**" (Confidence Interval או בקיצור CI) והגודל שלו תלוי ב-Confidence Level שאנו דורשים (במקרה זה – 95%). אילו נדרוש Confidence Level = 99% הטווח יגדל על מנת להבטיח זאת. עבור התפלגות נורמלית, <u>הדרישה של *Confidence Level = 95%* מתאימה בקירוב לתחום של שתי סטיות תקן מכל צד של הממוצע (ראו איור 3).</u>

<u>כלל ההחלטה להשוואה:</u> לאחר שנגדיר את התחום עבור כל מדידה בהתאם לסטיית התקן שלה, נבדוק אם יש חפיפה בין התחומים. אם התחומים חופפים, נאמר שניתן להסביר את ההבדל בין הערכים שקיבלנו בכל מדידה על ידי האקראיות שבמדידות, ושבפועל מדדנו את אותו ערך בשתי שיטות. אם אין חפיפה בין התחומים, נסיק מכך שבכל שיטה נמדד גודל אחר והערכים שקיבלנו אינם זהים מבחינה סטטיסטית.

יש להדגיש כי רווח בר-הסמך שאנו מחשבים אינו בהכרח "ממורכז" סביב הערך האמיתי, אלא אנו בונים אותו סביב הערך שמדדנו, שבהכרח מכיל שגיאה. לכן כאשר אנו משתמשים בו, הטענה היחידה שלנו היא שהערך האמיתי מוכל **איפשהו** בתוך הקטע, ייתכן אפילו קרוב לאחד מהקצוות שלו. תחת ההנחה הזאת, אם יש חפיפה בין שני קטעים איך בכך אישור שהערך האמיתי באמת נמצא בקטע החופף, אבל אם אין חפיפה, הערכים שאנו משווים בהכרח שונים.

דגש נוסף הוא שחוזק הטענות שלנו מוגדר לפי רמת הביטחון בה אנו משתמשים. אם אנו מבססים את הטענות שלנו ברמת ביטחון של 95%, המשמעות היא שרק במקרה אחד מתוך 20 מקרים אנו נסיק באופן זה טענה שגויה וזו רמת ביטחון שמקובלת עלינו.

<u>דוגמה למבחן השוואה בין שני גדלים</u>

נניח שמדדנו שני גדלים בשתי שיטות ואנו מעוניינים להשוות ביניהם. יש בידינו למשל:

$$A = 5.33 \pm 0.07sec; B = 4.9 \pm 0.5sec$$

תחת ההנחה שהגדלים שמדדנו מתפלגים נורמלית ושניתן להשתמש בשגיאה בתור סטיית התקן של ההתפלגות, מחשבים רווחי בר-סמך עם רמת ביטחון של 95% לשני הגדלים:

$$A \in [5.33 - 2 \cdot 0.07, 5.33 + 2 \cdot 0.07] = [5.19, 5.47]$$

$$B \in [4.9 - 2 \cdot 0.5, 4.9 + 2 \cdot 0.5] = [3.9, 5.9]$$

ניתן לראות שיש חפיפה בין הקטעים ולכן נסיק כי ברמת הדיוק המתאפשרת בניסוי, שני הגדלים זהים מבחינה סטטיסטית.

נשים לב שאילו השגיאות היו קטנות יותר (ניסוי יותר מדויק) ייתכן שלא הייתה מתקבלת חפיפה ואז היינו מסיקים כי הניסוי מספיק מדויק כדי להראות שהגדלים שמדדנו שונים באופן מובהק סטטיסטית זה מזה ויש למצוא הסבר לכך.

3.6.2 קריטריון שוונה (Chauvenet Criterion) לזיהוי מדידות חריגות

כאשר אוספים מדגם של מדידות, לעיתים אחד או יותר מערכים רחוקים באופן שנראה בלתי סביר מהממוצע של המדידות. ערך כזה שרחוק באופן חריג מהממוצע נקרא **"חריג חשוד טעות"** (***Outlier***) משום שהוא חשוד שהתקבל כתוצאה מטעות במדידה או ברישום התוצאה. אם מתאפשר לנו כחוקרים לזהות מה גרם ל-*Outlier* אז ניתן להשמיט אותו מתוצאות ולהמשיך בחישובים הסטטיסטיים. עם זאת, במקרים רבים הגורם לא ידוע ואז יש צורך בקריטריון סטטיסטי אובייקטיבי על מנת לקבוע האם הערך באמת אינו סביר. הקריטריון לזיהוי ערך כזה, בהנחה שההתפלגות היא נורמלית, נקרא "קריטריון שוונה".

קריטריון שוונה מחשב את ההסתברות לקבל את הערך החריג או ערכים חריגים עוד יותר ממנו. אם ההסתברות הזו קטנה מחצי, כלומר היינו אמורים לקבל פחות מ-"חצי דגימה" של הערך הנ"ל, אז יש להשמיטו כיוון שלא סביר שמדדנו אותו באמת. עבור סט של דגימות שהממוצע שלהן הוא $\langle x \rangle$ וסטיית התקן של המדידות היא s, ההסתברות לקבל ערך חשוד $x_{suspect}$ היא:

$$P = 1 - \mathrm{erf}\left(\frac{\left|x_{suspect} - \langle x \rangle\right|}{s\sqrt{2}}\right)$$

ולכן מספר הדגימות החריגות שנצפה לקבל בסט שמכיל n דגימות הוא $n \cdot P$. אם מספר הדגימות קטן מחצי אז נוכל להשמיט את הדגימה החשודה לפי קריטריון שוונה.

<u>מונחים להרחבה בתחום של מבחני מובהקות סטטיסטית:</u>

Significance Level, Student's Distribution, T-Test, P-Value, Null Hypothesis

<u>רשימת מקורות:</u>

[1] G. L. Squires, *Practical Physics*, Cambridge University Press, 4th edition (2001), Ch. 3

[2] מקור ההיסטוגרמה:

Caravaca, Francisco, et al. "Hydration status assessment by multi-frequency bioimpedance in patients with advanced chronic kidney disease."*Nefrologia* 31.5 (2011): 537-544.

[3] הסבר להוכחה למשפט הגבול המרכזי (Central Limit Theorem):

Lemons, D. S., & Langevin, P. (2002). *An introduction to stochastic processes in physics*. JHU Press. Ch.5, Pg. 36-38.

[4] מקור התמונה: John Canning

http://johncanning.net/wp/?p=1202

4 אנליזת רגרסיה (Regression Analysis)

אנליזת רגרסיה[10] היא כלי למציאת הקשר בין משתנה בלתי תלוי למשתנה תלוי מתוך אוסף נתונים. זיהוי או אפיון הקשר הזה יכול להיעשות למספר מטרות שונות. ניתן להשתמש בקשר שנמצא ברגרסיה על מנת לאפיין את הנתונים כאשר אין מודל תיאורטי ידוע ובכך לבנות מודל אמפירי (ניסיוני). שימוש נוסף הוא חיזוי של ערכים סבירים עבור נתונים שלא נמדדו, על סמך אלו שכן נמדדו בניסוי (שימוש זה נקרא Interpolation או Extrapolation). בנוסף, אם ידוע מודל תיאורטי הקושר בין המשתנה התלוי לבלתי תלוי וחוזה קשר ספציפי, הרגרסיה מאפשרת לבחון עד כמה הקשר הנ"ל מתאר את אוסף הנתונים שנמדדו.

הקשר שאנליזת הרגרסיה מוצאת הוא תמיד **קשר פונקציונאלי** בין המשתנה התלוי לבלתי תלוי לפי קריטריונים סטטיסטיים. הרגרסיה משתמשת באחד מאוסף של אלגוריתמים שונים על מנת לחשב מהו העקום "המתאים ביותר" לתיאור הנתונים שנמדדו. הקריטריון מהי "התאמה" משתנה מאלגוריתם לאלגוריתם, ולכן גם התוצאות שהם מפיקים עבור אותו סט נתונים הן שונות. חלק מהאלגוריתמים הם איטרטיביים (Iterative), ומתכנסים לפתרון לאחר מספר איטרציות (חזרות). באלגוריתמים מסוג זה, ניתן לעיתים לקבל תוצאות שונות (עקומי התאמה שונים) גם בהפעלה חוזרת של האלגוריתם על אותו סט נתונים.

בהינתן מודל המתאר את הקשר בין המשתנים (תבנית של פונקציה או משפחה של פונקציות), אנליזת רגרסיה על הנתונים מחשבת את הפרמטרים של המודל ואת אי הוודאות (השגיאה) שלהם. החוקר מזין לתוך אלגוריתם הרגרסיה את המודל שמכיל מספר פרמטרים לא ידועים וכן את אוסף הנתונים שהוא מדד. תוצאת הרגרסיה היא הערכים של הפרמטרים שהוא הגדיר במודל והשגיאות בהם. כמו כן, תוצאת הרגרסיה כוללת מידע סטטיסטי שמתאר את טיב ההתאמה שהושגה באמצעות האלגוריתם.

4.1 רגרסיה לינארית

רגרסיה לינארית פשוטה מניחה כי המודל המקשר בין המשתנים בניסוי הוא לינארי, כלומר מהצורה:

$$y = a \cdot x + b$$

כאשר x, y הם המשתנים הנמדדים בניסוי ו-a, b הם פרמטרי המודל.

עבור סט נתונים מדוד $(x_1, y_1), (x_2, y_2), \dots, (x_N, y_N)$ ביצוע רגרסיה לינארית על נתוני המדידה מפעיל אלגוריתם שמחשב את ערכם של a, b. הפרמטרים מחושבים כך שה"חיזוי" של המודל לפי ערך נתון של x_i קרוב ככל האפשר לערך שאכן נמדד עבור y_i. במילים אחרות, המשערך (approximator) $\hat{y}_i$ עבור מדידת x_i מסוימת הוא:

$$\hat{y}_i = a \cdot x_i + b$$

והאלגוריתם מחשב את a, b כך ש-$\hat{y}_i$ קרוב ככל הניתן ל-y_i שנמדד.

ההגדרה מהו "ערך קרוב" מתבססת על הגדרת **קריטריון מרחק** מסוים, עליו האלגוריתם מבצע מינימיזציה. ברגרסיה לינארית פשוטה, קריטריון המרחק הוא "**שגיאה ריבועית מינימאלית**" *(Least Squares)*. כלומר, מתבצעת מינימיזציה על סכום ריבועי ה"מרחקים" של המשערכים מערכי y שנמדדו בניסוי (*Sum of Squares of Residuals*):

$$Minimize\ SS_{residual} = \sum_{i=1}^{N\ (num.\ of\ measurements)} (y_i - \hat{y}_i)^2$$

[10] מקור המונח "רגרסיה" הוא במחקר מ-1885 של Sir Francis Galton לתיאור התופעה בה ילדים להורים בעלי גובה חריג (גבוהים או נמוכים) הם בעלי גובה יותר ממוצע מהוריהם. הדבר הביא לתיאוריה שלו של "דעיכה לממוצע" ("Regression toward mediocrity") ולאחר מכן לשימושים של התפיסה במחקרים אחרים של יחסים וקשרים. (עמ' 1)

כדי להעריך את ה"טיב" של המשערך - עד כמה הקו הלינארי מתאר היטב את המידע שמובע במדידות - מגדירים:

$$SS_{total} = \sum_{i=1}^{N}(y_i - \langle y \rangle)^2 \ ; \ \langle y \rangle \equiv mean(y_i)$$

$$\mathrm{R}^2 \equiv 1 - \frac{SS_{residual}}{SS_{total}}$$

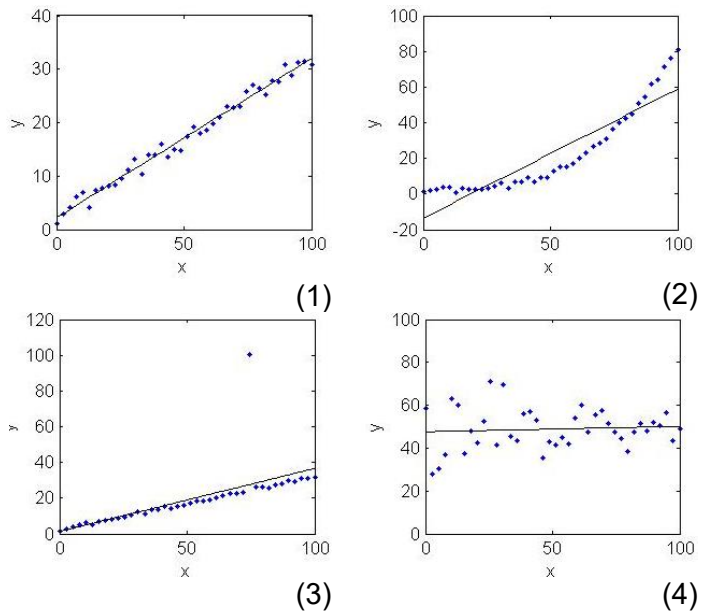
כאשר R^2 נקרא "*R squared*" ומהווה קריטריון להתאמת המודל הלינארי. ככל שהוא שואף ל-1 ההתאמה טובה יותר. לפי הנוסחה ניתן לראות שככל שהערכים המשוערכים רחוקים מהמדידות, *R* קטן. כמו כן, אם התלות של y ב-x חלשה (והרגרסיה תחשב ש $a \approx 0$) אז *R* ישאף ל-0.

לסיכום, הפעלת רגרסיה לינארית על סט של מדידות מספקת את התוצאות הבאות: $\mathrm{R}^2, a, \delta a, b, \delta b$.

4.1.1 הפירוש של R^2

בפני עצמו, *R squared* אינו מהווה קריטריון מוצלח להערכת דיוק הרגרסיה. לדוגמא, בכל הגרפים הבאים (איור 5) יכול להתקבל אותו ערך של *R squared*, בעוד שרק בגרף הראשון הרגרסיה מתארת נאמנה את התוצאות. לפיכך, יש להתחשב בערכו של *R squared* בצירוף עם בדיקה נוספת.

לדוגמה, בדיקה וויזואלית של הגרפים מראה כי בגרף השני התלות בין x ל-y אינה לינארית ויש לבצע רגרסיה אחרת (פולינומית, למשל) על הנתונים. בגרף השלישי, אחת מהמדידות רחוקה מהמגמה הכללית ויש לבטל או לתקן את הנקודה לפני ביצוע הרגרסיה כדי שקו הרגרסיה יתאים למרבית התוצאות. בגרף הרביעי, ה"קפיצות" של ערכי y מעידות שכלל אין תלות בינן לבין ערכי x, ובעצם אין משמעות לביצוע רגרסיה.



איור 5: רגרסיה לינארית להתפלגויות נתונים שונות.

עם זאת, כאשר מבצעים רגרסיה לינארית בצורה מושכלת ותוך התחשבות בשיקולים נוספים כגון אלו, תוצאות הרגרסיה הן בעלות משמעות חזקה למדי. מכיוון שאלגוריתם הרגרסיה הלינארית אינו איטרטיבי, הוא מחושב מהר ותוצאותיו הן חד משמעיות ועקביות, זאת בניגוד לסוגי רגרסיה אחרים. בשל כך, יש יתרון בביצוע ליניאריזציה (*linearization*) למודל התיאורטי. הצגת הקשר בין המשתנה התלוי לבלתי תלוי בצורה לינארית מאפשרת שימוש ברגרסיה הלינארית ומספקת דרך חד משמעית ופשוטה לפענח את תוצאות הניסוי.

נדגיש כי בשונה ממבחן χ^2 שהזכרנו, קריטריון R^2 עונה על השאלה **כמה מהמידע שמיוצג בממדידות נשמר כאשר עוברים לייצוג באמצעות קואורדינטה בודדת (קו ישר) במקום שתיים (גרף דו-מימדי)?** והוא אינו קריטריון סטטיסטי/הסתברותי.

4.1.2 *ביצוע רגרסיה לינארית ב- MATLAB (גרסה A2014 והלאה)*

על מנת לבצע רגרסיה לינארית ב-*Matlab* יש לקרוא לפונקציה *fitlm* ולספק לה כפרמטרים וקטור של המשתנה הבלתי תלוי (x) ווקטור של המשתנה התלוי (y). הפונקציה מחזירה אובייקט שמייצג את הרגרסיה והשדות של

האובייקט מכילים את הפרמטרים הרצויים: $R^2, a, \delta a, b, \delta b$. כמו כן, השדה *Fitted* מכיל וקטור של הערכים החזויים לפי הרגרסיה ($\hat{y}$) עבור ערכי x הנתונים וכך ניתן לשרטט את עקום הרגרסיה בקלות.

```
linear_regression = fitlm(x,y);
R_squared = linear_regression.Rsquared.Ordinary;
b_coefficient = linear_regression.Coefficients.Estimate(1);
a_coefficient = linear_regression.Coefficients.Estimate(2);
b_error = linear_regression.Coefficients.SE(1);
a_error = linear_regression.Coefficients.SE(2);
plot(x,y,'b.',x,linear_regression.Fitted,'k');
```

על מנת לאלץ את הקבוע החופשי להיות 0 (שקו הרגרסיה יעבור בראשית) יש לקרוא לפונקציה באופן הבא:

```
linear_regression = fitlm(x,y,'Intercept',false);
```

4.2 רגרסיה לא לינארית

כפי שברגרסיה לינארית מניחים כי המודל הוא לינארי ומוצאים את הפרמטרים של המודל, ברגרסיה לא לינארית המודל הוא פונקציה לא לינארית של המשתנה הבלתי תלוי ומכיל פרמטרים שונים. הרגרסיה הלא לינארית משתמשת גם היא בקריטריון התאמה של שגיאה ריבועית מינימאלית אך מביאה השגיאה למינימום באופן אחר (Non-Linear Least Squares).

בניגוד לרגרסיה לינארית, אין נוסחה סגורה לחישוב הערך האופטימאלי של כל אחד מהפרמטרים ולכן משתמשים באלגוריתם איטרטיבי המתחיל מ"ניחוש" של ערכי הפרמטרים, מעדכן אותם לפי ההתאמה שלהם לנתונים ושוב בודק את ההתאמה. לאחר מספר חזרות כאלו של עדכונים האלגוריתם מתכנס לפתרון "יציב" ובכך מוצא את ההתאמה הטובה ביותר לנתונים תחת המגבלות של המודל. אופן עדכון הפרמטרים והקריטריון הקובע מתי הפתרון הנוכחי מספיק קרוב על מנת להפסיק את האיטרציות ולסיים את האלגוריתם נקבעים על ידי אופן הפעלת האלגוריתם וסוג האלגוריתם [1].

ב-MATLAB, ביצוע רגרסיה כללית נעשה באמצעות הפונקציה fit והכנסת הפרמטרים המתאימים. לחלופין קיים curve fitting toolbox שפותחים באמצעות קריאה ל cftool. cftool מציג את האופציות של ביצוע הרגרסיה הלא לינארית בצורה נוחה וישירה ובסיום העבודה עימו ניתן לשמור את אובייקט ה-cfit (curve fit) ל-workspace להמשך עיבוד הנתונים והצגתם [2].

<u>רשימת מקורות:</u>

[1] הסבר לשיטת האיטרציות של non linear least squares

Weisstein, Eric W. "Nonlinear Least Squares Fitting." From *MathWorld*--A Wolfram Web Resource. http://mathworld.wolfram.com/NonlinearLeastSquaresFitting.html

[2] Curve Fitting Toolbox™ User's Guide

http://www.mathworks.com/help/pdf_doc/curvefit/curvefit.pdf, pg. 86-88

5 נספחים

טבלת ההתפלגות של χ^2 עבור הדוגמה לשימוש במבחן χ^2:

Percentage Points of the Chi-Square Distribution

Degrees of Freedom	Probability of a larger value of x^2								
	0.99	0.95	0.90	0.75	0.50	0.25	0.10	0.05	0.01
1	0.000	0.004	0.016	0.102	0.455	1.32	2.71	3.84	6.63
2	0.020	0.103	0.211	0.575	1.386	2.77	4.61	5.99	9.21
3	0.115	0.352	0.584	1.212	2.366	4.11	6.25	7.81	11.34
4	0.297	0.711	1.064	1.923	3.357	5.39	7.78	9.49	13.28
5	0.554	1.145	1.610	2.675	4.351	6.63	9.24	11.07	15.09
6	0.872	1.635	2.204	3.455	5.348	7.84	10.64	12.59	16.81
7	1.239	2.167	2.833	4.255	6.346	9.04	12.02	14.07	18.48
8	1.647	2.733	3.490	5.071	7.344	10.22	13.36	15.51	20.09
9	2.088	3.325	4.168	5.899	8.343	11.39	14.68	16.92	21.67
10	2.558	3.940	4.865	6.737	9.342	12.55	15.99	18.31	23.21
11	3.053	4.575	5.578	7.584	10.341	13.70	17.28	19.68	24.72
12	3.571	5.226	6.304	8.438	11.340	14.85	18.55	21.03	26.22
13	4.107	5.892	7.042	9.299	12.340	15.98	19.81	22.36	27.69
14	4.660	6.571	7.790	10.165	13.339	17.12	21.06	23.68	29.14
15	5.229	7.261	8.547	11.037	14.339	18.25	22.31	25.00	30.58
16	5.812	7.962	9.312	11.912	15.338	19.37	23.54	26.30	32.00

קוד ה-*MATLAB* של הדוגמה:

```
%creating samples (x,y)
N = 8;
x_error = 0.5;
y_error = 2;
a=4;b=0.5;
x=linspace(1,10,N)+normrnd(0,x_error,[1,N]);
y= a*x+b+normrnd(0,y_error,[1,N]);

%processing the samples
fit_obj = fitlm(x,y);
figure;
errorbarxy(x,y,repmat(x_error,1,8),repmat(y_error,1,8));
hold on
plot(x,fit_obj.Fitted,'k');
legend('samples','fit');
xlabel('x[m/sec]');
ylabel('y[J]');
b_estimate = fit_obj.Coefficients.Estimate(1);
a_estimate = fit_obj.Coefficients.Estimate(2);
error_estimation = sqrt(y_error^2+(a_estimate*x_error)^2);
T=sum(((y-(a_estimate*x+b_estimate))/error_estimation).^2);
d = N-2;
```